Proceedings of Machine Learning Research vol vvv:1–13, 2024

Learning-Based Estimation of Fitness Landscape Ruggedness for Directed Evolution

Sebastian Towers SEBASTIAN.TOWERS@ENG.OX.AC.UK
Jessica James JESSICA.JAMES@ENG.OX.AC.UK
Harrison Steel HARRISON.STEEL@ENG.OX.AC.UK
Idris Kempf† IDRIS.KEMPF@ENG.OX.AC.UK
Department of Engineering Science, University of Oxford, Parks Road, Oxford OX1 3PJ, UK
† *Corresponding author*

Abstract

Directed evolution is a method for engineering biological systems or components, such as proteins, wherein desired traits are optimised through iterative rounds of mutagenesis and selection of fit variants. The process of protein directed evolution can be envisaged as navigation over high-dimensional landscapes with numerous local maxima, mapping every possible variant of a protein to its fitness. The performance of any strategy in navigating such a landscape is dependent on several parameters, including its ruggedness. However, this information is generally unavailable at the outset of an experiment, and cannot be computed using analytical methods. Here we propose a learning-based method for estimating landscape ruggedness from a mutating population, using only population average performance data. This method uses a short period of exploration at the beginning of an experiment to predict the ruggedness, subsequently guiding the choice of high-performing parameters for directed evolution control. We then simulate this approach on two real-world protein fitness landscapes, demonstrating an improvement upon the performance of standard strategies, particularly on rugged landscapes. In addition to improving the overall outcomes of directed evolution, this method has the advantage of being readily deployable in laboratory settings, even in configurations that exclusively capture average population measures. Given the rapidly expanding application space of engineered proteins, the products of improved directed evolution are relevant in medicine, agriculture and manufacturing.

Keywords: Protein engineering, directed evolution, machine learning.

1. Introduction

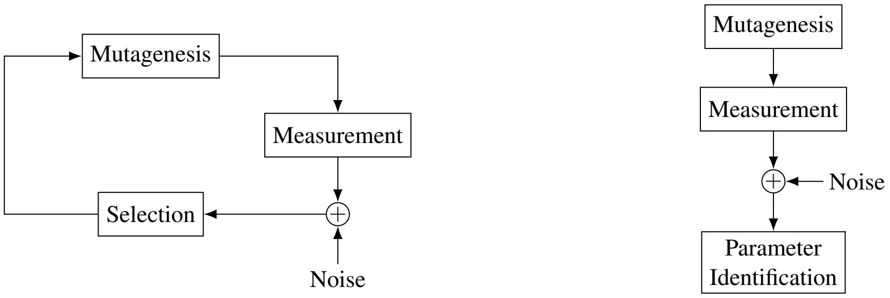
Proteins are fundamental biological components that perform various functions within living organisms. Proteins are variable-length chains of sub-units known as amino acids, for which there are 20 different types. The length and sequence of amino acids in each protein is dictated by a gene, which is composed of DNA. In an ideal world, new proteins could be engineered by *de novo* protein design (Jumper et al., 2021; Baek et al., 2021), which aims at understanding how the sequence of amino acids maps to structure and function of the protein. However, due to the huge combinatorial possibilities (20^N for a protein of length N), *de novo* protein design remains difficult, even with recent advances in computing power and deep learning methods (Dauparas et al., 2022; Ferruz et al., 2022; Singer et al., 2022; Anishchenko et al., 2021; Wicky et al., 2022).

In contrast, directed evolution is a process by which biological components, such as proteins, are engineered and improved through iterative rounds of selection and mutagenesis, emulating the natural evolution process (Arnold, 1998). Directed evolution has had many successful applications, including the development of drugs (Nixon et al., 2014) and biofuels (Heater et al., 2019). The problem of directed evolution can be interpreted as navigation over a protein *fitness landscape* (Wright, 1932). Fitness landscapes are high-dimensional structures in which the sequence of the protein represents a coordinate that maps to a fitness value, which, in the context of this work, is defined as the property the directed evolutionary process aims to optimise. Fitness landscapes are known to exhibit variable degrees of ruggedness, which can create local optima that constrain paths of evolution (Wu et al., 2016).

The standard approach to directed evolution is to select the best performing variants with each iteration. However, this approach can be prone to getting trapped in local optima (Carpenter et al., 2023). In recent years, numerous optimisation methods have been developed, leveraging machine learning to actively navigate a protein fitness landscape (Wu et al., 2019; Wittmann et al., 2021; Yang et al., 2019; Frisby and Langmead, 2021; Hu et al., 2023; Fox et al., 2003). Although effective, these optimisation approaches require sequencing of the entire population of variants with each iteration. This makes them labour- and resource-intensive, and not applicable to novel continuous directed evolution methods where DNA sequence data is not generally available during an experiment (Molina et al., 2022). In a previous work, we proposed strategies that can be used to optimise directed evolution without the need to sequence all variants (James et al., in press). We found, however, that the performance of each strategy is dependent on the properties of the underlying landscape, information that is generally not available at the outset of an experiment.

It has previously been shown that neural networks can be used to infer evolutionary parameters, such as rate of accumulation of beneficial mutations (Avecilla et al., 2022). Statistical models have also been generated for inferring protein fitness landscape properties from directed evolution trajectories with sequencing information (D'Costa et al., 2023). In this paper, we use a neural network to estimate properties of a protein fitness landscape without sequencing information. The method requires measurements of the average fitness from a population mutating away from a starting point, which is easily implementable in various experimental configurations. Such data can be collected by mutating a single population of bacteria (e.g. by UV), and recording the average population fitness value (e.g. fluorescence) at regular intervals. The resulting measurements are input into a fully connected neural network (FCN), which has been pre-trained on theoretical fitness landscapes to predict ruggedness. The use of an FCN is required as fitness landscapes represent a highly nonlinear mapping from DNA sequence to protein functions and properties, which cannot be captured by traditional methods such as linear regression. This estimate is then used to select directed evolution control parameters that are sensitive to the ruggedness, such as identified in our previous research (James et al., in press). We analyse the prediction accuracy of the FCN with respect to the number of fitness values provided, and demonstrate that the ruggedness can be reliably estimated, resulting in a performance increase of directed evolution experiments in highly rugged landscapes. Finally, we apply our estimation procedure to two real-world landscapes and demonstrate that it can be used in practical settings, even though trained on theoretical models.

The paper is organized as follows. Section 2 frames directed evolution as a control problem. In Section 3, the proposed method for parameter inference is developed. The paper is concluded by evaluating the proposed method on empirical landscapes.



(a) Closed loop.

(b) Open loop.

Figure 1: Block diagrams for directed evolution in closed-loop and open-loop configurations.

2. Problem Formulation

Directed evolution can be represented by the feedback loop shown in Fig. 1a, which includes measurement noise. Let $P_k = \left[p_k^1 \quad \dots \quad p_k^n\right]^\mathrm{T} \in \mathcal{P}^n$ denote the n members of the evolved population at iteration $k \in \mathbb{N}$, $F : \mathcal{P} \mapsto \mathbb{R}$ the fitness function, $S : \mathbb{R}^n \mapsto \{0,1\}^{n\times n}$ the selection process, and $M : \mathcal{P} \mapsto \mathcal{P}$ the mutagenesis process. For a gene of N loci, each with A possible alleles, $\mathcal{P}$ represents a discrete sequence space with A^N different sequences, $A \in \mathbb{N}^+$. The function F measures the performance of a population member and has multiple local optima in general. The selection process takes $F(P_k)$, where F is applied element-wise, as inputs, and outputs a selection matrix $S(\cdot)$ with exactly one 1 per row, so that the selected variants are obtained from $S(F(P_k))P_k$. Note that the sequences are not observed directly in practice, only their fitnesses $F(P_k)$. The function M is modelled as a stochastic function that changes each element of p_k^i with probability θ and leaves it unchanged with probability $1-\theta$, so that the number of mutations approximately follows a Poisson distribution with mean $N\theta$.

With these definitions, the process from Fig. 1 can be modelled as

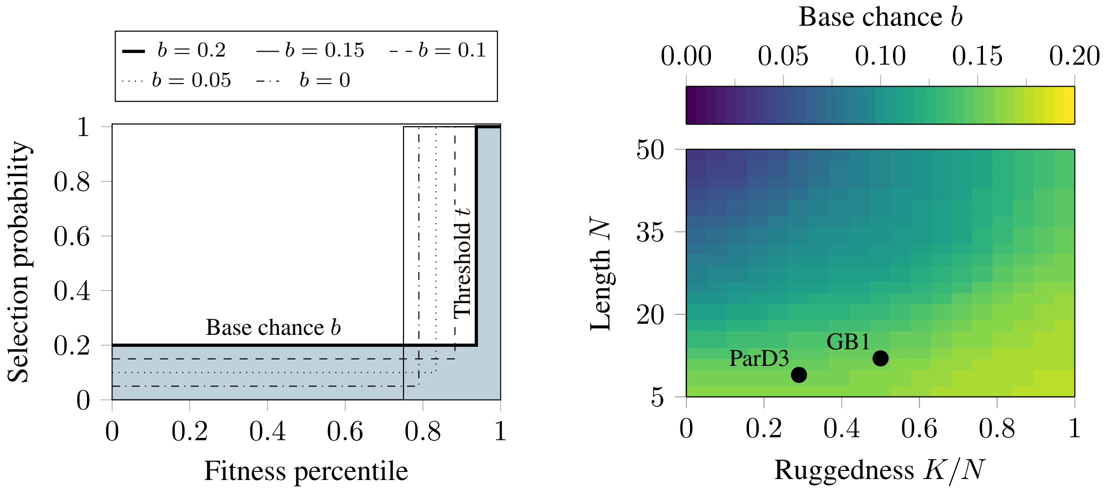
$$P_{k+1} = M\left(S\left(F(P_k) + n_k\right)P_k\right), \tag{1}$$

where $n_k \in \mathbb{R}^n$ refers to the noise, and the functions $F(\cdot)$ and $M(\cdot)$ are applied element-wise. For the remainder of the paper, the effect of noise is ignored. The aim of directed evolution is to maximise the maximum fitness of the population, i.e.

$$\max_{i=1,\dots,n} F(p_H^i), \tag{2}$$

where $H \in \mathbb{N}^+$ is the fixed duration of the experiment. Because F lacks strong structure in general, M is a stochastic function, and the sequences p_k^i are not directly observed, problem (2) cannot be solved using standard optimisation techniques.

The standard approach to selection in directed evolution is to just choose the fittest variants in each generation. This approach is prone to getting trapped in local optima, particularly in rugged landscapes (Carpenter et al., 2023). In order to reduce this propensity, in a previous work we proposed an alternative function for selection shown in Fig. 2a (James et al., in press). The selection function is defined by two parameters: a threshold fitness percentile, $t \in [0,1]$, above which variants are always selected, and a base chance of selection, $b \in [0,1]$, for variants with lower fitness percentiles. The expected fraction of cells f selected at each iteration is represented by the shaded area



(a) Selection function.

(b) Best-performing base chances.

Figure 2: Stochastic selection functions (with $t = (1-f)/(1-b)$ and $f = 0.25$) and corresponding best-performing base chances on the NK landscape (adapted from (James et al., in press)). Figure (b) also shows the estimated K/N for the GB1 and ParD3 landscapes using the method from Section 3.

in Fig. 2a and given by $f := 1 - t(1-b)$. Throughout the paper it is assumed that $f = 0.25$ (James et al., in press), so that for a given $b \in [0, f]$, $t = (1-f)/(1-b)$. This selection function trades off exploration with exploitation, and is therefore less prone to getting trapped in local optima than the aforementioned standard approach. Depending on the properties of fitness landscape, it has been shown that the choice of b significantly affects the outcome of the experiment (James et al., in press). While low base chances (greedy selection) perform well on landscapes with few maxima, high base chances benefit the outcome of the experiment on rugged landscapes. This is captured in Fig. 2b, which shows the best-performing b as a function of the gene length N and a ruggedness measure K/N on the theoretical NK landscape (Section 3.1) for a fixed number of iterations $H = 100$. For heavily rugged landscapes ($K/N \approx 1$), high base chances are favourable, whereas for less rugged landscapes ($K/N \ll 1$), low base chances are favourable.

3. Estimating Landscape Ruggedness

The fitness landscapes of biological entities evolved in real-world experiments are not known a priori. The question arises whether landscape properties can be inferred from fitness measurements taken before the start of the experiment in order to choose appropriate control parameters for selection and mutagenesis. To achieve this, a method for inferring a ruggedness measure of the landscape is developed, which is in turn used to select the parameter b. Given that F can be arbitrarily complex, an FCN is trained to estimate the ruggedness from fitness values measured in the open-loop configuration from Fig. 1b, where the selection procedure has been omitted. First, information on the landscape is collected by repeatedly mutating the population and measuring the average fitness. Second, these measurements are preprocessed and fed into the trained FCN, which outputs an estimate of the ruggedness. Finally, the ruggedness estimate is used to choose an appropriate base chance from the look-up table in Fig. 2b.

3.1. NK Landscapes

The successful training of the FCN hinges on the availability of data. At present, there exist few empirical data sets that map sequence space to fitness measurements (Section 4). These empirical data sets are therefore not in sufficient quantity for training the FCN, nor are they accompanied by a clear definition of ruggedness for training. This problem can be circumvented by using theoretical models of fitness landscapes, such as Fisher's geometric model (Tenaillon, 2014; Fisher, 1931), the holey landscape model (Gavrilets, 1997), the multilinear model (Hansen and Wagner, 2001), the Rough Mount Fuji model (Neidhart et al., 2014) and the NK model (Kauffman and Levin, 1987; Kauffman and Weinberger, 1989), which is used here on account of its tuneable ruggedness and implementation of epistasis (interaction between sub-units, which is prevalent in a protein context).

In the NK model, each gene is represented by a sequence of length N. Every site interacts with K other sites in the gene, influencing the resulting fitness. The number of interactions, determined by K, is what allows ruggedness to be tuned. When $K = 0$, the landscape is linear and has a single peak. The other extreme, $K = N - 1$, is maximally rugged, with all fitness values independent from one another. Values of K between the two interpolate between these two extremes. Note that, even though the underlying generation process is known, finding the global optimum of an NK-landscape is an NP-hard problem for $K > 1$ (Wright et al., 2000).

3.2. Training the Fully Connected Neural Network

To estimate the ruggedness, the population is mutated for G generations in the open-loop configuration from Fig. 1b, so that $P_{k+1} = M(P_k) = M^{(k+1)}(P_0)$. To account for numerical differences between the landscapes, the measurements $F(p_k^i)$ are normalised by the observed mean, $\bar{\mu}$, as $x_k^i := (F(p_k^i) - \bar{\mu})/\bar{\sigma}$, where $\bar{\mu} = \sum_{k=1}^{G} \sum_{i=1}^{n} F(p_k^i)/(Gn)$ and $\bar{\sigma}^2 = \sum_{k=1}^{G} \sum_{i=1}^{n} (F(p_k^i) - \bar{\mu})^2/(Gn)$. The mean, μ_k, of the population fitness of generation k are then computed as $\mu_k := \sum_{i=1}^{n} x_k^i/n$, from which the features provided to the FCN are computed as

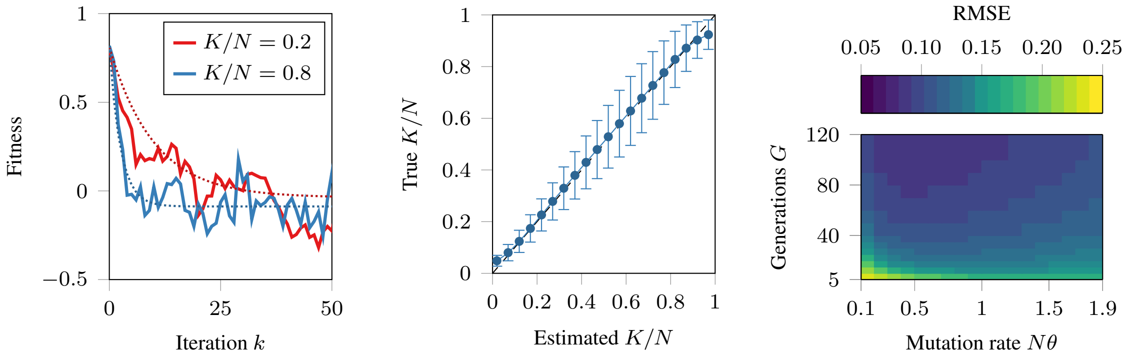
$$u := \begin{bmatrix} \mu_1 & \dots & \mu_G \end{bmatrix}^{\mathrm{T}} \in \mathbb{R}^G. \tag{3}$$

The choice (3) is designed to be as simple as possible, while capturing factors relevant to ruggedness estimation. Fig. 3a shows two example trajectories for a smooth ($K/N = 0.2$) and a rugged ($K/N = 0.8$) landscape with $N = 100$, $N\theta = 0.5$, and with $n = 50$ (solid) and $n = 4000$ (dotted). It can be seen that both trajectories converge to the overall mean fitness, $\bar{F}$, of their corresponding landscape, but at different speeds. For a rugged landscape, μ_k converges more quickly to $\bar{F}$ than for a smooth landscape. This is formalised for the NK landscape in the following proposition:

Proposition 1 *Let $F : A^N \to \mathbb{R}$ be an NK landscape, and let p_k^i be a single cell mutating at rate θ per generation. Then:*

$$\mathbb{E}\left[F(p_k^i)|F(p_0^i)\right] - \bar{F} \approx e^{-\theta k K}(F(p_0^i) - \bar{F})$$

Proof See Appendix A. ■



(a) Examples features (3).

(b) Mean error for $G = 50$.

(c) RMSE of $\phi_{\text{fcn}}(u)$.

Figure 3: Example features ($N = 100$ and $n = 50$) and performance of the ruggedness estimator evaluated on a test set with 1.2×10^6 datapoints.

In particular for large n, μ_k can be interpreted as an population estimate of $\mathbb{E}\left[F(p_k^i)|F(p_0^i)\right]$, so that $\mu_k \to \mathbb{E}\left[F(p_k^i)|F(p_0^i)\right]$ as $n \to \infty$. As K is a measure of ruggedness, this implies that the more rugged the landscape, the more quickly μ_k converges to $\bar{F}$. Note that the approximation from Proposition 1 coincides with the dotted line from Fig. 3a, which corresponds to (3) with $n = 4000$.

In order to infer K/N from (3), an FCN, $\phi_{\text{fcn}}(u) \approx \mathbb{E}\left[K/N \mid u\right]$, is trained on a dataset of 1.2×10^6 $(u, K/N)$ pairs generated from a distribution of NK landscapes. The FCN has two hidden layers of size 128 and is trained using gradient descent. The main computational cost is data generation, which takes 10 min using an NVIDIA A40, whereas the training of the FCN takes 2 min. The estimation error of $\phi_{\text{fcn}}(u)$ obtained on a test set with 1.2×10^6 datapoints is shown in Fig. 3b for different values of K/N and $G = 50$. It can be seen that $\phi_{\text{fcn}}(u)$ performs well for extreme values of K/N, but worse for intermediary values.

The accuracy of $\phi_{\text{fcn}}(u)$ is further analysed in Fig. 3c for $5 \leq G \leq 120$ and mutation rates $0.1 \leq N\theta \leq 1.9$, which shows the root mean square error (RMSE) averaged over the values of K/N marked in Fig. 3b. Fig. 3c shows that the accuracy increases as G does, but plateaus quickly, which is related to the convergence properties of μ_k shown in Fig. 3a. Fig. 3c also shows that a lower mutation rate requires a larger G for a high accuracy.

4. Translation to Empirical Landscapes

The performance of $\phi_{\text{fcn}}(u)$ is tested on two different empirical landscapes. Empirical landscapes are experimental data sets describing the fitness of all possible variants of a protein region. Measurement of such fitness landscapes can be a highly resource-intensive task, as they increase in size exponentially with the addition of each amino acid position. At present, there is a limited number of empirical protein fitness landscapes, with the largest combinatorially complete example not exceeding four amino acid positions (Weinreich et al., 2006; Khan et al., 2011; Chou et al., 2014; Bank et al., 2016; Wu et al., 2016; Lite et al., 2020; Papkou et al., 2023). There is a necessity, therefore, to train $\phi_{\text{fcn}}(u)$ using theoretical NK landscapes, and to reserve the empirical landscapes for testing. NK landscapes are a simplistic representation of true protein fitness landscapes, given the fact that K is a fixed constant and not variable over the protein, and that the distribution of fitness values is normal, as opposed to being heavily skewed towards zero. Despite this, it is found that the model fares well when applied to real fitness landscapes, and may improve with more tailored theoretical fitness landscapes.

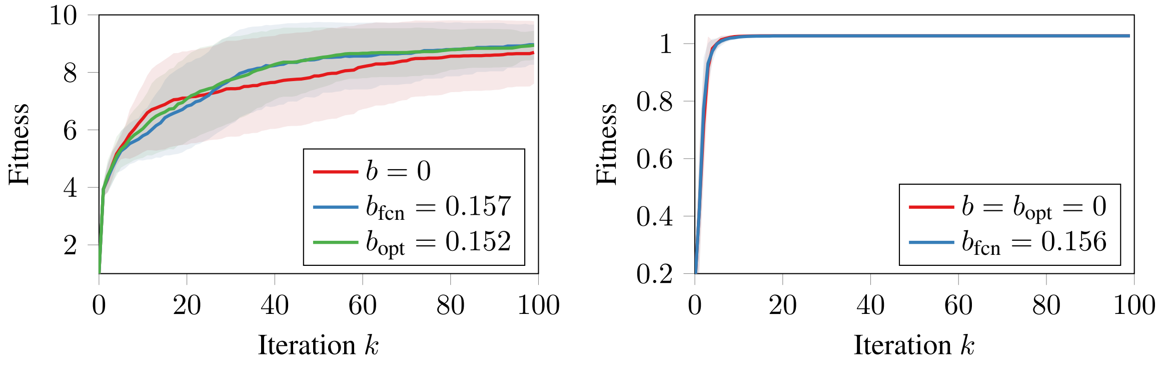
The first empirical fitness landscape the ruggedness estimator is tested on is that of a four amino acid region of GB1 (20^4 combinations) (Wu et al., 2016). GB1 is an antibody-binding protein isolated from Streptococcal bacteria. The fitness values of this landscape correspond to how well each GB1 variant binds to the antibody. The second empirical landscape tested on is that of a three amino acid region of an antitoxin protein known as ParD3 (20^3 combinations) (Lite et al., 2020). Fitness values in this landscape correspond to strength of binding to the toxin ParE2.

Ruggedness (K/N) estimation is performed on GB1 and ParD3 landscapes using a mutating population of size $n = 50$, mutating from the natural (wildtype) sequence over $G = 50$ generations, with a mutation rate $N\theta = 0.5$. As the process underlying the ruggedness estimation is stochastic, $\phi_{\text{fcn}}(u)$ is evaluated 100 times. The FCN $\phi_{\text{fcn}}(u)$ estimated K/N values of 0.5 ± 0.11 and 0.28 ± 0.04 on GB1 and ParD3, respectively (see Fig. 2b). The estimated K/N values are combined with N to look up an estimate for optimal base chance. Here, N corresponds to the length of the DNA sequence for the protein mutating region, so that $N = 12$ for GB1 and $N = 9$ for ParD3. The inferred base chance values are $b_{\text{fcn}} = 0.157$ on GB1, and $b_{\text{fcn}} = 0.156$ on ParD3.

Finally, directed evolution simulations are performed on GB1 and ParD3 using the selected base chance values and compared in Fig. 4. In each case, the simulation is compared to the standard approach to directed evolution, which is to select only the top fraction of variants each generation ($b = 0$), as well as to the "optimal" base chance b_{opt} obtained in the same way Fig. 2b was obtained. On GB1, a $5.6\,\%$ improvement in fitness after 100 generations is observed. As a landscape that is predicted to be more rugged, it is proposed that this improvement is due to the increase in b, reducing the propensity to get trapped in local optima. On closer inspection, this was confirmed as the mean fitness of the new strategy corresponds to between the highest and second highest peak on the landscape, whereas the mean fitness of the standard approach corresponds to between the second and third highest peak on the landscape. On ParD3, all strategies achieved the global maximum (1.023) within $\sim$10 generations. This supports the estimated lower K/N value (0.28), which suggests that ParD3 is a less rugged landscape and thus easier to navigate. In each case, the strategy that uses b_{fcn} either matches or out-performs the standard approach, and matches the strategy that uses b_{opt}.

5. Conclusion

In this paper, it was shown that the ruggedness of protein fitness landscapes can be estimated from measured average fitness values, μ_k. The estimated ruggedness parameter can be used to select parameters for control of directed evolution. To estimate the ruggedness parameter, an FCN was trained on a range of NK landscapes with known ruggedness parameters. The performance of the FCN-selected parameters were tested on the GB1 and ParD3 empirical landscapes, and compared to a fixed parameter (standard) approach to directed evolution. It was shown that the proposed method leads to improvements on the more rugged GB1 landscape, and matches the already high performance of the standard approach on ParD3. In the absence of prior knowledge regarding the shape of a fitness (or other) landscape, the proposed method allows one to determine a tailored strategy that improves the likelihood of a desired outcome, contrasting fixed-parameter approaches that are prone to getting trapped in local optima.



(a) GB1.

(b) ParD3.

Figure 4: Mean fitness and standard deviation for simulated directed evolution experiments with population size $n = 300$ averaged over 100 different starting points.

The assumption underlying the ruggedness estimation was that μ_0 significantly differs from the overall mean fitness, $\bar{F}$, allowing the features to capture a trend as $\mu_k \to \bar{F}$. When $\mu_0 \approx \bar{F}$, the features observed in smooth and rugged landscapes do not differ, in which case the FCN performs worse. Although in practice it is unlikely that $\mu_0 \approx \bar{F}$ for larger N, future research could address this limitation, e.g., by adding higher moments of the fitness distribution to the features.

We believe that the ruggedness estimation can be improved by increasing the quality of training data, e.g., with a fitness landscape model more tailored to the case of protein evolution than the NK landscapes. Future research could develop a theoretical fitness landscape model that incorporates the specific properties of proteins and their evolution. This method could be validated in real-world directed evolution experiments, which commonly feature much larger values of N than the empirical landscapes used in this paper. Additionally, future research could combine the ruggedness estimation with the base chance look-up to obtain an end-to-end solution for parameter selection.

Although this method has been applied to the literal case of directed evolution, it could also apply to other non-linear, non-convex optimisation problems, e.g., for controlling the parameters of genetic algorithms.

Appendix A.

In order to prove Proposition 1, several intermediary definitions and results are provided. First, the random process corresponding to a mutating gene is defined. Note that the more general continuous time setting is used in the following.

Definition 2 (Randomly drifting locus) *Let $X_t^i \in A$ be a random process. X_t^i is referred to as a randomly drifting locus with A alleles, and mutation rate α if it is a continuous-time Markov chain, with an transition rate of $\frac{\alpha}{A-1}$ between any two non-identical states.*

Lemma 3 *If X_t^i is a randomly drifting locus with A alleles, and mutation rate α, then, $\mathbb{P}(X_0^i = X_t^i) = \frac{1}{A} + (1 - \frac{1}{A})e^{-\alpha(\frac{A}{A-1})t}$.*

Proof All alleles not equal to X_0^i are symmetric. Hence we may consider them as a single collective state, and by symmetry the chance that X_t^i will be any specific allele given $X_t^i \neq X_0^i$ will be $\frac{1}{A-1}$ by symmetry. The result from Lemma 3 may be obtained from solving the following differential equation:

$$\frac{d\mathbb{P}(X_0^i = X_t^i)}{dt} = -\alpha \mathbb{P}(X_0^i = X_t^i) + \frac{\alpha}{A-1}(1 - \mathbb{P}(X_0^i = X_t^i)).$$

■

The following definition extends Definition 2 to genotype level.

Definition 4 (Randomly drifting gene) *Let $X_t = \begin{bmatrix} X_t^1 & X_t^2 & \dots & X_t^N \end{bmatrix} \in A^N$ be a random process. If all of the processes X_t^i are statistically independent, randomly drifting loci with A alleles and mutation rate α, then X_t is referred to as a randomly drifting gene on N loci, A alleles, and mutation rate α. This may be written shorthand as $X_t \sim \mathcal{D}(N, A, \alpha)$*

Lemma 5 *Any subset of the loci of a randomly drifting gene is also a randomly drifting gene.*

Proof This follows from Definition 4. ■

Lemma 6 *If $X_t \sim \mathcal{D}(N, A, \alpha)$, then $\mathbb{P}(X_t = X_0) \approx e^{-\alpha N t}$.*

Proof $\mathbb{P}(X_t = X_0) = \mathbb{P}(X_t^0 = X_0^0)^N$, by independence of the loci. Using Lemma 3 and approximating $e^x \approx 1 + x$, we obtain:

$$\mathbb{P}(X_t^0 = X_0^0) = \frac{1}{A} + (1 - \frac{1}{A})e^{-\alpha(\frac{A}{A-1})t} \approx \frac{1}{A} + (1 - \frac{1}{A})(1 - \alpha(\frac{A}{A-1})t) = 1 - \alpha t$$

Again approximating $e^x \approx 1 + x$ yields the desired result for $\mathbb{P}(X_t = X_0)$ as

$$\mathbb{P}(X_t = X_0) = \mathbb{P}(X_t^0 = X_0^0)^N \approx (1 - \alpha t)^N \approx e^{-\alpha N t}.$$

■

Next, the conditions that the fitness landscapes must satisfy for Proposition 1 to hold are formally defined.

Definition 7 (K-loci landscape) *We say $F : A^N \to \mathbb{R}$ is a K-loci landscape if $\exists \Phi^\nu : A^K \to \mathbb{R} \forall \nu \in \mathcal{P}_K(N)$ such that $\forall X = \begin{bmatrix} X^1 & X^2 & \dots & X^N \end{bmatrix} \in A^N$:*

$$F(X) = \sum_{\nu \in \mathcal{P}_K(N)} \Phi^\nu(X[\nu])$$

where $\mathcal{P}_K(N) = \{\nu \subseteq \{1 \dots N\} \mid K = |\nu|\}$, and $X[\nu] \in A^K$ is shorthand for $\begin{bmatrix} X^{\nu_i} & X^{\nu_i} & \dots & X^{\nu_K} \end{bmatrix}$, where ν_i is shorthand for element i of ν (with the standard ordering).

Definition 8 (Isotropic K-loci landscape) *Let $F : A^N \to \mathbb{R}$ be a distribution over K-loci landscapes. We say F is isotropic if, $\forall \nu^1, \nu^2 \in \mathcal{P}_K(N)$ and $\forall X_1, X_2 \in A^N$:*

1. *$\Phi^{\nu_1}(X_1[\nu_1])$ and $\Phi^{\nu_2}(X_2[\nu_2])$ are independent unless $\nu_1 = \nu_2$ and $X_1[\nu_1] = X_2[\nu_2]$.*
2. *$\Phi^{\nu_1}(X_1[\nu_1])$ and $\Phi^{\nu_1}(X_2[\nu_1])$ have the same distribution.*

Furthermore, we write $\bar{\Phi}^{\nu_1} = \mathbb{E}\left[\Phi^{\nu_1}(X_1[\nu_1])\right] = \mathbb{E}\left[\Phi^{\nu_1}(X_2[\nu_1])\right]$, and $\bar{F} = \mathbb{E}\left[F(X_1)\right] = \mathbb{E}\left[F(X_2)\right] = \sum_{\nu\in\mathcal{P}_K(N)} \bar{\Phi}^{\nu}$

Lemma 9 *All NK landscapes are isotropic K-loci landscapes.*

Proof We provide a proof by construction. Let $\kappa = \kappa_1, \kappa_2 \ldots \kappa_N$ be the N interaction loci in an NK landscape. Write $\mathcal{U}(0, 1, k)$ for the Irwin–Hall distribution, which is the sum of k independent standard uniform distributions. Let $k_\nu = |\{i \mid \kappa_i = \nu\}|$, for $X \in A^N$, we may set $\Phi^\nu(X[\nu]) \sim \mathcal{U}(0, 1, k_\nu)$. This induces an NK landscape, whilst also trivially satisfying both requirements for an isotropic K-loci landscape. ■

Finally, the main theoretical result is provided:

Proposition 10 *Suppose $X_t \sim \mathcal{D}(N, A, \alpha)$, and $F : A^N \to \mathbb{R}$ is an Isotropic K-loci landscape. Then, $\mathbb{E}\left[F(X_t) \mid F(X_0)\right] \approx e^{-\alpha K t} F_0 + (1 - e^{-\alpha K t})\bar{F}$.*

Proof We write $F_t = F(X_t)$ as shorthand. Using Definition 8, we obtain:

$$\mathbb{E}\left[F_t \mid F_0\right] = \mathbb{E}\left[\sum_{\nu\in\mathcal{P}_K(N)} \Phi^\nu(X_t[\nu]) \middle| F_0\right] = \sum_{\nu\in\mathcal{P}_K(N)} \mathbb{E}\left[\Phi^\nu(X_t[\nu]) \mid F_0\right]. \tag{4}$$

Focusing on a single ν, by Lemma 5, $X_t[\nu] \sim \mathcal{D}(K, A, \alpha)$, allowing Lemma 6 to be used:

$$\begin{aligned}\mathbb{E}\left[\Phi^\nu(X_t[\nu]) \mid F_0\right] &= \mathbb{E}\left[\Phi^\nu(X_t[\nu]) \mid F_0 \cap X_t[\nu] = X_0[\nu]\right] \mathbb{P}(X_t[\nu] = X_0[\nu]) \\ &\quad + \mathbb{E}\left[\Phi^\nu(X_t[\nu]) \mid F_0 \cap X_t[\nu] \neq X_0[\nu]\right] \mathbb{P}(X_t[\nu] \neq X_0[\nu]) \\ &\approx \mathbb{E}\left[\Phi^\nu(X_0[\nu]) \mid F_0\right] e^{-\alpha K t} + \bar{\Phi}^\nu(1 - e^{-\alpha K t})\end{aligned} \tag{5}$$

Substituting (5) into the right-hand side of (4) yields

$$\begin{aligned}\mathbb{E}\left[F_t | F_0\right] &\approx \sum_{\nu\in\mathcal{P}_K(N)} \mathbb{E}\left[\Phi^\nu(X_0[\nu]) \mid F_0\right] e^{-\alpha K t} + \bar{\Phi}^\nu(1 - e^{-\alpha K t}), \\ &= e^{-\alpha K t}\mathbb{E}\left[\sum_{\nu\in\mathcal{P}_K(N)} \Phi^\nu(X_0[\nu]) \middle| F_0\right] + (1 - e^{-\alpha K t}) \sum_{\nu\in\mathcal{P}_K(N)} \bar{\Phi}^\nu, \\ &= e^{-\alpha K t} F_0 + (1 - e^{-\alpha K t})\bar{F},\end{aligned} \tag{6}$$

which proves Proposition 10. Proposition 1 follows from Lemma 9 and discretising (6). ■

Acknowledgments

I.K. acknowledges support from EPSRC (EP/X017982/1).

References

I. Anishchenko et al. De novo protein design by deep network hallucination. *Nature*, 600(7889): 547–552, December 2021. doi: 10.1038/s41586-021-04184-w.

F. H. Arnold. Design by Directed Evolution. *Accounts of Chemical Research*, 31(3):125–131, March 1998. doi: 10.1021/ar960017f.

G. Avecilla et al. Neural networks enable efficient and accurate simulation-based inference of evolutionary parameters from adaptation dynamics. *PLOS Biology*, 20(5):e3001633, May 2022. doi: https://doi.org/10.1371/journal.pbio.3001633.

M. Baek et al. Accurate prediction of protein structures and interactions using a three-track neural network. *Science*, 373(6557):871–876, August 2021. doi: 10.1126/science.abj8754.

C. Bank et al. On the (un)predictability of a large intragenic fitness landscape. *Proceedings of the National Academy of Sciences*, 113(49):14085–14090, December 2016. doi: 10.1073/pnas.1612676113.

A. C. Carpenter et al. Have you tried turning it off and on again? Oscillating selection to enhance fitness-landscape traversal in adaptive laboratory evolution experiments. *Metabolic Engineering Communications*, 17:e00227, December 2023. doi: 10.1016/j.mec.2023.e00227.

H.-H. Chou et al. Mapping the fitness landscape of gene expression uncovers the cause of antagonism and sign epistasis between adaptive mutations. *PLOS Genetics*, 10(2):e1004149, February 2014. doi: 10.1371/journal.pgen.1004149.

J. Dauparas et al. Robust deep learning–based protein sequence design using ProteinMPNN. *Science*, 378(6615):49–56, October 2022. doi: 10.1126/science.add2187.

S. D'Costa et al. Inferring protein fitness landscapes from laboratory evolution experiments. *PLOS Computational Biology*, 19(3):e1010956, March 2023. doi: https://doi.org/10.1371/journal.pcbi.1010956.

N. Ferruz, S. Schmidt, and B. Höcker. ProtGPT2 is a deep unsupervised language model for protein design. *Nature Communications*, 13(1):4348, July 2022. doi: 10.1038/s41467-022-32007-7.

R. A. Fisher. The genetical theory of natural selection. *Journal of the Royal Statistical Society*, 33: 94–100, October 1931. doi: 10.2307/2341823.

R. Fox et al. Optimizing the search algorithm for protein engineering by directed evolution. *Protein Engineering, Design and Selection*, 16(8):589–597, August 2003. doi: 10.1093/protein/gzg077.

T. S. Frisby and C. J. Langmead. Bayesian optimization with evolutionary and structure-based regularization for directed protein evolution. *Algorithms for Molecular Biology*, 16(1):13, July 2021. doi: 10.1186/s13015-021-00195-4.

S. Gavrilets. Evolution and speciation on holey adaptive landscapes. *Trends in Ecology & Evolution*, 12(8):307–312, August 1997. ISSN 0169-5347. doi: 10.1016/S0169-5347(97)01098-7.

T. F. Hansen and G. P. Wagner. Modeling genetic architecture: a multilinear theory of gene interaction. *Theoretical Population Biology*, 59(1):61–86, February 2001. doi: 10.1006/tpbi.2000.1508.

B. S. Heater et al. Directed evolution of a genetically encoded immobilized lipase for the efficient production of biodiesel from waste cooking oil. *Biotechnology for Biofuels*, 12(1):165, June 2019. doi: 10.1186/s13068-019-1509-5.

R. Hu et al. Protein engineering via Bayesian optimization-guided evolutionary algorithm and robotic experiments. *Briefings in Bioinformatics*, 24(1), January 2023. doi: https://doi.org/10.1093/bib/bbac570.

J. James et al. Optimision strategies for directed evolution without sequencing. *bioRxiv*, December 2023, in press.

J. Jumper et al. Highly accurate protein structure prediction with AlphaFold. *Nature*, 596(7873): 583–589, August 2021. doi: 10.1038/s41586-021-03819-2.

S. Kauffman and S. Levin. Towards a general theory of adaptive walks on rugged landscapes. *Journal of Theoretical Biology*, 128(1):11–45, September 1987. doi: 10.1016/S0022-5193(87)80029-2.

S. Kauffman and E. D. Weinberger. The NK model of rugged fitness landscapes and its application to maturation of the immune response. *Journal of Theoretical Biology*, 141(2):211–245, November 1989. doi: 10.1016/s0022-5193(89)80019-0.

A. I. Khan et al. Negative epistasis between beneficial mutations in an evolving bacterial population. *Science*, 332(6034):1193–1196, June 2011. doi: 10.1126/science.1203801.

T. V. Lite et al. Uncovering the basis of protein-protein interaction specificity with a combinatorially complete library. *eLife*, 9:e60924, October 2020. doi: 10.7554/eLife.60924.

R. S. Molina et al. In vivo hypermutation and continuous evolution. *Nature Reviews Methods Primers*, 2(1):1–22, May 2022. doi: 10.1038/s43586-022-00119-5.

J. Neidhart et al. Adaptation in tunably rugged fitness landscapes: The rough mount fuji model. *Genetics*, 198(2):699–721, October 2014. doi: 10.1534/genetics.114.167668.

A. E. Nixon, D. J. Sexton, and R. C. Ladner. Drugs derived from phage display. *mAbs*, 6(1):73–85, January 2014. doi: 10.4161/mabs.27240.

A. Papkou et al. A rugged yet easily navigable fitness landscape. *Science*, 382(6673):eadh3860, November 2023. doi: 10.1126/science.adh3860.

J. M. Singer et al. Large-scale design and refinement of stable proteins using sequence-only models. *PLOS ONE*, 17(3):e0265020, March 2022. doi: 10.1371/journal.pone.0265020.

O. Tenaillon. The utility of Fisher's geometric model in evolutionary genetics. *Annual review of ecology, evolution, and systematics*, 45:179–201, November 2014. doi: 10.1146/annurev-ecolsys-120213-091846.

D. M. Weinreich et al. Darwinian evolution can follow only very few mutational paths to fitter proteins. *Science*, 312(5770):111–114, April 2006. doi: 10.1126/science.1123539.

B. I. M. Wicky et al. Hallucinating symmetric protein assemblies. *Science*, 378(6615):56–61, October 2022. doi: 10.1126/science.add1964.

B. J. Wittmann, Y. Yue, and F. H. Arnold. Informed training set design enables efficient machine learning-assisted directed protein evolution. *Cell Systems*, 12(11):1026–1045.e7, November 2021. doi: 10.1016/j.cels.2021.07.008.

A. Wright, R. K. Thompson, and J. Zhang. The computational complexity of n-k fitness functions. 4:373–379, 2000. doi: 10.1109/4235.887236.

S. Wright. The Roles of Mutation, Inbreeding, crossbreeding and Selection in Evolution. *Proceedings of the XI International Congress of Genetics*, 8:209–222, 1932.

N. C. Wu et al. Adaptation in protein fitness landscapes is facilitated by indirect paths. *eLife*, 5: e16965, July 2016. doi: 10.7554/eLife.16965.

Z. Wu et al. Machine learning-assisted directed protein evolution with combinatorial libraries. *Proceedings of the National Academy of Sciences*, 116(18):8852–8858, April 2019. doi: 10.1073/pnas.1901979116.

K. K. Yang, Z. Wu, and F. H. Arnold. Machine-learning-guided directed evolution for protein engineering. *Nature Methods*, 16(8):687–694, August 2019. doi: 10.1038/s41592-019-0496-6.