Machine learning-assisted directed protein evolution with combinatorial libraries

Zachary Wu[a], S. B. Jennifer Kan[a], Russell D. Lewis[b], Bruce J. Wittmann[b], and Frances H. Arnold[a,b,1]

[a]Division of Chemistry and Chemical Engineering, California Institute of Technology, Pasadena, CA 91125; and [b]Division of Biology and Bioengineering, California Institute of Technology, Pasadena, CA 91125

Contributed by Frances H. Arnold, March 18, 2019 (sent for review February 4, 2019; reviewed by Marc Ostermeier and Justin B. Siegel)

To reduce experimental effort associated with directed protein evolution and to explore the sequence space encoded by mutating multiple positions simultaneously, we incorporate machine learning into the directed evolution workflow. Combinatorial sequence space can be quite expensive to sample experimentally, but machine-learning models trained on tested variants provide a fast method for testing sequence space computationally. We validated this approach on a large published empirical fitness landscape for human GB1 binding protein, demonstrating that machine learning-guided directed evolution finds variants with higher fitness than those found by other directed evolution approaches. We then provide an example application in evolving an enzyme to produce each of the two possible product enantiomers (i.e., stereodivergence) of a new-to-nature carbene Si–H insertion reaction. The approach predicted libraries enriched in functional enzymes and fixed seven mutations in two rounds of evolution to identify variants for selective catalysis with 93% and 79% *ee* (enantiomeric excess). By greatly increasing throughput with in silico modeling, machine learning enhances the quality and diversity of sequence solutions for a protein engineering problem.

protein engineering | machine learning | directed evolution | enzyme | catalysis

Nature provides countless proteins with untapped potential for technological applications. Rarely optimal for their envisioned human uses, nature's proteins benefit from sequence engineering to enhance performance. Successful engineering is no small feat, however, as protein function is determined by a highly tuned and dynamic ensemble of states (1). In some cases, engineering to enhance desirable features can be accomplished reliably by directed evolution, in which beneficial mutations are identified and accumulated through an iterative process of mutation and testing of hundreds to thousands of variants in each generation (2–4). However, implementing a suitable screen or selection can represent a significant experimental burden.

Given that screening is the bottleneck and most resource-intensive step for the majority of directed evolution efforts, devising ways to screen protein variants in silico is highly attractive. Molecular-dynamics simulations, which predict dynamic structural changes for protein variants, have been used to predict changes in structure (5) and protein properties caused by mutations (6). However, full simulations are also resource-intensive, requiring hundreds of processor hours for each variant, a mechanistic understanding of the reaction at hand, and, ideally, a reference protein structure. A number of other less computationally intensive physical models have also been used to identify sequences likely to retain fold and function for further experimental screening (7–9).

An emerging alternative for screening protein function in silico is machine learning, which comprises a set of algorithms that make decisions based on data (10). By building models directly from data, machine learning has proven to be a powerful, efficient, and versatile tool for a variety of applications, such as extracting abstract concepts from text and images or beating humans at our most complex games (11, 12). Previous applications of machine learning in protein engineering have identified beneficial mutations (13) and optimal combinations of protein fragments (14) for increased enzyme activity and protein stability, as reviewed recently (15). Here we use machine learning to enhance directed evolution by using combinatorial libraries of mutations to explore sequence space more efficiently than conventional directed evolution with single mutation walks. The size of a mutant library grows exponentially with the number of residues considered for mutation and quickly becomes intractable for experimental screening. However, by leveraging in silico models built based on sampling of a combinatorial library, machine learning assists directed evolution to make multiple mutations simultaneously and traverse fitness landscapes more efficiently.

In the machine learning-assisted directed evolution strategy presented here, multiple amino acid residues are randomized in each generation. Sequence–function information sampled from the large combinatorial library is then used to predict a restricted library with an increased probability of containing variants with high fitness. The best-performing variants from the predicted libraries are chosen as the starting points for the next round of evolution, from which further improved variants are identified. We first investigate the benefits of in silico screening by machine learning using the dataset collected by Wu et al. (16), who studied the effects on antibody binding of mutations at four positions in

Significance

Proteins often function poorly when used outside their natural contexts; directed evolution can be used to engineer them to be more efficient in new roles. We propose that the expense of experimentally testing a large number of protein variants can be decreased and the outcome can be improved by incorporating machine learning with directed evolution. Simulations on an empirical fitness landscape demonstrate that the expected performance improvement is greater with this approach. Machine learning-assisted directed evolution from a single parent produced enzyme variants that selectively synthesize the enantiomeric products of a new-to-nature chemical transformation. By exploring multiple mutations simultaneously, machine learning efficiently navigates large regions of sequence space to identify improved proteins and also produces diverse solutions to engineering problems.

Author contributions: Z.W., S.B.J.K., R.D.L., and F.H.A. designed research; Z.W. and B.J.W. performed research; Z.W. contributed new reagents/analytic tools; Z.W., S.B.J.K., R.D.L., and B.J.W. analyzed data; and Z.W., S.B.J.K., R.D.L., B.J.W., and F.H.A. wrote the paper.

Reviewers: M.O., Johns Hopkins University; and J.B.S., UC Davis Health System.

The authors declare no conflict of interest.

Data deposition: The data reported in this paper have been deposited in the ProtaBank database, https://www.protabank.org, at https://www.protabank.org/study_analysis/mnqBQFjF3/.

[1]To whom correspondence should be addressed. Email: frances@cheme.caltech.edu.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1901979116/-/DCSupplemental.

Published online April 12, 2019.

human GB1 binding protein (theoretical library size, 20^4 = 160,000 variants). We then use machine learning-assisted directed evolution to engineer an enzyme for stereodivergent carbon–silicon bond formation, a new-to-nature chemical transformation.

Results

Directed Evolution and Machine Learning. In directed evolution, a library of variants is constructed from parental sequences and screened for desired properties, and the best variants are used to parent the next round of evolution; all other variants are discarded. When machine learning assists directed evolution, sequences and screening data from all of the variants can be used to train a panel of models (covering linear, kernel, neural network, and ensemble methods; *SI Appendix*, *Model Training*). The models with the highest accuracy are then used to screen variants in a round of in silico evolution, whereby the models simulate the fitnesses of all possible sequences and rank the sequences by fitness. A restricted library containing the variants with the highest predicted fitnesses is then constructed and screened experimentally.

This work explores the full combinatorial space of mutations at multiple positions. Fig. 1 illustrates the approach considering a set of four mutated positions. In a conventional directed evolution experiment with sequential single mutations, the identification of optimal amino acids for N positions in a set requires N rounds of evolution (Fig. 1*A*). An alternative directed evolution approach is to randomly sample the combinatorial space and recombine the best mutations found at each position in a subsequent combinatorial library (Fig. 1*B*). Machine learning-assisted evolution samples the same combinatorial space with comutated positions in silico, enabling larger steps through sequence space in each round (Fig. 1*C*). In this approach, data from a random sample of the combinatorial library, the input library, are used to train machine learning models. These models are used to predict a smaller set of variants, the predicted library, which can be encoded with degenerate codons to test experimentally (17). The best-performing variant is then used as the parent sequence for the next round of evolution with mutations at new positions.

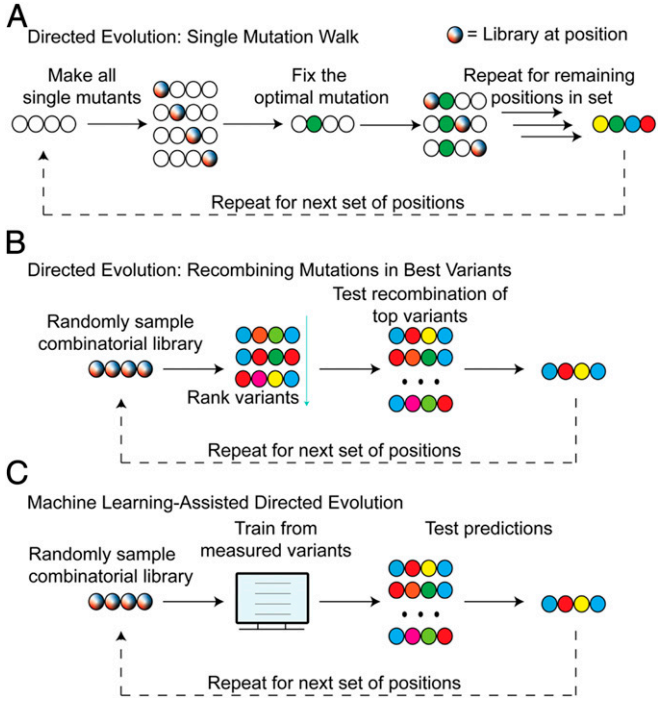


Fig. 1. (*A*) Directed evolution with single mutations. If limited to single mutations, the identification of optimal amino acids for N positions requires N rounds of evolution. (*B*) Directed evolution by recombining mutations found in best variants from a random combinatorial search. (*C*) Machine learning-assisted directed evolution. As a result of increased throughput provided by screening in silico, four positions can be explored simultaneously in a single round, enabling a broader search of sequence–function relationships and deeper exploration of epistatic interactions.

Validation on an Empirical Fitness Landscape. We first investigated this machine learning-assisted approach on the large empirical fitness landscape of Wu et al. (16), who studied protein G domain B1 (GB1) binding to an antibody. Specifically, we compare the final fitnesses reached by simulated directed evolution with and without machine learning based on testing the same number of variants. The empirical landscape used here consists of measurements of 149,361 of a total of 160,000 (20^4) variants from NNK/NNS saturation mutagenesis at four positions known to interact epistatically. The fitness of protein GB1 was defined as the enrichment of folded protein bound to the antibody IgG-Fc measured by coupling mRNA display with next-generation sequencing. The landscape contains a fitness maximum at 8.76, with a fitness value of 1 set for the parent sequence and 19.7% of variants at a reported value of 0. On this landscape, the simulated single-mutant walk (described later) reached 869 fitness peaks, 533 of which outperformed the wild-type (WT) sequence and 138 of which had fitness less than 5% of the WT fitness. A full description of the epistatic landscape is provided in the thorough analysis of Wu et al. (16).

We first simulated single-mutation evolutionary walks starting from each of the 149,361 variants reported. The algorithm proceeded as follows: in each single-mutation walk, all possible single amino acid mutations were tested at each of the four mutated positions. The best amino acid was then fixed at its observed position, and that position was restricted from further exploration. This process continued iteratively with the remaining positions until an amino acid was fixed at each position. As a greedy search algorithm that always follows the path with strongest improvements in fitness, this single-mutation walk has a deterministic solution for each starting variant. Assuming each amino acid occurs with equal frequency and that the library has complete coverage, applying the threefold oversampling rule to obtain approximately 95% library coverage (18, 19) results in a total of 570 variants screened (*SI Appendix*, *Library Coverage*).

Another technique widely used in directed evolution is recombination. For a given set of positions to explore, one method is to randomly sample the combinatorial library and recombine the mutations found at each position in the top M variants. This process is shown in Fig. 1*B*. For N positions, the recombinatorial library then has a maximum of M^N variants, and we selected the top three variants, for a maximum recombinatorial library size of 81. An alternative recombination approach is to test all possible single mutants from a given parent sequence and recombine the top three mutations at each position, for a fixed recombinatorial library size of 81. However, this alternative recombination does not perform as well on the GB1 data set (*SI Appendix*, Fig. S1*B*). Compared with these recombination strategies, the machine learning-assisted approach has the distinct advantage of providing estimates for the variability at each position (as opposed to taking the top three mutations at each).

To compare the distribution of fitness values of the optimal variants found by the described directed evolution methods, shallow neural networks were trained with 470 randomly selected input variants, from which 100 predictions were tested, for a total screening burden equivalent to the single-mutation walk. Although the number of variants tested was determined by comparison with another method (a single-mutant walk) and the ratio of training variants vs. predicted variants was set through

experimental convenience (the size of a deep-well plate), from a modeling perspective, these design choices could be improved to increase the expected fitness improvement (*SI Appendix*, Fig. S1*A*). Histograms of the highest fitnesses found by these approaches are shown in Fig. 2*A* and reiterated as empirical cumulative distribution functions in Fig. 2*B*.

As shown in Fig. 2, with the same number of variants screened, machine learning-assisted evolution reaches the global optimum fitness value in 8.2% of 600 simulations, compared with 4.9% of all starting sequences reaching the same value through a single-mutant walk and 4.0% of simulated recombination runs. Additionally, on this landscape, the machine-learning approach requires approximately 30% fewer variants to achieve final results similar to the single-mutant walk with this analysis. Perhaps more importantly, a single-mutant walk is much more likely to end at low fitness levels compared with approaches that sample the combinatorial library directly. To this end, the machine learning approach has an expected fitness value of 6.42, compared with 5.41 and 5.93 for the single-step walk and recombination, respectively.

Interestingly, the accuracy of the machine learning models as determined on a test set of 1,000 random variants not found in the training set can be quite low (Pearson's $r = 0.41$, SD = 0.17). However, this level of accuracy as measured by Pearson's r appears to be sufficient to guide evolution. Although perfect accuracy does not seem to be necessary, if the accuracy of the models is so low that predictions are random guesses, this approach cannot be expected to outperform a single-mutant walk (*SI Appendix*, Fig. S1*A*). As an algorithm, evolution is focused on identifying optimal variants, and the development of a measure of model accuracy biased toward correctly identifying optimal variants will likely improve model selection. This validation experiment gave us confidence that machine learning-assisted directed evolution can find improved protein variants efficiently.

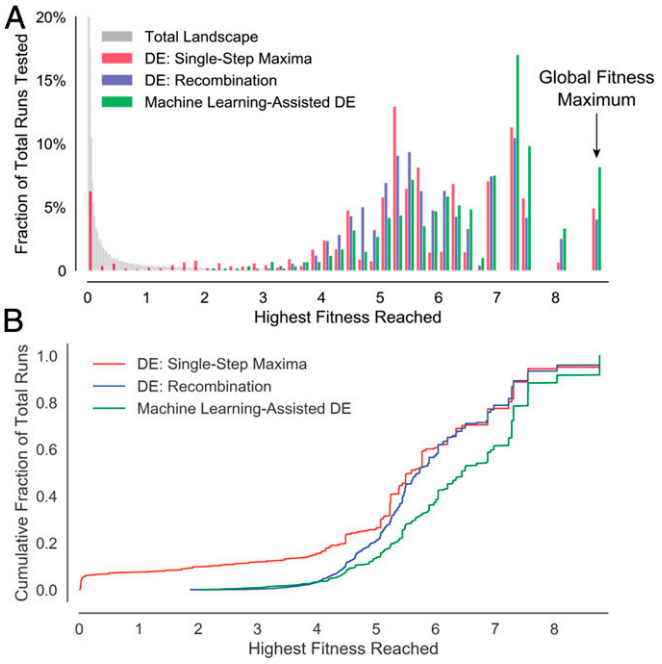


Fig. 2. (*A*) Highest fitness values found by directed evolution and directed evolution assisted by machine learning. The distribution of fitness peaks found by iterative site-saturation mutagenesis from all labeled variants (149,361 of 20^4 possible covering four residues) is shown in red. The distribution of fitness peaks found by 10,000 recombination runs with an average of 570 variants tested is shown in blue. The distribution of the highest fitnesses found from 600 runs of the machine learning-assisted approach is shown in green. A total of 570 variants are tested in all approaches. For reference, the distribution of all measured fitness values in the landscape is shown in gray. (*B*) The same evolutionary distributions are shown as empirical cumulative distribution functions, where the ordinate at any specified fitness value is the fraction of evolutionary runs that reach a fitness less than or equal to that specified value. Machine learning-assisted evolution walks are more likely to reach higher fitness levels compared with conventional directed evolution.

Application to Evolution of Enantiodivergent Enzyme Activity. We next used machine learning-assisted directed evolution to engineer an enzyme to produce each of two possible product enantiomers. For this demonstration, we selected the reaction of phenyldimethyl silane with ethyl 2-diazopropanoate (Me-EDA) catalyzed by a putative nitric oxide dioxygenase (NOD) from *Rhodothermus marinus* (*Rma*), as shown in Fig. 3. Carbon–silicon bond formation is a new-to-nature enzyme reaction (20), and *Rma* NOD with mutations Y32K and V97L catalyzes this reaction with 76% *ee* (enantiomeric excess) for the (*S*)-enantiomer in whole-cell reactions (*SI Appendix*, Table S1).

Silicon has potential for tuning the pharmaceutical properties of bioactive molecules (21, 22). Because enantiomers of bioactive molecules can have stark differences in their biological effects (23), access to both is important (24). Screening for enantioselectivity, however, typically requires long chiral separations to discover beneficial mutations in a low-throughput screen (25). We thus tested whether machine learning-assisted directed evolution can efficiently generate two catalysts to make each of the product (*S*)- and (*R*)-enantiomers starting from a single parent sequence.

We chose the parent *Rma* NOD (UniProt ID D0MGT2) (26) enzyme for two reasons. First, *Rma* NOD is native to a hyperthermophile and should be thermostable. Because machine learning-assisted directed evolution makes multiple mutations per iteration, a starting sequence capable of accommodating multiple potentially destabilizing mutations is ideal (27). Second, although we previously engineered a cytochrome *c* (*Rma* cyt *c*) to >99% *ee* for the (*R*)-enantiomer, WT *Rma* cyt *c* serendipitously started with 97% *ee* (20). We hypothesized that a parent enzyme with less enantiopreference [76% *ee* for the (*S*)-enantiomer in whole cells] would be a better starting point for engineering enantiodivergent variants.

During evolution for enantioselectivity, we sampled two sets of amino acid positions: set I contained mutations to residues K32, F46, L56, and L97; and set II contained mutations to residues P49, R51, and I53 after fixing beneficial mutations identified from set I. For both sets, we first tested and sequenced an initial set of randomly selected mutants (i.e., the input library) to train models. We next tested a restricted set of mutants predicted to have high selectivity (i.e., the predicted library). The targeted positions are shown in a structural homology model in Fig. 4*A*. Set I positions were selected based on proximity to the putative active site, whereas set II positions were selected based on their proximity to the putative substrate entry channel.

Machine learning models are more useful when trained with data broadly distributed across input space, even if those data are noisy (28). When designing a training set for machine learning-assisted directed evolution, it is thus important to maximize the input sequence diversity by avoiding disproportionate amino acid representation (e.g., from codon usage). We therefore used NDT codons for the input libraries. NDT libraries encode 12 amino acids having diverse properties with 12 unique codons (18), thus minimizing the probability that an amino acid is overrepresented in the initial training set (29). Notably, the parent amino acid at a site is still considered by the model even if it is not encoded by the NDT codons, as sequence–function data are available for the parent sequence.

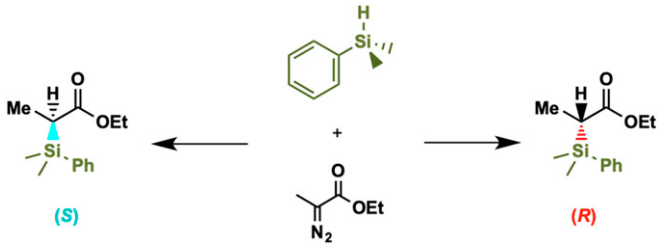


Fig. 3. Carbon–silicon bond formation catalyzed by heme-containing *Rma* NOD to form individual product enantiomers with high selectivity.

The evolution experiment is summarized in Fig. 4*B*. In the first round, *Rma* NOD Y32K V97L (76% *ee*) was used as a parent for NDT mutagenesis at the set I positions. From 124 sequence–function relationships sampled randomly, models were trained to predict a restricted set of selective variants. Specifically, a variety of models covering linear, kernel, shallow neural network, and ensemble methods were tested on each library, from which the optimum models were used to rank every sequence in the theoretical library by its predicted fitness. Under strict limitations in experimental throughput, and with one 96-well plate as the smallest batch size, we settled on two plates of input data for each round of evolution and one plate of tested predictions. However, increased throughput allows for increased likelihood of reaching the landscape's optimum (*SI Appendix*, Fig. S1*A*). The lower numbers of variants input in Fig. 4*C* compared with two full 96-well plates of sequencing reflect failed sequencing reads of these two plates.

From the predicted libraries for both enantiomers, two variants, called VCHV (86% *ee*) and GSSG (62% *ee*) for their amino acids at positions 32, 46, 56, and 97, were identified by screening 90 variants for each. VCHV and GSSG were then used as the parent sequences for the second round of mutagenesis at the three positions in set II. VCHV was the most selective variant in the initial screen, but was less selective in final validation. The approach of experimentally testing a library predicted by models trained on a randomly sampled input library was repeated. From those predicted libraries, we obtained two variants with measured enantioselectivities of 93% and 79% *ee* for the (*S*)- and (*R*)-enantiomers, respectively. These two enantioselective enzymes were achieved after obtaining 445 sequence–function relationships for model training and testing an additional 360 predicted variants, for a total of 805 variants tested experimentally covering seven positions, as summarized in Fig. 4*C*.

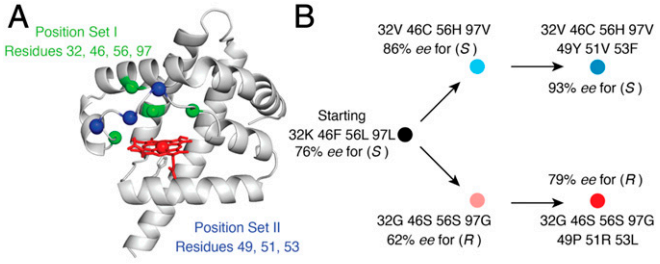
Machine Learning Identifies Diverse Improved Sequences. Comparison on the empirical protein GB1 dataset showed that machine learning-assisted directed evolution is more likely than directed evolution alone to identify improved variants. However, another benefit of this approach is the ability to identify a diverse set of sequences for accomplishing a specific task. Having diverse solutions is attractive, as some of those variants may satisfy other design requirements, such as increased total activity, altered substrate tolerance, specific amino acid handles for further protein modification, or sequence diversity for intellectual property considerations (30). By enabling exploration of the combinatorial space, machine learning-assisted directed evolution is able to identify multiple solutions for each engineering objective.

Tables 1 and 2 summarize the most selective variants in the input and predicted libraries for position sets I and II. The input library for set I is the same for both product enantiomers. The parent sequences for set II, VCHV and GSSG, are identified in the tables. The improvement in total activity measured in whole cells compared with the starting variant (32K, 46F, 56L, 97L) obtained after two rounds of machine learning-assisted directed evolution is also shown in Table 2. Although evolved for enantioselectivity, the variants have increased levels of (cellular) activity. Negative controls with cells expressing nonheme proteins yield a racemic mixture of product enantiomers as a result of a low level of nonselective background activity from free heme or heme proteins. Increasing the cellular activity of the *Rma* NOD protein can overcome this background activity and appears in the screen as improved selectivity if the protein is selective. Thus, enhanced activity is one path to higher selectivity. The two variants most selective for the (*S*)-enantiomer differ by less than 1% in *ee*. However, the 49P 51V 53I variant from VCHV has higher total activity under screening conditions. By providing multiple solutions in a combinatorial space for a single design criterion, machine learning is able to identify variants with other beneficial properties.

The solutions identified by this approach can also be nonobvious. For example, the three most (*S*)-selective variants in the initial input for position set I are YNL**L**, CSV**L**, and CVH**V**. The three most selective sequences from the restricted, predicted library are VGV**L**, CFN**L**, and VCH**V**. If only considering the last residue in bold, the predicted library can be sampled from the top variants in the input library. However, for each of the other three positions, there is at least one mutation that is not present in the top three input sequences.

Machine Learning Predicts Regions of Sequence Space Enriched in Function. Although the machine learning-assisted approach is more likely to reach sequences with higher fitness, as demonstrated in simulations using the human GB1 dataset, there may well be instances in which other evolution strategies serendipitously discover variants with higher fitness more quickly. Therefore, as the purpose of library creation is to increase the likelihood of success, we caution against focusing solely on examples of individual variants with higher fitness and propose an alternative analysis.

Sequence–fitness landscapes are typically represented with fitness values on the vertical axis, dependent on some ordering of the corresponding protein sequences. Representing this high-dimensional space, even when it is explored with single mutations, is complicated and requires the sequencing of each variant (31). However, in functional protein space, the engineer is primarily



C

	Variants Input	Fraction of Total Library	Leave-One-Out CV Score	Predictions Tested
Position Set I	124	0.60 %	0.549	90 for (R) 90 for (S)
Position Set II from VCHV	155	8.3 %	0.630	90
Position Set II from GSSG	166	8.9 %	0.605	90

Fig. 4. (*A*) Structural homology model of *Rma* NOD and positions of mutated residues made by SWISS-MODEL (47). Set I positions 32, 46, 56, and 97 are shown in red, and set II positions 49, 51, and 53 are shown in blue. (*B*) Evolutionary lineage of the two rounds of evolution. (*C*) Summary statistics for each round, including the number of sequences obtained to train each model, the fraction of the total library represented in the input variants, each model's leave-one-out cross-validation (CV) Pearson correlation, and the number of predicted sequences tested.

Table 1. Summary of the most (*S*)- and (*R*)-selective variants in the input and predicted libraries in set I (K32, F46, L56, L97)

Variant	Residue 32	46	56	97	Selectivity, % *ee* (enantiomer)
Input variants	Y	N	L	L	84 (*S*)
	C	S	V	L	83 (*S*)
	C	V	H	V	82 (*S*)
	C	R	S	G	56 (*R*)
	I	S	C	G	55 (*R*)
	N	V	R	I	47 (*R*)
Predicted variants	V	G	V	L	90 (*S*)
	C	F	N	L	90 (*S*)
	V*	C*	H*	V*	86 (*S*)
	G†	S†	S†	G†	62 (*R*)
	G	F	L	R	24 (*R*)
	H	C	S	R	17 (*R*)

*Parent sequence used for set II for (*S*)-selectivity.
†Parent sequence used for set II for (*R*)-selectivity.

concerned with fitness. Therefore, an alternative representation of a library is a 1D distribution of fitness values sampled at random for each encoded library. In other words, the sequences are disregarded for visualization, and the library is represented by the distribution of its fitness values. Each subplot in Fig. 5 shows the input and predicted (output) library as kernel density estimates in each round of evolution for (*R*)- and (*S*)-selectivity as fitness. This representation shows the main benefit of incorporating machine learning into directed evolution, which is the ability to focus expensive experiments on regions of sequence space enriched in desired variants.

A few things are immediately clear with this visualization. First, the distribution of random mutations made in the input libraries is shifted toward the parent in Fig. 5, as has been shown previously (32). In other words, random mutations made from an (*R*)-selective variant are more likely to be (*R*)-selective. More importantly, the machine-learning algorithm is able to focus its predictions on areas of sequence space that are enriched in high fitness, as can be seen in the shift in distribution from input to predicted libraries. Specifically, 90 predicted variants were tested for predicted library sizes of 864 for the (*S*)-enantiomer and 630 for the (*R*)-enantiomer in position set I. In position set II, the predicted library sizes were much smaller, at 192 and 90 variants for the (*S*)- and (*R*)-enantiomer, respectively. Ninety variants were tested for these predicted libraries, which were sequenced for redundancy (because of the smaller theoretical library size) to yield 47 and 39 unique variants. Thus, machine learning optimized directed evolution by sampling regions of sequence space dense in functionality. Notably, the machine learning algorithms appear to have more pronounced benefits in position set II, likely as a result of the smaller number of positions explored and larger number of sequence–function relationships obtained.

Discussion

We have shown that machine learning can be used to quickly screen a full recombination library in silico by using sequence–fitness relationships randomly sampled from the library. The predictions for the most-fit sequences are useful when incorporated into directed evolution. By sampling large regions of sequence space in silico to reduce in vitro screening efforts, we rapidly evolved a single parent enzyme to generate variants that selectively form both product enantiomers of a new-to-nature C–Si bond-forming reaction. Rather than relying on identifying beneficial single mutations as other methods such as ProSAR do (13), we modeled epistatic interactions at the mutated positions by sampling the combinatorial sequence space directly and incorporating models with nonlinear interactions.

Machine learning increases effective throughput by providing an efficient computational method for estimating desired properties of all possible proteins in a large library. Thus, we can take larger steps through sequence space by identifying combinations of beneficial mutations, circumventing the need for indirect paths (16) or alterations of the nature of selection (31), and potentially avoiding negative epistatic effects resulting from the accumulation of large numbers of mutations (33) that require reversion later in the evolution (34). This gives rise to protein sequences that would not be found just by recombining the best amino acids at each position. Allowing simultaneous incorporation of multiple mutations accelerates directed evolution by navigating different regions of the fitness landscape concurrently and avoiding scenarios in which the search for beneficial mutations ends in low-fitness regions of sequence space.

Importantly, machine learning-assisted directed evolution also results in solutions that appear quite distinct. For example, proline is conserved at residue 49 in two of the most (*S*)-selective variants. Proline is considered unique for the conformational rigidity it confers, and at first may seem structurally important, if not critical, for protein function. However, tyrosine and arginine are also tolerated at position 49 with less than 1% loss in enantioselectivity. This suggests that there are diverse solutions in protein space for specific properties, as has also recently been shown in protein design (8). Computational models make abstractions to efficiently model physical processes, and the level of abstraction must be tailored to the task, such as protein structure prediction (35). Although predictive accuracy could be improved by more computationally expensive simulations or by collecting more data for machine learning, improved variants can already be identified by sampling from a space predicted to be dense in higher-fitness variants. Nevertheless, full datasets collected with

Table 2. Summary of the most (*S*)- and (*R*)-selective variants in the input and predicted libraries in position set II (P49, R51, I53)

Variant	Residue 49	51	53	Selectivity, % *ee* (enantiomer)	Cellular activity increase over KFLL
Input variants					
From VCHV	P*	R*	I*	86 (*S*)	—
	Y	V	F	86 (*S*)	—
	N	D	V	75 (*S*)	—
From GSSG	P†	R†	I†	62 (*R*)	—
	N	S	Y	56 (*R*)	—
	N	I	I	55 (*R*)	—
Predicted variants					
From VCHV	Y	V	V	93 (*S*)	2.8-fold
	P	V	I	93 (*S*)	3.2-fold
	P	V	V	87 (*S*)	3.1-fold
From GSSG	P	R	L	79 (*R*)	2.2-fold
	P	G	L	75 (*R*)	2.1-fold
	P	F	F	70 (*R*)	2.2-fold

Mutations that improve selectivity for the (*S*)-enantiomer appear in the background of [32V, 46C, 56H, 97V (VCHV)] and for the (*R*)-enantiomer are in [32G, 46S, 56S, 97G (GSSG)]. Activity increase over the starting variant, 32K, 46F, 56L, 97L (KFLL), is shown for the final variants.
*Parent sequence used for set II for (*S*)-selectivity.
†Parent sequence used for set II for (*R*)-selectivity.

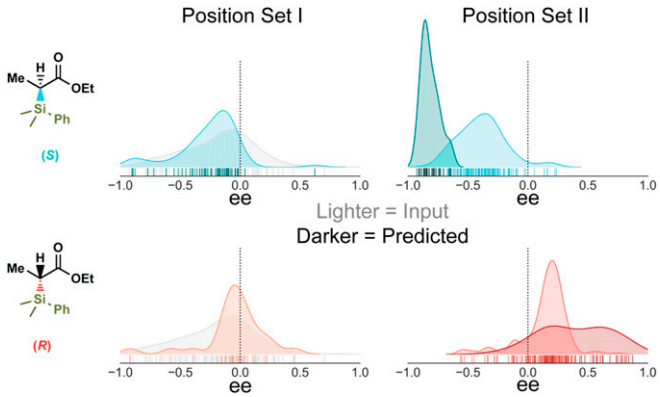


Fig. 5. A library's fitness values can be visualized as a 1D distribution, in this case as kernel density estimates over corresponding rug plots. This figure shows subplots for each library illustrating the changes between input (lighter) and predicted (darker) libraries for the (*S*)-enantiomers (cyan) and (*R*)-enantiomers (red). The initial input library for set I is shown in gray. The predicted (darker) libraries for each round are shifted toward the right and left of the distributions for the (*S*)- and (*R*)-enantiomers, respectively. For reference, dotted lines are shown for no enantiopreference (i.e., 0% *ee*).

higher-throughput methods such as deep mutational scanning (36) serve as valuable test beds for validating the latest machine-learning algorithms for regression (37, 38) and design (39) that require more data.

An evolution strategy similar in spirit to that described here was recently applied to the evolution of GFP fluorescence (40). However, the implementations are quite different. Saito et al. (40) used Gaussian processes to rank sequences based on their probability of improvement, or the probability that a variant outperforms those in the training set. We take a different approach of identifying the optimal variants, focusing efforts in the area of sequence space with highest fitness. Additionally, because it is difficult to know a priori which models will be most accurate for describing a particular landscape, we tested multiple types of models, from linear to ensemble models, to predict the optimal sequences. Modeling the effects of previously identified point mutations has also recently been studied for evolution of enantioselectivity of an enzyme (41). This study and others focused on increasing the accuracy of protein modeling by developing other physical descriptors (42, 43) or embedded representations (44), suggesting that machine learning will assist directed evolution beyond the baseline implementation employed here.

By providing an efficient estimate for desired properties, machine learning models are able to leverage the information from limited experimental resources to model proteins, without the need for a detailed understanding of how they function. Machine learning-assisted directed evolution with combinatorial libraries provides a tool for understanding the protein sequence–function relationship and for rapidly engineering useful proteins. Protein engineers have been sentenced to long treks through sequence space in the search for improved fitness. Machine learning can help guide us to the highest peaks.

Materials and Methods

Approach Validation on an Empirical Fitness Landscape. Fitness values were provided for 149,361 of 160,000 total possible sequences covering four positions in human protein GB1, where fitness was defined as the enrichment of folded protein bound to IgG-Fc antibody as measured by coupling mRNA display with next-generation sequencing (16). We used only measured sequences and did not incorporate imputed values of variants that were not measured directly. Three directed evolution approaches were simulated on this landscape: (*i*) a single-mutation walk, (*ii*) simulated recombination, and (*iii*) directed evolution with machine learning. For (*i*), the algorithm proceeds as follows: (*1*) from a starting sequence, every possible single mutation (19*N* variants for *N* positions) is made and evaluated; (*2*) the best single mutation is fixed in the reference sequence, and the position it was found in is locked from further editing; (*3*) steps (*1*) and (*2*) are repeated until every position has been tested, for a total of four rounds to cover four positions. (*ii*) Simulated recombination proceeds by selecting 527 random variants and recombining the mutations found in the top three variants, for an average of 570 variants tested over 10,000 simulations. (*iii*) Directed evolution with machine learning proceeds as follows: (*1*) 470 randomly selected sequences in the combinatorial space are used to train shallow neural networks with randomized hyperparameter search from fourfold cross-validation based on Pearson's *r*. Errors are then calculated based on 1,000 randomly selected variants that were not present in the training set. (*2*) The optimal model is used to predict the top 100 sequences, or the approximate screening capacity of a plate. (*3*) The highest true fitness value in this predicted set of 100 sequences and the training set of 470 is the maximum fitness value found. This process was repeated with different numbers of random sequences in (*i*) to simulate lower model accuracies, the results of which are shown in *SI Appendix*, Fig. S1A. In Fig. 2, 100 variants was used as the size of the predicted library test for its similarity to the screening capacity of a 96-well plate. With 570 total variants (*SI Appendix*, *Library Coverage*), this leaves 470 variants for the input library in (*1*) for an equal screening burden, assuming 95% coverage of 19 mutations from WT at each position.

Library Cloning, Expression, and Characterization of *Rma* NOD. The gene encoding *Rma* NOD was obtained as a gBlock and cloned into pET22b(+) (cat. no. 69744; Novagen). Standard PCR amplification and Gibson assembly were used for libraries with degenerate codons specified by SwiftLib (17). Encoded vs. sequenced codon distributions are shown in *SI Appendix*, Fig. S2. Expression was performed in 96 deep-well plates in 1 mL HyperBroth (AthenaES) using *Escherichia coli* BL21 *E. cloni* EXPRESS (Lucigen) with 100 μg/mL ampicillin from a 20-fold dilution of overnight culture. Expression cultures were induced after 2.5 h of outgrowth with 0.5 mM isopropyl β-D-1-thiogalactopyranoside, and heme production was enhanced with supplementation of 1 mM 5-aminolevulinic acid.

The relative product activity was measured by using 10 mM Me-EDA and 10 mM $PhMe_2SiH$ with whole cells resuspended in 400 μL nitrogen-free M9-N buffer, pH 7.4 (47.7 mM Na_2HPO_4, 22.0 mM KH_2PO_4, 8.6 mM NaCl, 2.0 mM $MgSO_4$, and 0.1 mM $CaCl_2$). Reactions were incubated anaerobically at room temperature for 6 h before extraction into 600 μL cyclohexane. Enantiomeric excess was measured by running the organic solution on a Jasco 2000 series supercritical fluid chromatography system with a CHIRALCEL OD-H (4.6 mm × 25 cm) chiral column (95% CO_2, 5% isopropanol, 3 min).

***Rma* NOD Model Training and Prediction Testing.** Screening information was paired with protein sequence obtained from rolling circle amplification followed by sequencing by MCLab. The sequence–function pairs, available on ProtaBank (45), were used to train a panel of models with default hyperparameters in the scikit-learn Python package (46), including K-nearest neighbors, linear (including Automatic Relevance Detection, Bayesian Ridge, Elastic Net, Lasso LARS, and Ridge), decision trees, random forests (including AdaBoost, Bagging, and Gradient Boosting), and multilayer perceptrons. The top three model types were selected, and grid-search cross-validation was used to identify the optimal hyperparameters. The top three hyperparameter sets for the top three model types were used to identify the top 1,000 sequences in each predicted library. Degenerate codons encoding amino acids occurring with highest frequencies in every model at each position were identified by SwiftLib (17), and 90 random variants were tested in vitro. This random sampling differs from that in the empirical fitness landscape, in which all sequences have been enumerated and can be easily tested. Even though sampling randomly means we may not have tested the optimal sequence as identified in trained models, we are able to generate fitness distributions as in Fig. 5 to describe this space.

ACKNOWLEDGMENTS. The authors thank Yisong Yue for initial guidance and Scott Virgil (Caltech Center for Catalysis and Chemical Synthesis) for providing critical instrument support; and Kevin Yang, Anders Knight, Oliver Brandenburg, and Ruijie Kelly Zhang for helpful discussions. This work is supported by National Science Foundation Grant GRF2017227007 (to Z.W.), the Rothenberg Innovation Initiative Program (S.B.J.K. and F.H.A.), and the Jacobs Institute for Molecular Engineering for Medicine at Caltech (S.B.J.K. and F.H.A.).

1. Petrović D, Kamerlin SCL (2018) Molecular modeling of conformational dynamics and its role in enzyme evolution. *Curr Opin Struct Biol* 52:50–57.
2. Romero PA, Arnold FH (2009) Exploring protein fitness landscapes by directed evolution. *Nat Rev Mol Cell Biol* 10:866–876.
3. Goldsmith M, Tawfik DS (2017) Enzyme engineering: Reaching the maximal catalytic efficiency peak. *Curr Opin Struct Biol* 47:140–150.
4. Zeymer C, Hilvert D (2018) Directed evolution of protein catalysts. *Annu Rev Biochem* 87:131–157.
5. Garcia-Borrás M, Houk KN, Jiménez-Oses G (2018) Computational design of protein function. *Computational Tools for Chemical Biology*, ed Martín-Santamaría S (Royal Society of Chemistry, London), pp 87–107.
6. Lewis RD, et al. (2018) Catalytic iron-carbene intermediate revealed in a cytochrome *c* carbene transferase. *Proc Natl Acad Sci USA* 115:7308–7313.
7. Dahiyat BI, Mayo SL (1997) De novo protein design: Fully automated sequence selection. *Science* 278:82–87.
8. Khersonsky O, et al. (2018) Automated design of efficient and functionally diverse enzyme repertoires. *Mol Cell* 72:178–186.e5.
9. Amrein BA, et al. (2017) *CADEE*: Computer-aided directed evolution of enzymes. *IUCrJ* 4:50–64.
10. Murphy KP (2012) *Machine Learning: A Probabilistic Perspective* (MIT Press, Cambridge, MA).
11. Jordan MI, Mitchell TM (2015) Machine learning: Trends, perspectives, and prospects. *Science* 349:255–260.
12. Silver D, et al. (2017) Mastering chess and shogi by self-play with a general reinforcement learning algorithm. arXiv:1712.01815v1.
13. Fox RJ, et al. (2007) Improving catalytic function by ProSAR-driven enzyme evolution. *Nat Biotechnol* 25:338–344.
14. Romero PA, Krause A, Arnold FH (2013) Navigating the protein fitness landscape with Gaussian processes. *Proc Natl Acad Sci USA* 110:E193–E201.
15. Yang KK, Wu Z, Arnold FH (2018) Machine learning in protein engineering. arXiv: 1811.10775v1.
16. Wu NC, Dai L, Olson CA, Lloyd-Smith JO, Sun R (2016) Adaptation in protein fitness landscapes is facilitated by indirect paths. *eLife* 5:e16965.
17. Jacobs TM, Yumerefendi H, Kuhlman B, Leaver-Fay A (2015) SwiftLib: Rapid degenerate-codon-library optimization through dynamic programming. *Nucleic Acids Res* 43:e34.
18. Reetz MT, Kahakeaw D, Lohmer R (2008) Addressing the numbers problem in directed evolution. *ChemBioChem* 9:1797–1804.
19. Bosley AD, Ostermeier M (2005) Mathematical expressions useful in the construction, description and evaluation of protein libraries. *Biomol Eng* 22:57–61.
20. Kan SBJ, Lewis RD, Chen K, Arnold FH (2016) Directed evolution of cytochrome c for carbon–silicon bond formation: Bringing silicon to life. *Science* 354:1048–1051.
21. Showell GA, Mills JS (2003) Chemistry challenges in lead optimization: Silicon isosteres in drug discovery. *Drug Discov Today* 8:551–556.
22. Franz AK, Wilson SO (2013) Organosilicon molecules with medicinal applications. *J Med Chem* 56:388–405.
23. Shi SL, Wong ZL, Buchwald SL (2016) Copper-catalysed enantioselective stereodivergent synthesis of amino alcohols. *Nature* 532:353–356.
24. Finefield JM, Sherman DH, Kreitman M, Williams RM (2012) Enantiomeric natural products: Occurrence and biogenesis. *Angew Chem Int Ed Engl* 51:4802–4836.
25. Reetz MT (2004) Controlling the enantioselectivity of enzymes by directed evolution: Practical and theoretical ramifications. *Proc Natl Acad Sci USA* 101:5716–5722.
26. The UniProt Consortium (2017) UniProt: The universal protein knowledgebase. *Nucleic Acids Res* 45:D158–D169.
27. Bloom JD, Labthavikul ST, Otey CR, Arnold FH (2006) Protein stability promotes evolvability. *Proc Natl Acad Sci USA* 103:5869–5874.
28. Fox R, et al. (2003) Optimizing the search algorithm for protein engineering by directed evolution. *Protein Eng* 16:589–597.
29. Kille S, et al. (2013) Reducing codon redundancy and screening effort of combinatorial protein libraries created by saturation mutagenesis. *ACS Synth Biol* 2:83–92.
30. Lissy NA (2003) Patentability of chemical and biotechnology inventions: A discrepancy in standards. *Washingt Univ Law Q* 81:1069–1095.
31. Steinberg B, Ostermeier M (2016) Environmental changes bridge evolutionary valleys. *Sci Adv* 2:e1500921.
32. Drummond DA, Iverson BL, Georgiou G, Arnold FH (2005) Why high-error-rate random mutagenesis libraries are enriched in functional and improved proteins. *J Mol Biol* 350:806–816.
33. Bershtein S, Segal M, Bekerman R, Tokuriki N, Tawfik DS (2006) Robustness-epistasis link shapes the fitness landscape of a randomly drifting protein. *Nature* 444:929–932.
34. Zhang RK, et al. (2019) Enzymatic assembly of carbon–carbon bonds via iron-catalysed sp3 C–H functionalization. *Nature* 565:67–72.
35. Kim DE, et al. (2014) One contact for every twelve residues allows robust and accurate topology-level protein structure modeling. *Proteins* 82(Suppl 2):208–218.
36. Fowler DM, Fields S (2014) Deep mutational scanning: A new style of protein science. *Nat Methods* 11:801–807.
37. Sinai S, Kelsic E, Church GM, Nowak MA (2017) Variational auto-encoding of protein sequences. arXiv:1712.03346v3.
38. Riesselman AJ, Ingraham JB, Marks DS (2018) Deep generative models of genetic variation capture the effects of mutations. *Nat Methods* 15:816–822.
39. Brookes DH, Listgarten J (2018) Design by adaptive sampling. arXiv:1810.03714v3.
40. Saito Y, et al. (2018) Machine-learning-guided mutagenesis for directed evolution of fluorescent proteins. *ACS Synth Biol* 7:2014–2022.
41. Cadet F, et al. (2018) A machine learning approach for reliable prediction of amino acid interactions and its application in the directed evolution of enantioselective enzymes. *Sci Rep* 8:16757.
42. Carlin DA, et al. (2016) Kinetic characterization of 100 glycoside hydrolase mutants enables the discovery of structural features correlated with kinetic constants. *PLoS One* 11:e0147596.
43. Barley MH, Turner NJ, Goodacre R (2018) Improved descriptors for the quantitative structure-activity relationship modeling of peptides and proteins. *J Chem Inf Model* 58:234–243.
44. Yang KK, Wu Z, Bedbrook CN, Arnold FH (2018) Learned protein embeddings for machine learning. *Bioinformatics* 34:4138.
45. Wang CY, et al. (2018) ProtaBank : A repository for protein design and engineering data. *Protein Sci* 27:1113–1124.
46. Pedregosa F, et al. (2011) Scikit-learn: Machine learning in Python. *J Mach Learn Res* 12:2825–2830.
47. Waterhouse A, et al. (2018) SWISS-MODEL: Homology modelling of protein structures and complexes. *Nucleic Acids Res* 46:W296–W303.