JACS Au AN OPEN ACCESS JOURNAL OF THE AMERICAN CHEMICAL SOCIETY

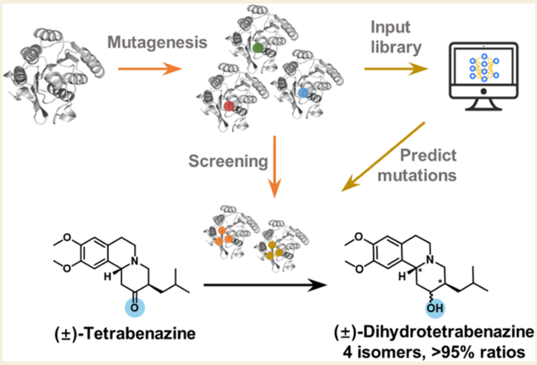
Application of Directed Evolution and Machine Learning to Enhance the Diastereoselectivity of Ketoreductase for Dihydrotetrabenazine Synthesis

Published as part of JACS Au virtual special issue "Biocatalysis in Asia and Pacific".

Chenming Huang, Li Zhang, Tong Tang, Haijiao Wang, Yingqian Jiang, Hanwen Ren, Yitian Zhang, Jiali Fang, Wenhe Zhang, Xian Jia, Song You,* and Bin Qin*

Cite This: *JACS Au* 2024, 4, 2547−2556 Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information



ABSTRACT: Biocatalysis is an effective approach for producing chiral drug intermediates that are often difficult to synthesize using traditional chemical methods. A time-efficient strategy is required to accelerate the directed evolution process to achieve the desired enzyme function. In this research, we evaluated machine learning-assisted directed evolution as a potential approach for enzyme engineering, using a moderately diastereoselective ketoreductase library as a model system. Machine learning-assisted directed evolution and traditional directed evolution methods were compared for reducing (±)-tetrabenazine to dihydrotetrabenazine via kinetic resolution facilitated by BsSDR10, a short-chain dehydrogenase/reductase from *Bacillus subtilis*. Both methods successfully identified variants with significantly improved diastereoselectivity for each isomer of dihydrotetrabenazine. Furthermore, the preparation of (2*S*,3*S*,11b*S*)-dihydrotetrabenazine has been successfully scaled up, with an isolated yield of 40.7% and a diastereoselectivity of 91.3%.

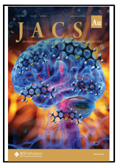
KEYWORDS: *ketoreductase, stereodivergent evolution, machine learning, dihydrotetrabenazine, kinetic resolution*

INTRODUCTION

The application of biocatalysis in the pharmaceutical industry is steadily increasing, driven by the emergence of novel engineered enzymes and proven enzymatic processes.[1,2] The biocatalytic strategy is widely used in the synthesis of active pharmaceutical ingredients due to the clear advantages it offers.[3−5] Biocatalysis can be used to functionalize certain compounds, providing milder and more environmentally friendly reaction conditions while offering exceptional selectivity. Biocatalysis has been instrumental in the synthesis of the key intermediates for well-known pharmaceuticals, such as atorvastatin (Lipitor), montelukast (Singulair), and duloxetine (Cymbalta).[6,7]

Protein engineering is often necessary for the use of biocatalysis and enzymes in non-natural reactions. This requires obtaining biocatalysts that are both active and stable. Directed evolution is a powerful technique that has gained popularity in the field of protein engineering over the last few decades.[8] (1) In traditional directed evolution, single mutations are introduced into wild-type enzymes or proteins sequentially. Subsequently, the resulting variants are subjected to screening for the desired properties (Figure 1a). An alternative approach is to randomly sample combinatorial libraries of mutations and recombine the most promising mutations at each position to generate optimal variants (Figure 1b). Both methods result in the generation of optimal variants, which serve as parental sequences for the subsequent stage of evolution. A variety of techniques, including error-prone PCR (epPCR),[9,10] saturation mutagenesis,[11] and others, can be used to explore the vast sequence space and discover new biocatalysts.[12,13] Traditional directed evolution has been extensively used to improve the regioselectivity,[14] diastereoselectivity,[15−17] and enzyme activity.[18,19] (2) Recently, machine learning has emerged as a powerful tool to support directed evolution, enabling the exploration of larger sequence spaces (Figure 1c). Numerous studies have demonstrated the capacity of machine learning to predict sequence-activity (selectivity) relationships.[20−28] While there has been considerable research in machine learning for protein engineering,[29−41] the approach is seldom used to achieve stereodivergent synthesis and only a

Received: March 29, 2024
Revised: June 13, 2024
Accepted: June 20, 2024
Published: June 26, 2024



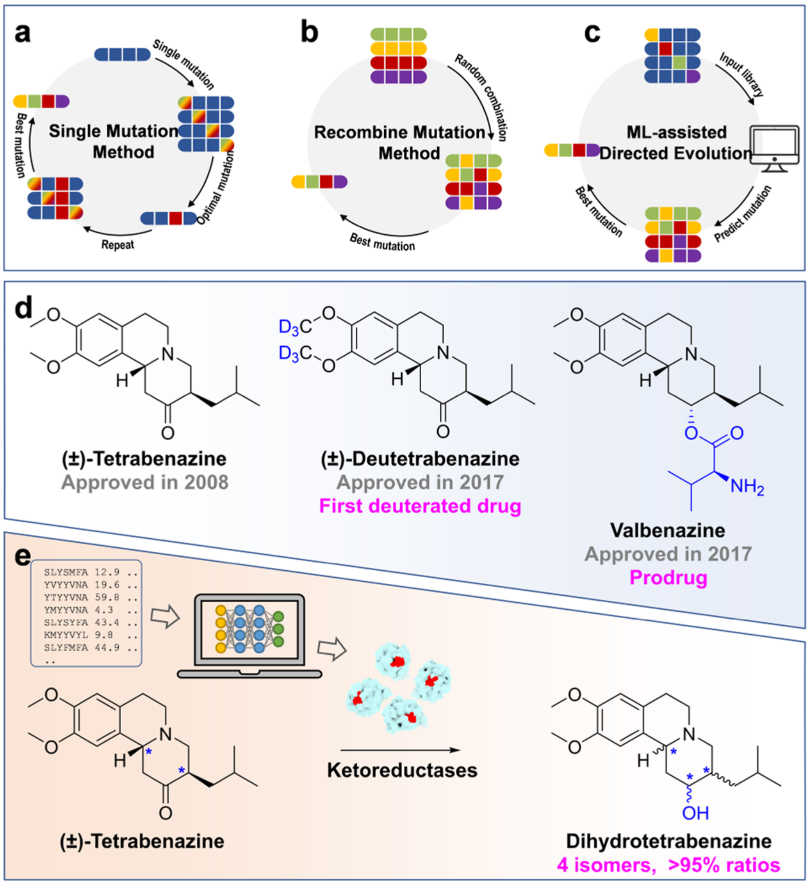


Figure 1. An overview of obtaining high-diastereoselectivity mutants for tetrabenazine using traditional directed evolution and machine learning-assisted directed evolution. (a) Single mutation-directed evolution required 19 mutations for a sequence of N amino acids. (b) Directed evolution using recombining mutations. (c) Machine learning-assisted directed evolution allows for the training of the mutation library on a machine learning model, enabling the simultaneous search of multiple positions and the exploration of the sequence−function relationship more broadly and deeply. (d) The chemical structures of (±)-tetrabenazine, (±)-deutetrabenazine, and valbenazine are presented, along with their respective release years and indications. (e) The use of machine learning in protein engineering to improve the diastereoselectivity of ketoreductase for kinetic resolution of (±)-tetrabenazine.

limited number of studies have compared directed evolution and machine-directed evolution.[30,42]

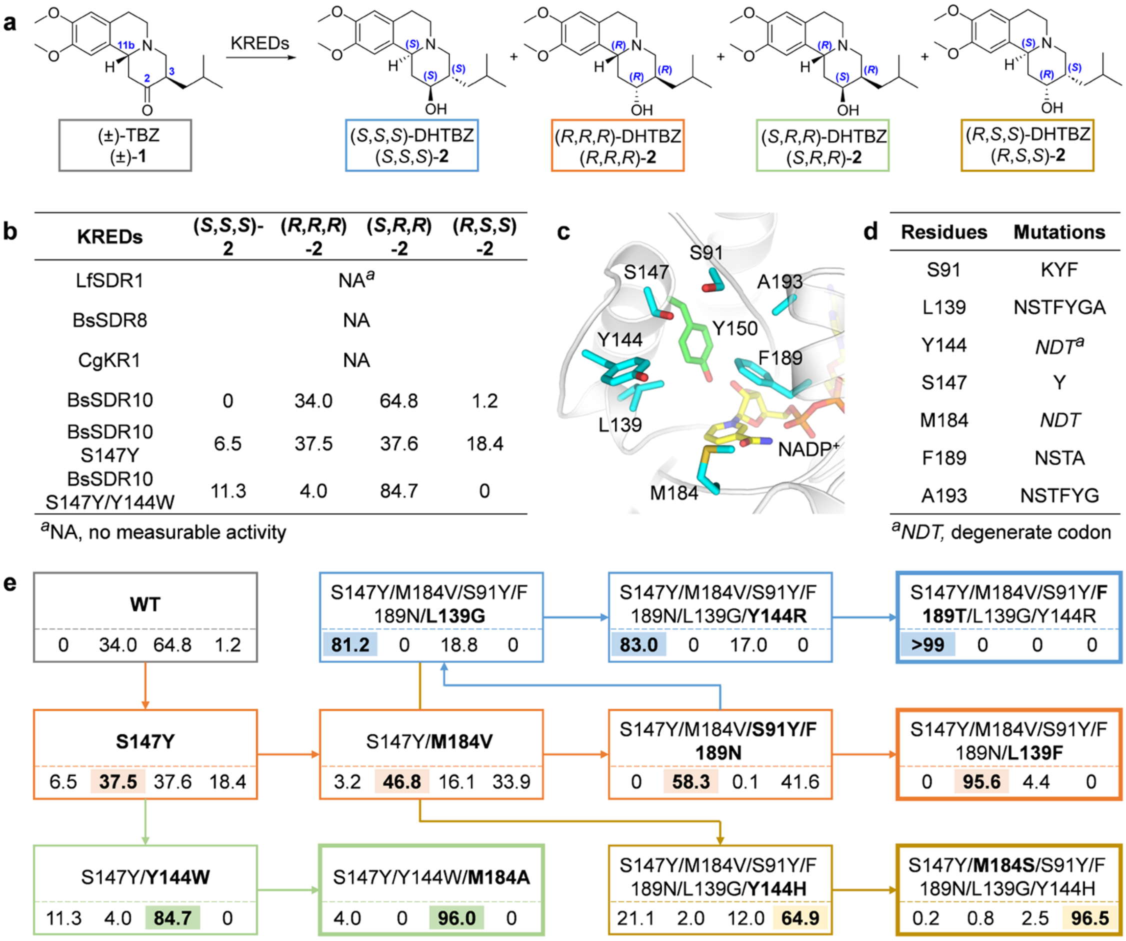
(±)-Tetrabenazine ((±)-TBZ), (±)-deutetrabenazine, and valbenazine (Figure 1d) are vesicular monoamine transporter-2 (VMAT-2) inhibitors that have been approved by the FDA for the treatment of tardive dyskinesia.[43−45] (±)-Deutetrabenazine contains deuterium, which increases the half-lives and prolongs the activity of (±)-TBZ. Valbenazine is designed as the prodrug of (2*R*,3*R*,11b*R*)-dihydrotetrabenazine (DHTBZ), the active metabolite of TBZ. Valbenazine is chemically synthesized through a four-step process that involves resolving the resolution of racemic (±)-TBZ to (3*R*,11b*R*)-TBZ, followed by the reducing (3*R*,11b*R*)-TBZ to produce (2*R*,3*R*,11b*R*)-DHTBZ. In comparison to the chemical synthesis of valbenazine, the ketoreductase (KRED)-catalyzed reductive kinetic resolution of (±)-TBZ has the potential to reduce the aforementioned resolution/reduction process to a single step. This is highly desirable due to its superior economic, sustainable, and environmentally friendly characteristics.

In this study, we compared machine learning-assisted directed evolution with traditional directed evolution in terms of diastereoselectivity (Figure 1e). Using traditional directed evolution methods, we generated various BsSDR10 variants capable of producing four isomers of DHTBZ, each with diastereoselectivity exceeding 95%. In parallel, machine learning-assisted directed evolution enabled us to derive mutants with high diastereoselectivity. Notably, these variants exhibited notable differences from those generated by traditional directed evolution. Upon evaluation, we found that machine learning was a more cost-effective approach and could provide valuable guidance on whether to use machine-directed evolution for efficiency. Ultimately, our study demonstrated two approaches for engineering enzyme catalysts and addressed the issue of obtaining dihydrotetrabenazine through biocatalysis.

■ RESULTS AND DISCUSSION

Identification of an Active Starting KRED for the Reduction of Tetrabenazine

To identify a starting KRED capable of accepting (±)-TBZ ((±)-**1**) with its relatively bulky structure (Figure 2a), we screened both wild-type and engineered KREDs from our in-house enzyme library. The initial activity of some of these enzymes for (±)-TBZ is shown in Figure 2b. Table S4 presents other variants that were screened. All KREDs expressed in *E. coli* Rosetta (DE3) were used for crude cell lysate reduction of (±)-TBZ in 1.5 mL Eppendorf tubes. Diastereoselectivity toward the target substrate was observed for both wild-type and engineered enzymes by high-performance liquid chroma-



KREDs	(S,S,S)-2	(R,R,R)-2	(S,R,R)-2	(R,S,S)-2
LfSDR1	NA[a]			
BsSDR8	NA			
CgKR1	NA			
BsSDR10	0	34.0	64.8	1.2
BsSDR10 S147Y	6.5	37.5	37.6	18.4
BsSDR10 S147Y/Y144W	11.3	4.0	84.7	0

[a]NA, no measurable activity

Residues	Mutations
S91	KYF
L139	NSTFYGA
Y144	*NDT*[a]
S147	Y
M184	*NDT*
F189	NSTA
A193	NSTFYG

[a]*NDT*, degenerate codon

Figure 2. Engineering of BsSDR10 using traditional directed evolution. (a) Kinetic resolution of (±)-TBZ. (b) Screening for KREDs capable of reducing (±)-TBZ. (c) The residues selected for engineering of BsSDR10. (d) Mutations constructed for each of the chosen residues. (e) Directed evolution of BsSDR10 for the synthesis of four distinct isomers of DHTBZ.

tography (HPLC). In particular, the wild-type enzyme BsSDR10, a short-chain dehydrogenase/reductase from *Bacillus subtilis*, exhibited the ability to generate products, signifying a superior potential when compared to other wild-type enzymes such as LfSDR1, BsSDR8, and CgKR1. Furthermore, the variant BsSDR10 S147Y/Y144W[46] (from the in-house enzyme library) exhibited a markedly higher preference for (2*S*,3*R*,11b*R*)-DHTBZ ((*S*,*R*,*R*)-**2**), resulting in an enhanced diastereoselectivity (84.7%) toward the targeted reaction (Figure 2b). Therefore, LfSDR1, BsSDR8, and CgKR1 were not selected as starting points for enzyme evolution. BsSDR10 was ultimately selected due to its potential for improved diastereoselectivity in engineering.

Engineering of BsSDR10 Using Traditional Directed Evolution

To accomplish our objective of reducing (±)-TBZ and obtaining four distinct isomers of DHTBZ, we initiated the process by utilizing traditional directed evolution methods to engineer BsSDR10. As mentioned above, wild-type BsSDR10 exhibits poor diastereoselectivity toward (±)-TBZ ((±)-**1**), which precludes the production of products with a single isomer. Nevertheless, by mutating only two residues (S147Y/Y144W), an initial diastereoselectivity of 84.7% for (*S*,*R*,*R*)-**2** toward (±)-**1** was achieved. To predict the protein structure of BsSDR10, we employed the SWISS-MODEL and AlphaFold 2, given the absence of a crystal structure. However, the resulting protein structure was deemed to be of insufficient accuracy, which prevented the substrates from fully docking into the modeled protein structure.

Our preliminary research results[46] indicate that positions S91, L139, Y144, S147, M184, F189, and A193 in BsSDR10 may be key residues that influence the size of the binding pocket and control the diastereoselectivity. Figure 2c shows that positions 139, 144, and 184 are located at the entrance of the binding pocket, while positions 91, 147, 189, and 193 are located at the bottom of the binding pocket. We hypothesized that the bulky residue M184 causes channel constriction, thereby preventing substrate **1** from accessing the binding pocket. Therefore, we selected smaller amino acids (Ala and Gly) for mutation at position 184. This mutation may have resulted in the larger volume binding pocket of BsSDR10. As previously stated, the BsSDR10 S147Y/Y144W variant, which exhibited 84.7% diastereoselectivity toward (*S*,*R*,*R*)-**2**, was thus mutated to S147Y/Y144W/M184A and S147Y/Y144W/M184G. The HPLC assays revealed that both variants exhibited high diastereoselectivity (96.0% and 98.3%) toward (*S*,*R*,*R*)-**2** (Figure 2e, green, and Table S7). These results are significant and provide exciting new opportunities for our experimentation and understanding.

Furthermore, another BsSDR10 S147Y variant was identified from the in-house enzyme library, which exhibited enhanced diastereoselectivity for (*R*,*R*,*R*)-**2** (Figure 2b). The

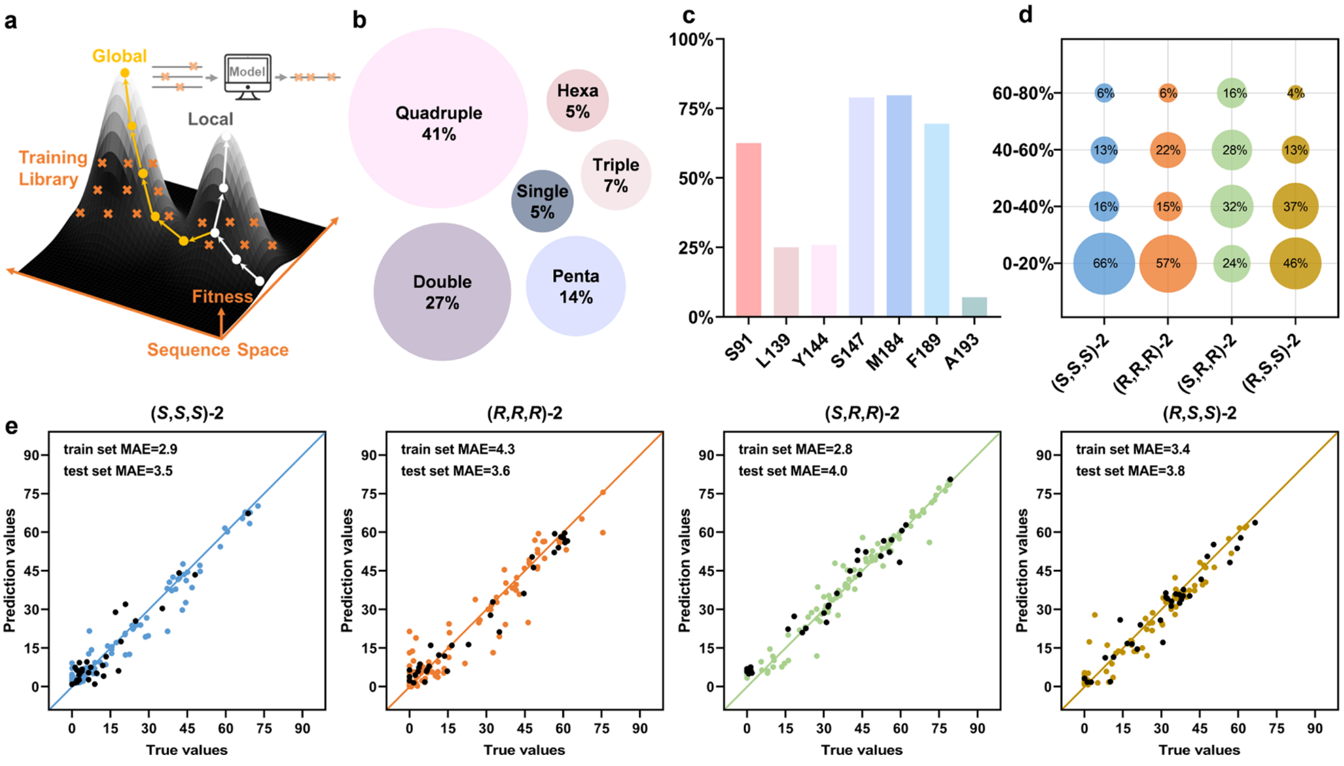


Figure 3. Distribution of the training data set and the performance of GRU on the data set. (a) Machine learning model-assisted directed evolution. ML models can help to explore the broader protein sequence space. (b) The distribution of the mutations in the data set. (c) The distribution of the residues in the data set. (d) The distribution of diastereoselectivity in the data set. (e) The performance of GRU on four isomers in the data set. The test set for four isomers is represented by black circles, while the train set is represented by blue, orange, green, and amber.

single mutant S147Y was then used as the starting point to construct a combinatorial library. For the first round of evolution, we continue to mutate at position M184 and incorporate the degenerate codon NDT (Figure 2d). After screening and sequencing of the variants, it was found that the diastereoselectivity of S147Y/M184V was increased from 37.5% to 46.8% for (*R*,*R*,*R*)-**2** (the results of the remaining variants are shown in Table S5). Since (*R*,*R*,*R*)-**2** and (*S*,*R*,*R*)-**2** have mutually flipped conformations in the binding pocket, the bulky residue F at position 189 might shrink one side of the pocket, while the smaller residue S at position 91 might enlarge the other side of the pocket. We hypothesized that protein engineering at positions 91 and 189 could reshape the binding pocket compared to S147Y/Y144W/M184A or S147Y/Y144W/M184G. Therefore, to test our hypothesis, we constructed combinatorial mutations at positions 91 and 189 based on the S147Y/M184V variant, with the aim of further improving the diastereoselectivity. As previously demonstrated, these two residues are key to control the stereopreferences of SDRs.[46]

For the S91 position of the BsSDR10 S147Y/M184V variant, we constructed mutations with larger volumes or hydrophobic amino acids (Phe, Lys, and Tyr) (Figure 2d). The purpose of this modification is to enhance hydrophobic interactions with the substrate and alter the pocket size. Meanwhile, since F189 is a large hydrophobic amino acid, we mutate F189 of the BsSDR10 S147Y/M184V variant into amino acids with different side-chain properties, including polar side chains such as Asn, Ser, and Thr, as well as transforming it into smaller volume amino acid (Ala) (Figure 2d). Several studies have shown that the proper combination of mutations is more important. Ultimately, the test results showed that the combinatorial mutation S147Y/M184V/S91Y/F189N increased the diastereoselectivity (58.3%) toward (*R*,*R*,*R*)-**2** (the remaining variants are shown in Table S5), thereby illustrating the extensibility of combinatorial mutations. Based on these findings, we selected S147Y/M184V/S91Y/F189N as the parent for the subsequent mutations. Similarly, the candidate residues at positions 193 and 139 were designed with the intention of covering a wide range of types based on differences in volume and polarity (Figure 2d). A combinatorial library containing the two positions was thus screened, resulting in the identification of an improved pentaploid mutant S147Y/M184V/S91Y/F189N/L139F (the remaining variants at position 139 are shown in Table S5). The HPLC assays showed that it has a high diastereoselectivity of 95.6% toward (*R*,*R*,*R*)-**2** (Figure 2e, orange).

Meanwhile, our results showed that the stereopreference of the variant S147Y/M184V/S91Y/F189N/L139F could be switched by mutation L139F to L139G, resulting in an 81.2% diastereoselectivity toward (*S*,*S*,*S*)-**2** for the S147Y/M184V/S91Y/F189N/L139G variant. This variant was then used as the starting point for a subsequent combinatorial library, which contained the Y144, an additional site located at the entrance of the binding pocket (Figure 2d). The test results showed that replacing Y144 with two positively charged residues, Arg (R) and His (H) (the remaining variants are shown in Table S5), could alter the diastereoselectivity. The S147Y/M184V/S91Y/F189N/L139G/Y144R variant exhibited an 83% diastereoselectivity toward (*S*,*S*,*S*)-**2**, while the S147Y/M184V/S91Y/F189N/L139G/Y144H variant showed a diastereoselectivity of 64.9% toward (*R*,*S*,*S*)-**2**. Following the initial investigation, a comprehensive exploration of mutational effects was conducted. Single mutations were performed at various sites, and it was found that mutagenesis of residues at positions 189 and 184 to polar uncharged Thr (T) and Ser (S) resulted in a significant increase in diastereoselectivity (the remaining variants are shown in Table S5). Ultimately, the S147Y/M184V/S91Y/F189T/L139G/Y144R variant showed a >99% diastereoselectivity toward (*S*,*S*,*S*)-**2** (Figure 2e, blue), whereas the S147Y/M184S/S91Y/F189N/L139G/Y144H variant showed a 96.5% diastereoselectivity toward (*R*,*S*,*S*)-**2** (Figure 2e, amber).

In summary, we were unable to obtain the protein crystal structure or successfully perform molecular docking of BsSDR10. Nevertheless, we were able to achieve high diastereoselectivity through traditional directed evolution (Figure 2e), resulting in four variants S147Y/M184V/S91Y/F189T/L139G/Y144R (>99% for (*S*,*S*,*S*)-**2**), S147Y/M184V/S91Y/F189N/L139F (95.6% for (*R*,*R*,*R*)-**2**), S147Y/Y144W/M184A (96.0% for (*S*,*R*,*R*)-**2**), and S147Y/M184S/S91Y/F189N/L139G/Y144H (96.5% for (*R*,*S*,*S*)-**2**). The diastereoselectivity data highlight the significant adaptability of BsSDR10, which implies a considerable diastereoselectivity of products by mutating amino acids strategically positioned around the binding pocket.

Engineering of BsSDR10 Using Machine Learning-Assisted Directed Evolution

The traditional direct evolution approach has proven effective in optimizing diastereoselectivity. However, the method is constrained in that it can only sample a subset of protein sequences. In contrast, machine learning-assisted directed evolution enables the exploration of larger sequence spaces (Figure 3a). A number of studies have been conducted on the use of machine learning to predict sequence-selectivity relationships.[20,29] In this study, the use of machine learning was employed to guide the directed evolution process in the absence of the protein crystal structure, contributing to the optimal diastereoselectivity of mutants. Research has demonstrated that machine learning-assisted directed evolution could reduce the screening burden and enhance efficiency compared to traditional directed evolution. We subsequently utilized the data from the aforementioned variants collected from the mutational scanning library to train prediction models and engineer BsSDR10 using machine learning-assisted directed evolution to further enhance its diastereoselectivity.

A total of over 300 variants of data were produced by the biocatalytic reactions mentioned above. These variants were generated through the use of an in-house enzyme library, as well as NDT-based saturation mutagenesis and combination mutagenesis approaches. Our initial selection focused on variants exhibiting diastereoselectivity ratios ranging from 0 to 80% for the four products, which constituted the primary data set. Due to the limited number of samples with diastereoselectivity in excess of 80%, their inclusion could potentially lead to overfitting of the model and compromise its ability to generalize. In addition, we also hope to identify highly diastereoselective mutants that differ from those obtained through traditional directed evolution.

The effectiveness of machine learning models can be enhanced when trained on data that is widely distributed across the input space.[47] The presence of similar variants with the same diastereoselectivity in a data set can result in data redundancy.[48,49] The simplification of the model by the removal of these redundant data points reduces the complexity of the model. Therefore, we removed data with identical diastereoselectivity values for different variants, as well as variants that showed no reaction toward the substrate. This resulted in 128 variants of data, all of which were confirmed by Sanger sequencing. Figure 3b shows that the percentage of single mutations was the lowest (5%), while quadruple mutations had the highest percentage (41%). The distribution of mutations at the seven active sites is displayed in Figure 3c, with the least proportion of mutations occurring at position 193 (7%). Figure 3d illustrates the diastereoselectivity intervals of the mutation library for four products, along with the percentage of mutations occurring at various intervals. The diastereoselectivity of variants for products (*S*,*S*,*S*)-**2**, (*R*,*R*,*R*)-**2**, (*S*,*R*,*R*)-**2**, and (*R*,*S*,*S*)-**2** was mostly distributed in the 0−20% interval, with percentages of 66, 57, 24, and 46%, respectively. However, variants with diastereoselectivity in the 60−80% range were the least common for (*R*,*S*,*S*)-**2**, with a 4% distribution. The absence of these data may have an impact on the subsequent construction of the machine learning model. Once the data set had been prepared, the input file for machine learning was created.

In the feature selection step, 242 of the 249 amino acids in BsSDR10 were excluded, as only seven of them were mutated in the 128 variants under study. Furthermore, the biochemical features of the amino acids at the seven positions, including volume, hydrophobicity, hydrophilicity, isoelectric point, and hydrogen bond (Table S6), were also incorporated. The average number of hydrations (bound water molecules) occurring each time indicates hydrophilicity.[50] The hydrophobic index is used to express the hydrophobicity.[51] Both the hydrogen bond donors and full nonbonding orbitals were introduced as features to describe the hydrogen bond.[52] The data were labeled to represent the ratio of the diastereoselectivity for four products of each variant. Subsequently, a model was constructed to predict variants with higher diastereoselectivity. A variety of models were used to achieve the objective of prediction, including random forest (RF), deep residual network 50 (ResNet50), decision tree regressor (DTR), Gaussian process regressor (GPR), ridge regression (RR), and gate recurrent unit (GRU). The data set was randomly partitioned into a training set (75%) and a test set (25%), and the hyperparameters were tuned for each model (Table S2), and subsequently, the models were trained using the data set.

The GRU predictor demonstrated superior performance relative to other models, as evidenced by its exceptional performance on test sets of four isomers (Figure 3e) and by its ability to yield the lowest mean absolute error (MAE) values (Figure S1). This model is then used to predict optimal variants. Given the lack of a crystal structure for BsSDR10 and the imprecise results of homology modeling, machine learning is expected to predict high-diastereoselectivity variants. The machine learning suggested that variants were ranked and several were selected based on their predicted higher diastereoselectivity toward four products: (*S*,*S*,*S*)-**2**, (*R*,*R*,*R*)-**2**, (*S*,*R*,*R*)-**2**, and (*R*,*S*,*S*)-**2**. We then investigated their diastereoselectivity through experiments.

To our delight, although the diastereoselectivity values were diverse, a number of variants exhibited high diastereoselectivity. The BsSDR10 S91F/L139G/S147Y/M184V/F189G/A193V variant (FGYYVGV) showed the highest diastereoselectivity (97.4%) toward (*S*,*S*,*S*)-**2** (Figure 4, blue circles), while the BsSDR10 S91F/L139M/S147Q/M184T/F189G/

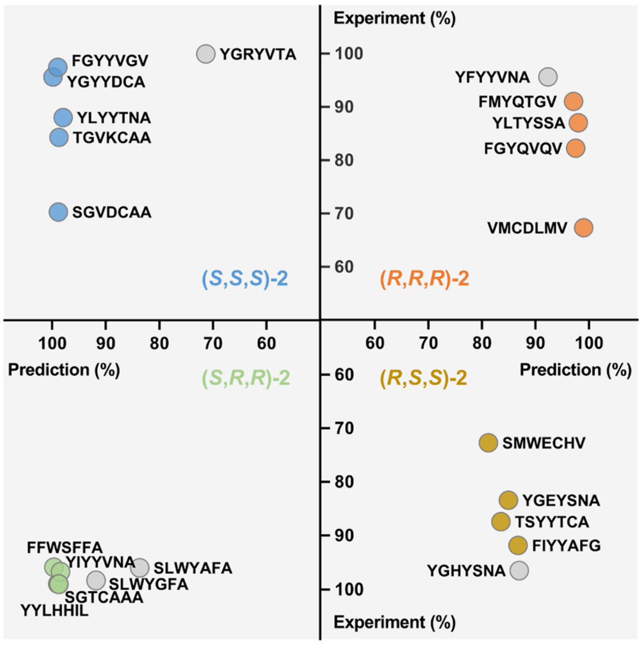


Figure 4. Results of asymmetric reduction for the best variants obtained by mutational scanning (gray) and machine learning-assisted directed evolution (blue, orange, green, and amber). The *x*-axis displays the predicted values from the model, while the *y*-axis shows the values from experimental results, with the origin starting at 50%. For the variants predicted by machine learning, (*S*,*S*,*S*)-**2** is represented in blue, (*R*,*R*,*R*)-**2** in orange, (*S*,*R*,*R*)-**2** in green, and (*R*,*S*,*S*)-**2** in amber. Additionally, variants obtained from traditional directed evolution are also predicted using machine learning models and are represented by gray circles (the seven residues represented in order are positions 91, 139, 144, 147, 184, 189, and 193 of the variants).

A193V variant (FMYQTGV) gave a 91% diastereoselectivity toward (*R*,*R*,*R*)-**2** (Figure 4, orange circles). Two variants, named BsSDR10 L139G/Y144T/S147C/M184A/F189A (SGTCAAA) (Figure 4, green circles) and BsSDR10 S91F/L139I/S147Y/M184A/A193G (FIYYAFG) (Figure 4, amber circles), gave diastereoselectivities of 99% and 91.8% for (*S*,*R*,*R*)-**2** and (*R*,*S*,*S*)-**2**, respectively. Additionally, there are also some mutants that exhibit some degree of diastereoselectivity (Table S7), while these variants differ significantly from the highly selective variants obtained in the traditional directed evolution mentioned above. This suggests that machine-directed evolution is not only capable of producing mutants with high diastereoselectivity but also of predicting previously unseen mutants. The model extends exploration to a wider range of sequences. However, the experimental diastereoselectivity of the majority of the variants obtained from machine learning, particularly those for (*S*,*S*,*S*)-**2**, (*R*,*R*,*R*)-**2**, and (*R*,*S*,*S*)-**2**, was found to be lower than the predicted diastereoselectivity.

Kinetic Resolution for Synthesis of DHTBZ

It is important to note that short reaction times with enzymes may not be an accurate indicator of mutant activity and diastereoselectivity. Therefore, it is essential to measure the conversion and diastereoselectivity over time. Following our enzyme engineering, conversion was recorded for the S147Y/M184V/S91Y/F189N/L139F, Y144W/S147Y/M184G, and S147Y/M184S/S91Y/F189N/L139G/Y144H variants (to facilitate comparison with the machine learning variants, they were named YFYYVNA, SLWYGFA, and YGHYSNA) produced through traditional directed evolution and another variant FGYYVGV generated by the machine learning-assisted directed evolution. Under optimal conditions, the conversion over time for the four variants is shown in Figure 5. The four variants exhibited a gradual increase in conversion over the tested time period, while YFYYVNA, FGYYVGV, and YGHYSNA showed a slight decrease in diastereoselectivity value.

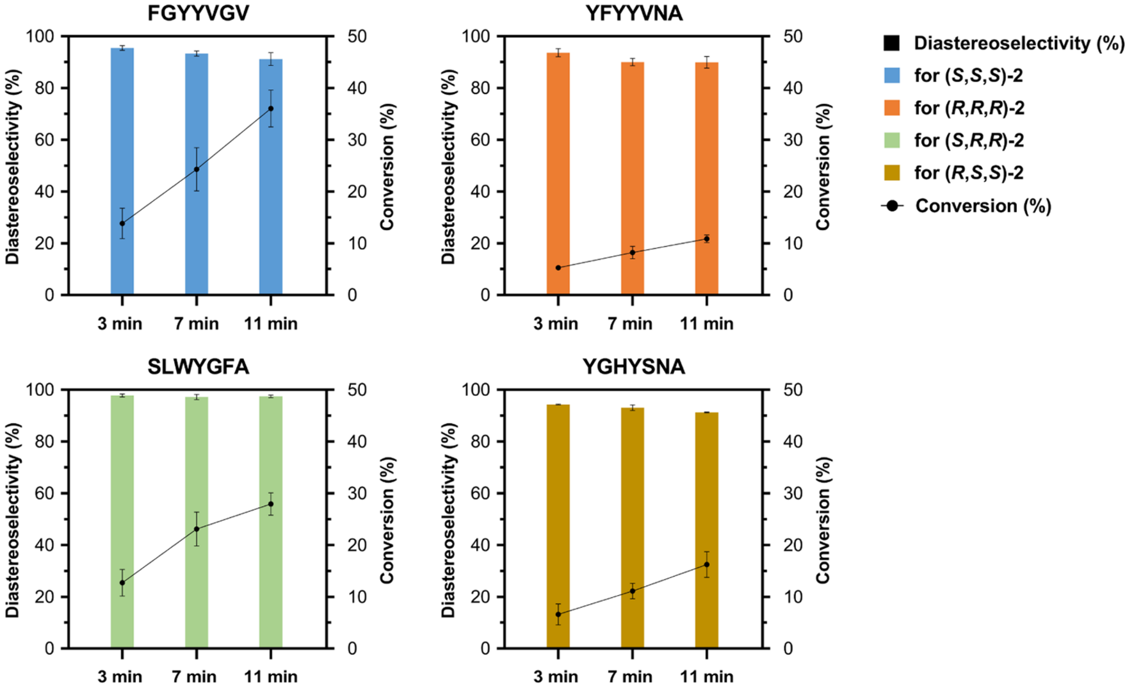


Figure 5. Time course of conversion and diastereoselectivity of the selected four variants producing each isomer of DHTBZ.

The FGYYVGV variant showed the highest conversion and significant diastereoselectivity values. Therefore, a comprehensive investigation was conducted to ascertain the potential applications of this variant in the synthesis of the isomer of (*S*,*S*,*S*)-**2**. The scale-up preparation was successfully carried out on 120 mL, with an isolated yield of 40.7% (68 mg) and a diastereoselectivity value of 91.3%. In addition, the desired isomer, (*R*,*R*,*R*)-**2**, was also prepared on a larger scale in a 120 mL system using the YFYYVNA variant. However, due to the relatively low conversion rate and the unstable of (*R*,*R*,*R*)-**2**, an isolated yield of 12.7% and a diastereoselectivity value of 92.5% were obtained.

DISCUSSION

We have shown that machine learning-assisted directed evolution is a straightforward and efficacious approach for engineering KREDs. This method enables the identification of mutants with comparable diastereoselectivity to traditional methods by screening only a subset of the mutant library. In our case, machine learning methods effectively address the limitation of screening by offering a new exploration of protein sequence space for high diastereoselectivity. Our findings indicate that machine learning can alleviate the physical strain of screening and expedite the acquisition of advantageous mutations.

This work primarily focuses on the prediction of diastereoselectivity, which is considerably more complex than the prediction of enzyme mutational activity. The complexity arises from the intricacy of the prediction, the quantity/amount of data required, and the processing of the resulting predictions. Diastereoselectivity depends not only on the interaction between the enzyme and the substrate but also on the relative position of the substrate in the active pocket. Furthermore, the techniques required for diastereoselectivity experiments are often more sophisticated and costly than those required for activity assays. Nevertheless, the ability to predict diastereoselectivity can offer valuable insights into how mutations impact substrate selection and aid in the development of highly specific enzymes with minimal side reactions.

Most enzymes require further optimization to achieve the desired properties at the discovery stage. Traditional directed evolution screening can only sample a small fraction of sequences in the protein fitness landscape, and it tends to ignore the nonadditive effects of accumulating multiple mutations (epistasis). As a result, directed evolution efforts may end up trapped in a local optimum. However, machine learning models can be used to learn mappings between protein sequences and their associated fitness values to overcome this limitation. The models can then predict the fitness of protein variants that have not been seen before, enhancing the efficiency of screening by performing protein evaluations *in silico* and increasing the breadth of exploration. This allows more sequences to be explored compared to traditional directed evolution methods.

Collecting data for machine learning and performing traditional directed evolution are both labor-intensive but they differ in their degree of difficulty. Traditional directed evolution involves a multistep experimental process that includes the design, preparation, selection, and evaluation of mutant strains. This process is time-consuming and often requires multiple iterations to achieve optimized results. In contrast, although acquiring extensive training data for machine learning can be labor-intensive, the workload subsequently diminishes after successfully generating a high-quality library of mutants through targeted mutation at active sites. Once a potentially predictive model is established, the prediction processes become faster. In addition, despite the difficulties inherent in model interpretation, machine learning is still regarded as a preferred strategy for enzyme evolution. Therefore, machine learning is more efficient and time-saving when sufficient experimental data and computational resources are available or when screening is impractical due to cost, time, or other constraints. Our study suggests that the machine learning model may not perform optimally in predicting (*R*,*S*,*S*)-selective and (*R*,*R*,*R*)-selective mutants due to the lack of high-diastereoselectivity data. This deficiency may impede the model training process and make it challenging to obtain mutants with higher diastereoselectivity.

At present, machine learning-assisted directed evolution still faces several challenges. One of the challenges is to enhance the prediction accuracy and reliability of machine learning methods. Further refinement is required in this study with regard to the experimental diastereoselectivity of some of the variants obtained by machine learning-assisted directed evolution, which was found to be lower than their predicted diastereoselectivity. Moreover, the performance of the variants obtained by machine learning was found to be inferior to that of the best variants obtained by traditional directed evolution for three of the four dihydrotetrabenazine isomers. This may be attributed to the limited quantity and quality of our training set, for instance, the relative paucity of variants exhibiting greater than 60% diastereoselectivity for the three products (*S*,*S*,*S*)-**2**, (*R*,*R*,*R*)-**2**, and (*R*,*S*,*S*)-**2**. This can be achieved by improving feature representation methods, selecting features that are better suited to protein sequences, and using more advanced machine learning algorithms and models. Additionally, there is also the potential to explore more training data and protein sequences to improve the generalization and adaptability of the model.

CONCLUSIONS

This research presents two methods for the production of variants of BsSDR10 with high diastereoselectivity in the reduction of (±)-tetrabenazine: traditional directed evolution and machine learning-assisted directed evolution. Moreover, both methods yielded four isomers of dihydrotetrabenazine. Furthermore, the results demonstrate that machine learning-assisted directed evolution can predict the diastereoselectivity of BsSDR10 variants. This is achieved through the processes of model construction, data training, selectivity prediction, and experimental validation. Machine learning-assisted directed evolution is well-suited to applications in enzyme evolution, offering the potential to efficiently explore a large sequence space. In addition, the FGYYVGV variant, obtained through machine learning-assisted directed evolution, demonstrated successful potential application in the scaled-up reaction, achieving an isolated yield of 40.7% and a diastereoselectivity of 91.3% for (*S*,*S*,*S*)-dihydrotetrabenazine.

ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/jacsau.4c00284.

General information, machine learning process, parameters of models, preparation of KRED, synthesis of

dihydrotetrabenazine, KRED-catalyzed reduction of tetrabenazine, diastereoselectivity of mutants, and sequences of primers (PDF)
Data set and features (XLSX)

AUTHOR INFORMATION

Corresponding Authors

Song You − *School of Life Sciences and Biopharmaceutical Sciences, Shenyang Pharmaceutical University, Shenyang 110016, People's Republic of China*; Email: yousong206@aliyun.com

Bin Qin − *Wuya College of Innovation, Shenyang Pharmaceutical University, Shenyang 110016, People's Republic of China*; orcid.org/0000-0002-9180-0550; Email: to-qinbin@163.com, binqin@syphu.edu.cn

Authors

Chenming Huang − *Wuya College of Innovation, Shenyang Pharmaceutical University, Shenyang 110016, People's Republic of China*

Li Zhang − *Wuya College of Innovation, Shenyang Pharmaceutical University, Shenyang 110016, People's Republic of China*

Tong Tang − *Wuya College of Innovation, Shenyang Pharmaceutical University, Shenyang 110016, People's Republic of China*

Haijiao Wang − *Wuya College of Innovation, Shenyang Pharmaceutical University, Shenyang 110016, People's Republic of China*

Yingqian Jiang − *Wuya College of Innovation, Shenyang Pharmaceutical University, Shenyang 110016, People's Republic of China*

Hanwen Ren − *Wuya College of Innovation, Shenyang Pharmaceutical University, Shenyang 110016, People's Republic of China*

Yitian Zhang − *Wuya College of Innovation, Shenyang Pharmaceutical University, Shenyang 110016, People's Republic of China*

Jiali Fang − *Wuya College of Innovation, Shenyang Pharmaceutical University, Shenyang 110016, People's Republic of China*

Wenhe Zhang − *School of Life Sciences and Biopharmaceutical Sciences, Shenyang Pharmaceutical University, Shenyang 110016, People's Republic of China*

Xian Jia − *School of Pharmaceutical Engineering, Shenyang Pharmaceutical University, Shenyang 110016, People's Republic of China*

Complete contact information is available at:
https://pubs.acs.org/10.1021/jacsau.4c00284

Author Contributions

C.H. and L.Z. conducted the literature search and wrote the original manuscript. S.Y. and B.Q. conceived the idea and provided financial support. B.Q. polished the final manuscript. All authors have given approval to the final version of the manuscript. CRediT: C.H. data curation, software, writing—original draft, and writing—review and editing; L.Z. data curation and writing—original draft; T.T. data curation; H.W. writing—original draft; Y.J. writing—original draft; H.R. writing—original draft; Y.Z. writing—original draft; J.F. writing—review and editing; W.Z. writing—review and editing; X.J. writing—review and editing; S.Y. writing—review and editing; and B.Q. writing—review and editing and methodology. CRediT: **Chenming Huang** data curation, software, writing-original draft, writing-review & editing; **Li Zhang** data curation, writing-original draft; **Tong Tang** data curation; **Haijiao Wang** writing-original draft; **Yingqian Jiang** writing-original draft; **Hanwen Ren** writing-original draft; **Yitian Zhang** writing-original draft; **Jiali Fang** writing-review & editing; **Wen-He Zhang** writing-review & editing; **Xian Jia** writing-review & editing; **Song You** writing-review & editing; **Bin Qin** funding acquisition, methodology, writing-review & editing.

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

This work was financially supported by the National Key R&D Program of China (2021YFC2102000) and the National Natural Science Foundation of China (32171273, 31971207).

REFERENCES

(1) Bell, E. L.; Finnigan, W.; France, S. P.; Green, A. P.; Hayes, M. A.; Hepworth, L. J.; Lovelock, S. L.; Niikura, H.; Osuna, S.; Romero, E.; Ryan, K. S.; Turner, N. J.; Flitsch, S. L. Biocatalysis. *Nat. Rev. Methods Primers* **2021**, *1*, 1−21.

(2) Wu, S.; Snajdrova, R.; Moore, J. C.; Baldenius, K.; Bornscheuer, U. T. Biocatalysis: Enzymatic Synthesis for Industrial Applications. *Angew. Chem., Int. Ed.* **2021**, *60*, 88−119.

(3) Huffman, M. A.; Fryszkowska, A.; Alvizo, O.; Borra-Garske, M.; Campos, K. R.; Canada, K. A.; Devine, P. N.; Duan, D.; Forstater, J. H.; Grosser, S. T.; et al. Design of an in vitro biocatalytic cascade for the manufacture of islatravir. *Science* **2019**, *366*, 1255−1259.

(4) Duboc, C.; Flitsch, S. L. Drug Discovery and Development. *JACS Au* **2024**, *4*, 276−278.

(5) Casamajo, A. R.; Yu, Y. Q.; Schnepel, C.; Morrill, C.; Barker, R.; Levy, C. W.; Finnigan, J.; Spelling, V.; Westerlund, K.; Petchey, M.; et al. Biocatalysis in Drug Design: Engineered Reductive Aminases (RedAms) Are Used to Access Chiral Building Blocks with Multiple Stereocenters. *J. Am. Chem. Soc.* **2023**, *145*, 22041−22046.

(6) Bornscheuer, U. T.; Huisman, G. W.; Kazlauskas, R. J.; Lutz, S.; Moore, J. C.; Robins, K. Engineering the third wave of biocatalysis. *Nature* **2012**, *485*, 185−194.

(7) Huisman, G. W.; Liang, J.; Krebber, A. Practical chiral alcohol manufacture using ketoreductases. *Curr. Opin. Chem. Biol.* **2010**, *14*, 122−129.

(8) Arnold, F. H. Innovation by Evolution: Bringing New Chemistry to Life (Nobel Lecture). *Angew. Chem., Int. Ed.* **2019**, *58*, 14420−14426.

(9) Leung, D. W.; Chen, E.; Goeddel, D. V. A Method for Random Mutagenesis of a Defined DNA Segment Using a Modified Polymerase Chain Reaction. *Technique* **1989**, *1*, 11−15.

(10) Cadwell, R. C.; Joyce, G. F. Randomization of genes by PCR mutagenesis. *Genome Res.* **1992**, *2*, 28−33.

(11) Hogrefe, H. H.; Cline, J.; Youngblood, G. L.; Allen, R. M. Creating randomized amino acid libraries with the QuikChange multi site-directed mutagenesis kit. *Biotechniques* **2002**, *33*, 1158−1165.

(12) Zeymer, C.; Zschoche, R.; Hilvert, D. Optimization of enzyme mechanism along the evolutionary trajectory of a computationally designed (retro-) aldolase. *J. Am. Chem. Soc.* **2017**, *139*, 12541−12549.

(13) Sandoval, B. A.; Meichan, A. J.; Hyster, T. K. Enantioselective hydrogen atom transfer: discovery of catalytic promiscuity in favin-dependent 'ene'-reductases. *J. Am. Chem. Soc.* **2017**, *139*, 11313−11316.

(14) Wu, L. J.; An, J. H.; Jing, X. R.; Chen, C.-C.; Dai, L. H.; Xu, Y.; Liu, W. D.; Guo, R.-T.; Nie, Y. Molecular Insights into the Regioselectivity of the Fe(II)/2-Ketoglutarate-Dependent Dioxyge-

nase-Catalyzed C−H Hydroxylation of Amino Acids. *ACS Catal.* **2022**, *12*, 11586−11596.

(15) Zhang, J. C.; Zhou, J. Y.; Xu, G. C.; Ni, Y. Stereodivergent evolution of *KpADH* for the asymmetric reduction of diaryl ketones with *para*-substituents. *Mol. Catal.* **2022**, *524*, No. 112315.

(16) Zhou, Q.; Chin, M.; Fu, Y.; Liu, P.; Yang, Y. Stereodivergent atom-transfer radical cyclization by engineered cytochromes P450. *Science* **2021**, *374*, 1612−1616.

(17) Xu, J.; Cen, Y. X.; Singh, W.; Fan, J. J.; Wu, L.; Lin, X. F.; Zhou, J. H.; Huang, M. L.; Reetz, M. T.; Wu, Q. Stereodivergent Protein Engineering of a Lipase To Access All Possible Stereoisomers of Chiral Esters with Two Stereocenters. *J. Am. Chem. Soc.* **2019**, *141*, 7934−7945.

(18) Zhang, H. L.; Chen, X.; Lv, T.; Li, Q.; Liu, W. D.; Feng, J. H.; Liu, X. T.; Yao, P. Y.; Wu, Q. Q.; Zhu, D. M. Engineering a Carbonyl Reductase to Simultaneously Increase Activity Toward Bulky Ketone and Isopropanol for Dynamic Kinetic Asymmetric Reduction via Enzymatic Hydrogen Transfer. *ACS Catal.* **2023**, *13*, 9960−9968.

(19) Kua, G. K. B.; Nguyen, G. K. T.; Li, Z. Enzyme Engineering for High-Yielding Amide Formation: Lipase-Catalyzed Synthesis of *N*-Acyl Glycines in Aqueous Media. *Angew. Chem.* **2023**, *135*, No. e202217878.

(20) Wu, Z.; Kan, S. B. J.; Lewis, R. D.; Wittmann, B. J.; Arnold, F. H. Machine learning-assisted directed protein evolution with combinatorial libraries. *Proc. Natl. Acad. Sci. U.S.A.* **2019**, *116*, 8852−8858.

(21) Bedbrook, C. N.; Yang, K. K.; Rice, A. J.; Gradinaru, V.; Arnold, F. H. Machine learning to design integral membrane channelrhodopsins for efficient eukaryotic expression and plasma membrane localization. *PLoS Comput. Biol.* **2017**, *13*, No. e1005786.

(22) Mazurenko, S.; Prokop, Z.; Damborsky, J. Machine Learning in Enzyme Engineering. *ACS Catal.* **2020**, *10*, 1210−1223.

(23) Saito, Y.; Oikawa, M.; Nakazawa, H.; Niide, T.; Kameda, T.; Tsuda, K.; Umetsu, M. Machine-Learning-Guided Mutagenesis for Directed Evolution of Fluorescent Proteins. *ACS Synth. Biol.* **2018**, *7*, 2014−2022.

(24) Yang, K. K.; Wu, Z.; Arnold, F. H. Machine-learning-guided directed evolution for protein engineering. *Nat. Methods* **2019**, *16*, 687−694.

(25) Alley, E. C.; Khimulya, G.; Biswas, S.; AlQuraishi, M.; Church, G. M. Unified rational protein engineering with sequence-based deep representation learning. *Nat. Methods* **2019**, *16*, 1315−1322.

(26) Li, G. Y.; Dong, Y. J.; Reetz, M. T. Can Machine Learning Revolutionize Directed Evolution of Selective Enzymes? *Adv. Synth. Catal.* **2019**, *361*, 2377−2386.

(27) Xu, G. C.; Dou, Z.; Chen, X. Z.; Zhu, L. D.; Zheng, X. Y.; Chen, X. Y.; Xue, J. Y.; Niwayama, S.; Ni, Y. Enhanced stereodivergent evolution of carboxylesterase for efficient kinetic resolution of near-symmetric esters through machine learning. *Res. Square* **2024**, DOI: 10.21203/rs.3.rs-3897762/v1.

(28) Malca, S. H.; Duss, N.; Meierhofer, J.; Patsch, D.; Niklaus, M.; Reiter, S.; Hanlon, S. P.; Wetzl, D.; Kuhn, B.; Iding, H.; Buller, R. Effective engineering of a ketoreductase for the biocatalytic synthesis of an ipatasertib precursor. *Commun. Chem.* **2024**, *7*, 46.

(29) Rives, A.; Meier, J.; Sercu, T.; Goyal, S.; Lin, Z. M.; Liu, J.; Guo, D. M.; Ott, M.; Zitnick, C. L.; Ma, J.; Fergus, R. Biological Structure and Function Emerge from Scaling Unsupervised Learning to 250 Million Protein Sequences. *Proc. Natl. Acad. Sci. U.S.A.* **2021**, *118*, No. e2016239118.

(30) Ma, E. J.; Siirola, E.; Moore, C.; Kummer, A.; Stoeckli, M.; Faller, M.; Bouquet, C.; Eggimann, F.; Ligibel, M.; Huynh, D.; et al. Machine-Directed Evolution of an Imine Reductase for Activity and Stereoselectivity. *ACS Catal.* **2021**, *11*, 12433−12445.

(31) Xu, Y. T.; Deeptak, V.; Sheridan, R. P.; Liaw, A.; Ma, J. S.; Marshall, N. M.; Mcintosh, J.; Sherer, E. C.; Svetnik, V.; Johnston, J. M. Deep Dive into Machine Learning Models for Protein Engineering. *J. Chem. Inf. Model.* **2020**, *60*, 2773−2790.

(32) Wittmann, B. J.; Yue, Y. S.; Arnold, F. H. Informed Training Set Design Enables Efficient Machine Learning-Assisted Directed Protein Evolution. *Cell Syst.* **2021**, *12*, 1026−1045.

(33) Gruver, N.; Stanton, S.; Kirichenko, P.; Finzi, M.; Maffettone, P.; Myers, V.; Delaney, E.; Greenside, P.; Wilson, A. G. Effective Surrogate Models for Protein Design with Bayesian Optimization In *Proceedings of the 2021 ICML Workshop of Computational Biology*, Virtual, July 24, 2021ICML, 2021; Paper 61. https://icml-compbio.github.io/2021/papers/WCBICML2021_paper_61.pdf.

(34) Hawkins-Hooker, A.; Depardieu, F.; Baur, S.; Couairon, G.; Chen, A.; Bikard, D. Generating Functional Protein Variants with Variational Autoencoders. *PLOS Comput. Biol.* **2021**, *17*, No. e1008736.

(35) Giessel, A.; Dousis, A.; Ravichandran, K.; Smith, K.; Sur, S.; McFadyen, L.; Zheng, W.; Licht, S. Therapeutic Enzyme Engineering Using a Generative Neural Network. *Sci. Rep.* **2022**, *12*, No. 1536.

(36) Stanton, S.; Maddox, W.; Gruver, N.; Maffettone, P.; Delaney, E.; Greenside, P.; Wilson, A. G.Accelerating Bayesian Optimization for Biological Sequence Design with Denoising Autoencoders. In *Proceedings of Machine Learning Research*; PMLR, 2022; pp 20459−20478.

(37) Johnson, S. R.; Fu, X. Z.; Viknander, S.; Goldin, C.; Monaco, S.; Zelezniak, A.; Yang, K. K. Computational Scoring and Experimental Evaluation of Enzymes Generated by Neural Networks. *bioRxiv* **2023**, 1−10, DOI: 10.1101/2023.03.04.531015.

(38) Greenhalgh, J. C.; Fahlberg, S. A.; Pfleger, B. F.; Romero, P. A. Machine Learning-Guided Acyl-ACP Reductase Engineering for Improved in Vivo Fatty Alcohol Production. *Nat. Commun.* **2021**, *12*, No. 5825.

(39) Hu, R. Y.; Fu, L. H.; Chen, Y. C.; Chen, J. Y.; Qiao, Y.; Si, T. Protein Engineering via Bayesian Optimization-Guided Evolutionary Algorithm and Robotic Experiments. *Brief. Bioinform.* **2023**, *24*, No. bbac570.

(40) Cui, Y. L.; Chen, Y. C.; Sun, J. Y.; Zhu, T.; Pang, H.; Li, C. L.; Geng, W. C.; Wu, B. Computational redesign of a hydrolase for nearly complete PET depolymerization at industrially relevant high-solids loading. *Nat. Commun.* **2024**, *15*, No. 1417.

(41) Ao, Y. F.; Pei, S. X.; Xiang, C.; Menke, M. J.; Shen, L.; Sun, C. H.; Dörr, M.; Born, S.; Höhne, M.; Bornscheuer, U. T. Structure- and Data-Driven Protein Engineering of Transaminases for Improving Activity and Stereoselectivity. *Angew. Chem., Int. Ed.* **2023**, *62*, No. e202301660.

(42) Voutilainen, S.; Heinonen, M.; Andberg, M.; Jokinen, E.; Maaheimo, H.; Pääkkönen, J.; Hakulinen, N.; Rouvinen, J.; Lähdesmäki, H.; Kaski, S.; et al. Substrate Specificity of 2-Deoxy-DRibose 5-Phosphate Aldolase (DERA) Assessed by Different Protein Engineering and Machine Learning Methods. *Appl. Microbiol. Biotechnol.* **2020**, *104*, 10515−10529.

(43) Citrome, L. Valbenazine for tardive dyskinesia: A systematic review of the efficacy and safety profile for this newly approved novel medication-What is the number needed to treat, number needed to harm and likelihood to be helped or harmed? *Int. J. Clin. Pract.* **2017**, *71*, No. e12964.

(44) Flick, A. C.; Leverett, C. A.; Ding, H. X.; Mcinturff, E.; Fink, S. J.; Helal, C. J.; O'donnell, C. J. Synthetic Approaches to the New Drugs Approved During 2017. *J. Med. Chem.* **2019**, *62*, 7340−7382.

(45) Boldt, K. G.; Biggers, M. S.; Phifer, S. S.; Brine, G. A.; Rehder, K. S. Synthesis of (+)- and (−)-Tetrabenazine from the Resolution of α-Dihydrotetrabenazine. *Synth. Commun.* **2009**, *39*, 3574−3585.

(46) Fang, J. L.; Xu, Y. P.; Ren, H. W.; Huang, C. M.; Zhang, W. H.; Jia, X.; You, S.; Qin, B. A Three-Step Chemoenzymatic Cascade Synthesis of Miconazole Analogues Based on the Asymmetric Synthesis of β-Heteroaryl Amino Alcohols via Ketoreductases. *Adv. Synth. Catal.* **2023**, *365*, 4181−4189.

(47) Fox, R.; Roy, A.; Govindarajan, S.; Minshull, J.; Gustafsson, C.; Jones, J. T.; Emig, R. Optimizing the search algorithm for protein engineering by directed evolution. *Protein Eng. Des. Sel.* **2003**, *16*, 589−597.

(48) Hamp, T.; Rost, B. More challenges for machine-learning protein interactions. *Bioinformatics* **2015**, *31*, 1521−1525.

(49) Garg, A.; Raghava, G. P. S. A machine learning based method for the prediction of secretory proteins using amino acid composition, their order and similarity-search. *In Silico Biol.* **2008**, *8*, 129−140.

(50) Kuhn, L. A.; Swanson, C. A.; Pique, M. E.; Tainer, J. A.; Getzoff, E. D. Atomic and Residue Hydrophilicity in the Context of Folded Protein Structures. *Proteins* **1995**, *23*, 536−547.

(51) Zimmerman, J. M.; Eliezer, N.; Simha, R. The Characterization of Amino Acid Sequencesin Proteins by Statistical Methods. *J. Theor. Biol.* **1968**, *21*, 170−201.

(52) Fauchère, J. L.; Charton, M.; Kier, L. B.; Verloop, A.; Pliska, V. Amino acid side chain parameters for correlation studies in biology and pharmacology. *Int. J. Pept. Protein Res.* **1988**, *32*, 269−278.