JACS Au
AN OPEN ACCESS JOURNAL OF THE AMERICAN CHEMICAL SOCIETY

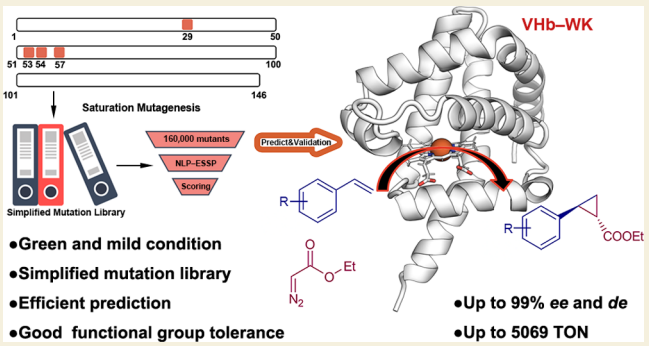
pubs.acs.org/jacsau
Article

Machine-Learning-Aided Engineering Hemoglobin as Carbene Transferase for Catalyzing Enantioselective Olefin Cyclopropanation

Hanqing Xie,[#] Kaifeng Liu,[#] Zhengqiang Li, Zhi Wang, Chunyu Wang, Fengxi Li,* Weiwei Han,* and Lei Wang*

Cite This: *JACS Au* 2024, 4, 4957−4967
Read Online
ACCESS | Metrics & More | Article Recommendations | Supporting Information

ABSTRACT: In this study, we developed a machine-learning-aided protein design strategy for engineering *Vitreoscilla* hemoglobin (VHb) as carbene transferase. A Natural Language Processing (NLP) model was used for the first time to construct an algorithm (EESP, enzyme enantioselectivity score predictor) and predict the enantioselectivity of VHb. We identified critical amino acid residue sites by molecular docking and established a simplified mutation library by site-saturated mutagenesis. Based on the simplified mutant library, the trianed EESP scored 160,000 virtual mutants, and 15 predicted high-score mutants were chosen for experimental validation. Among these mutants, VHb-WK (Y29W/P54K) demonstrated the highest diastereoselectivity and enantioselectivity of carbene transferase for the olefin cyclopropanation in aqueous conditions. Subsequently, molecular dynamics simulations were performed to explore the interaction between protein and substrates, finding that the high enantioselectivity of VHb-WK stems from the interactions of R47, Q53, and K84, which narrows the entrance of the enzyme's pocket, favoring the restriction of the formation of reaction intermediates. Integrating the NLP model and enzyme modification offers significant advantages by reducing economic costs and workloads associated with the protein engineering process.



KEYWORDS: *vitreoscilla hemoglobin, machine learning, cyclopropanation, carbene transferase, molecular dynamics simulations*

INTRODUCTION

The development of biocatalytic strategies as a more sustainable alternative to traditional organic catalysis is eliciting increased research interest in the academia and industry.[1−3] Remarkably, many engineered enzymes can catalyze biological reactions that proved to be difficult even for some of the most sophisticated organic catalysts.[4−6] Carbene transferases are engineered heme-dependent proteins capable of transferring carbene moieties to different target compounds.[7] Hemoproteins such as cytochrome, peroxidase, myoglobin, and hemoglobin can catalyze many carbene-mediated formations, including cyclopropanation,[8−13] X-H insertion,[14−19] and ylide formation.[20−22]

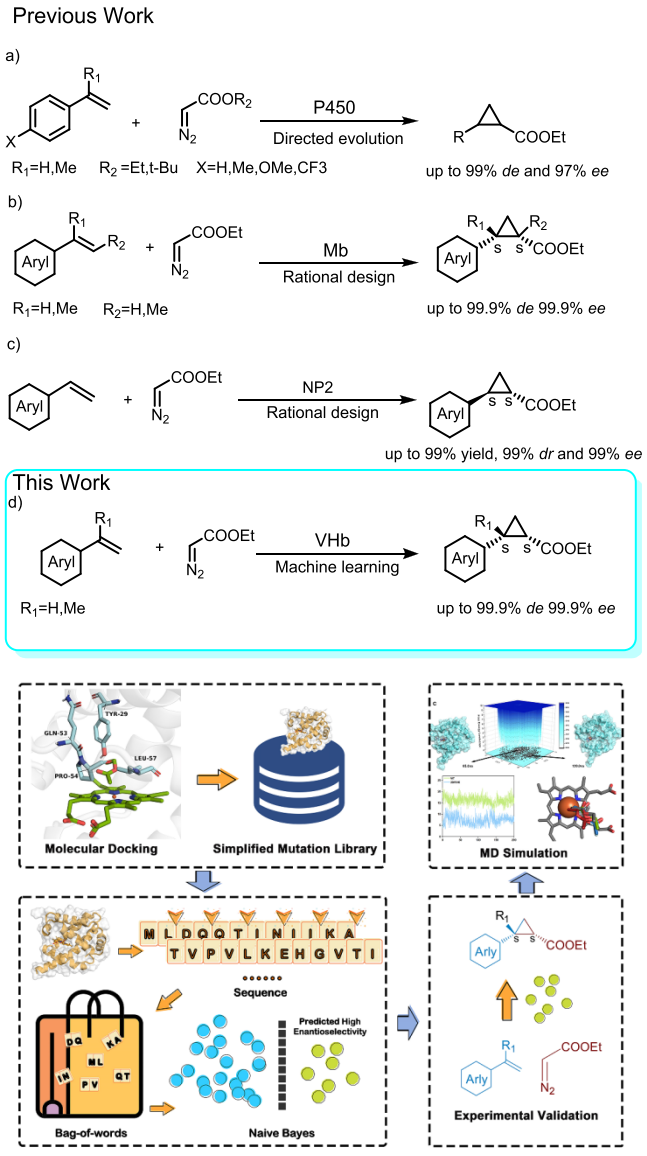
Cyclopropane moieties are versatile intermediates for a range of useful ring-opening transformations. They are commonly found in natural products and bioactive compounds.[23−25] Accordingly, the demand for cyclopropane products is significant, and extensive efforts have been devoted to developing synthesis methods for them.[26−28] Olefin cyclopropanation is gaining considerable interest in biocatalyst development, leading to the development of convenient and mild techniques for the construction of these valuable compounds. Recently, we have developed an efficient olefin cyclopropanation catalyzed by wild-type *Vitreoscilla* hemoglobin (VHb) by using *in situ* generated diazoacetonitrile.[11] Our approach exhibited exceptional stereoselectivity with diastereomeric excess (*de*) and enantiomeric excess (*ee*) of product up to 99.9%. This intriguing result demonstrated that VHb can be an ideal carbene transferase, similar to other hemoproteins. It also inspired us to further engineer VHb for catalyzing olefin cyclopropanations with other diazo reagents as well as to expand the practical applications of VHb in organic synthesis.

Diazo compounds are versatile building synthons in organic synthesis.[19,29−31] From this class of compounds, ethyl diazoacetate (EDA) is a significant reagent and frequently used in practical application.[14−16] Thus, the cyclopropanation of aromatic olefins and EDA is considered to be the model system for validating new cyclopropanation biocatalysts.[12] Arnold and co-workers were the pioneers in biocatalytic olefin cyclopropanation, utilizing engineered $P450_{BM3}$ (Scheme 1a).[8]

Received: November 4, 2024
Revised: November 13, 2024
Accepted: November 14, 2024
Published: November 21, 2024

https://doi.org/10.1021/jacsau.4c01045
JACS Au 2024, 4, 4957−4967

Scheme 1. Methods for the Study of VHb-Catalyzed Cyclopropanation Reaction



Fasan's group also rationally designed a series of myoglobin (Mb) variants that can promote the reaction with excellent stereoselectivity properties (Scheme 1b).[9] Liao et al. expanded the range of enabling protein scaffolds from mostly α-helical to β-barrel by establishing nitric oxide transport hemoprotein Nitrophorin 2 (NP2) as a new promiscuous heme enzyme for the cyclopropanation of styrenes with ethyl diazoacetate (Scheme 1c).[10]

Directed evolution is an important method for enzyme engineering.[7,8,32] To realize the intended catalytic effect, the utilization of directed evolution necessitates the generation of thousands of different variants.[33−35] Thus, it may not be feasible for all enzymes and reaction. Machine learning offers a time-saving alternative with its capacity to derive rules from extensive data. It has proven to be beneficial in enzyme modification and understanding enzyme−substrate interactions. For instance, the 3D self-supervised CNN (Convolutional Neural Networks), MutCompute, identified mutations enhancing the stability and catalytic efficiency of PET enzyme mutants.[36] Similarly, a BaggingTree model predicted the activity of PylRS mutants, leading to the discovery of new NCAA (Noncanonical amino acids) substrates and thus broadening their substrate spectrum.[37] In protein stability, machine-learning algorithms like FireProt and PROSS outperformed traditional mutation strategies.[38] A transformer model has also accurately predicted enzyme−substrate pairs, demonstrating its superiority over family specific models.[39] Despite these successes, machine-learning's applicability is often confined to specific systems. Incorporating Natural Language Processing (NLP) into protein research offers a paradigm shift in how we understand and analyze proteins.[40,41] NLP, renowned for its adaptability, allows for the conceptualization of protein sequences in linguistic terms: amino acids become the "alphabet", their arrangements form "words", and the resulting structures articulate "sentences" with specific functions. This linguistic analogy underscores the transformative potential of NLP in enzymology.

In the present study, we utilized VHb mutants [obtained with an enzyme enantioselectivity score predictor (EESP)] for the biocatalytic cyclopropanation of aromatic olefins with EDA. In four steps, we successfully addressed the challenge of engineering a hemoglobin with its desired catalytic effect (Scheme 1d). First, the reaction conditions for biocatalytic cyclopropanation were optimized, and the stereoselectivity of VHb in this reaction was investigated. Second, we determined key residues in catalytic pockets by molecular docking. A simplified mutant library was constructed by modifying these residues, and the library comprised four single-site-saturated mutants. Third, by training with the simplified mutation library (77 mutants), we designed an EESP to predict the enantioselectivity of mutants, which was inspired by the bag-of-words model in NLP. Finally, we screened the high-score mutants and obtained optimal engineered VHb-WK (Y29W/P54K). We then conducted molecular dynamics simulations to determine the impact of the catalytic pocket entrance on the enantioselectivity of VHb-WK. Additionally, we highlighted the potential impact of residue characteristics and the property relationship among key residues.

RESULTS AND DISCUSSION

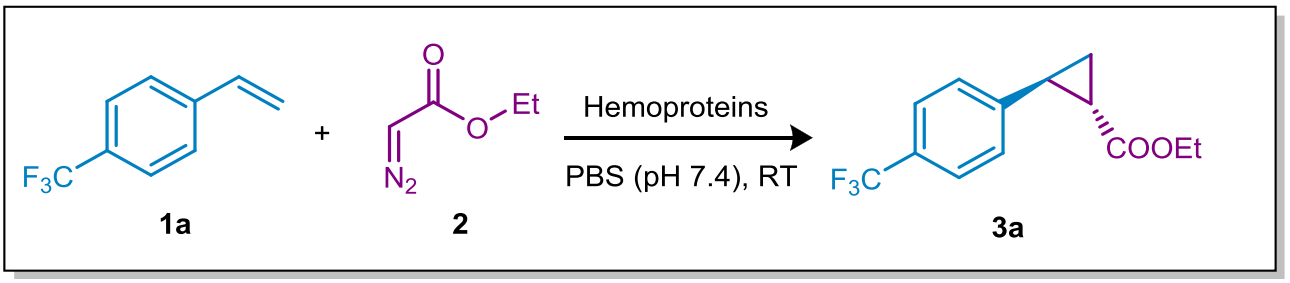
Design of the VHb Catalytic System

Fluorine-containing molecules, commonly referred to as organofluorines, have emerged as one of the most significant classes of compounds in medicinal chemistry.[42] To identify a suitable starting point for machine-learning-aided engineering of an olefin cyclopropanation enzyme, we initially adopted the cyclopropanation of 4-trifluoromethylstyrene (**1a**) with ethyl diazoacetate (EDA) (**2**) as the model reaction and screened a range of hemoproteins. Commercially available hemoproteins displayed poor yields ranging from 21.2 to 51.1% and poor enantioselectivities (0−12% *ee*) (Table 1, entries 1−5). Compared with other hemoproteins, wild-type (WT) VHb achieved a yield of 53% and 15% *ee* (Table 1, entry 6). The structures of the products were determined through extensive NMR spectroscopy analysis. The production of a single stereoisomer with the *trans*-(1*S*,2*S*) absolute configuration was confirmed by HPLC analysis (Supporting Information).

Based on these results, we aimed to engineer VHb with the objective of enhancing its enantioselectivity. Considering the results of this work and previous investigations, a reasonable mechanism for this VHb-catalyzed cyclopropanation was proposed, as shown in Scheme 2a. The reaction was postulated to proceed via the formation of an iron carbenoid intermediate

https://doi.org/10.1021/jacsau.4c01045
JACS Au 2024, 4, 4957−4967

Table 1. Hemoproteins Catalyzed Cyclopropanation of 4-Trifluoromethyl Styrene with Ethyl Diazoacetate[a]



entry	catalyst	yield (%)	turnover number (TON)	*de* %	*ee* %
1	Hb (from porcine blood)	38.2	1960	99	15
2	Mb (from horse heart)	26.7	1368	99	12
3	Hb (from bovine blood)	51.1	2620	99	7
4	cytochrome C (from horse heart)	21.2	1088	99	0
5	Hb (from human blood)	47.2	2420	99	12
6	WT VHb	53.0	2714	99	15
7	VHb (Y29F)	75.6	3870	99	65
8	VHb (Y29W)	58.2	2979	99	58
9	VHb (Q53V)	45.0	2304	99	65
10	VHb (Q53E)	72.5	3710	99	64
11	VHb (P54N)	87.2	4465	99	89
12	VHb (P54F)	77.8	3983	99	85
13	VHb (P54L)	24.2	1239	99	80
14	VHb (L57D)	81.6	4178	99	57
15	VHb (L57H)	45.8	2344	99	55

[a]Reaction conditions: 5 mM 4-trifluoromethylstyrene (**1a**), 5 mM EDA (**2**), 5 mM sodium dithionite, promoting solvent: MeOH (50 μL), purified protein (0.025% mol) in 4 mL PBS (pH7.4), rt, 5 h. The ratio of enantiomers on the reaction mixture and yield of products was determined by HPLC.

when EDA was catalyzed by the heme cofactor in VHb. The subsequent interaction between 4-trifluoromethylstyrene and the iron carbenoid resulted in the formation of cyclopropanation products. The enantioselective outcome of the reaction was controlled at this stage, and the observed stereoselectivity can be influenced by steric hindrance within the active site of VHb. Based on this mechanism and the inspection of the VHb crystal structure, residues 29, 53, 54, and 57 were selected as promising targets for mutagenesis because of the proximity of these residues to the carbene intermediate α-C, with a distance within 5 Å, could have significant implications in a substrate attack (Scheme 2b). Accordingly, we constructed a mutation library targeting four specific amino acid residue sites comprising the four single-site mutants. Based on the obtained data (77 in Table S4), we carefully evaluated the catalytic performance of these mutant variants and identified several impressive mutants (Table 1, entries 7−15).[43]

Machine Learning

We organized the enantioselectivity of 77 mutants and marked the mutants with *ee* of the product less than 50% and greater than 50% with 0 and 1, respectively. We randomly extracted 56 as the training set and 14 as the independent validation set.

We borrowed the idea of the bag-of-words model in NLP and constructed the complete sequence of the mutants. We then split it within the sequence, treating a sequence as a sentence and using the amino acid length as the scanning window. We chose a bag-of-words combination with a window length of 2. Based on machine-learning methods, we constructed a polynomial naive Bayesian classifier to predict the high and low enantioselectivities of VHb catalysis.

Mutants in the data set are labeled as 0 or 1, where 0 represents an *ee* value less than 50% and 1 represents an *ee* value greater than 50%. Each mutant can be represented as a sequence of amino acids, which are processed into word vectors in the model.

We provided a sequence to be tested, which can be represented as a word vector $\langle w_1, w_2, \cdots, \text{and } w_m\rangle$ after processing. The model estimates the conditional probability $P(d|c)$ using the formula:

$$P(d|c) = (\sum_{i=1}^{m} f_i)! \prod_{i=1}^{m} \frac{P(w_i|c)^{f_i}}{f_i!}$$

Here, m represents the total number of amino acids in sequence d, w_i (i = 1,2, ⋯, m) is the i-th amino acid in sequence d, f_i is the frequency of w_i in category d, and $P(w_i|c)$ describes the conditional probability of w_i appearing in category c. It can be calculated by the following formula:

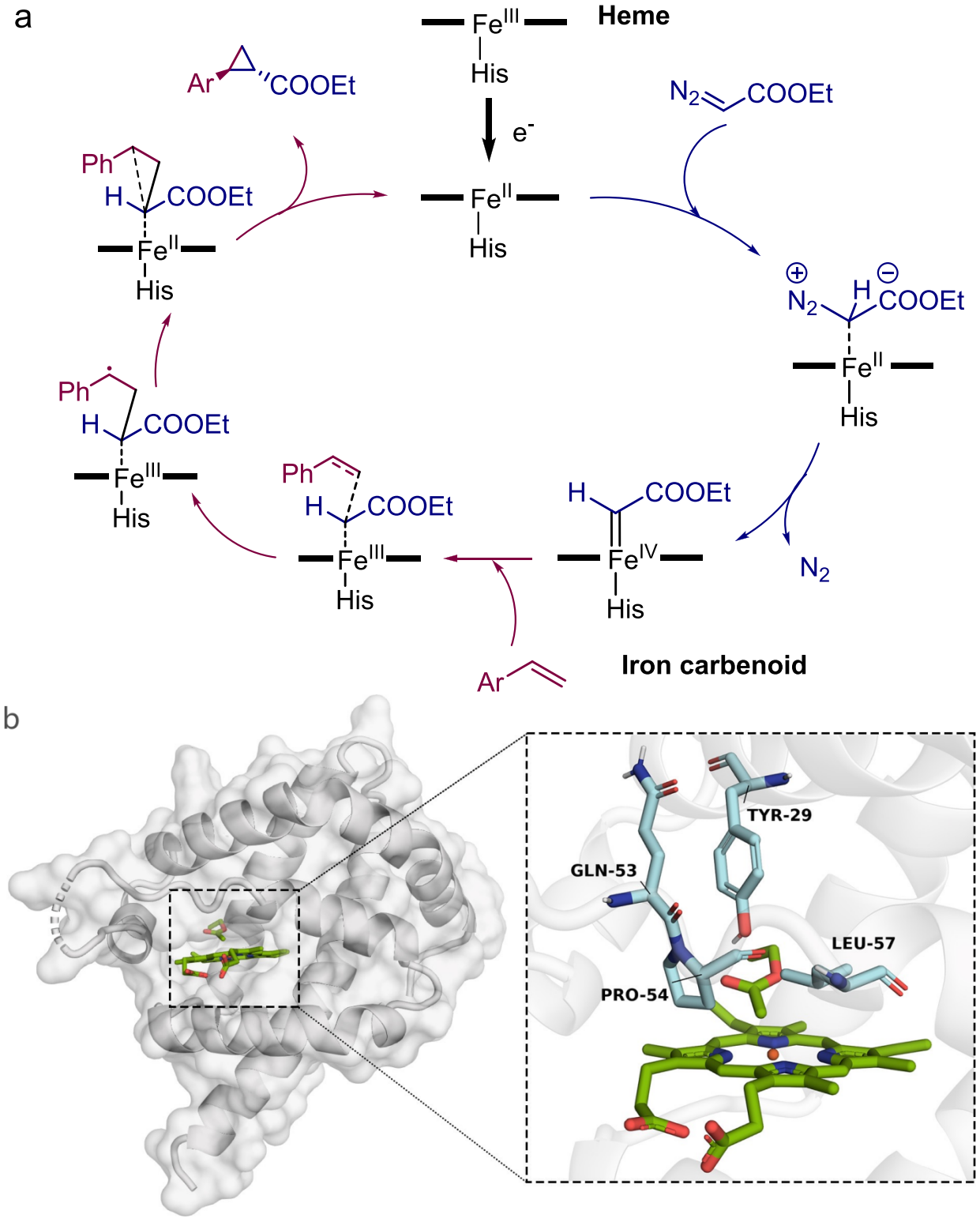
$$P(w_i|c) = \frac{\sum_{j=1}^{n} f_{ji}\,\delta(c_j, c) + 1}{\sum_{j=1}^{m}\sum_{j=1}^{n} f_{ji}\,\delta(c_j, c) + m}$$

where n represents the number of sequences in the training data set, c_j is the category the j-th sequence belongs to, m is the number of amino acids, f_{ji} is the frequency of amino acid i in the j-th sequence, and δ is a binary function. When its two parameters c_j and c are equal, it outputs 1; otherwise, it outputs 0.

The scoring value actually refers to the posterior probability $P(c|d)$, the probability that a given mutant sequence belongs to a certain category (such as an *ee* value greater than 50%). According to Bayes' theorem, this posterior probability can be expressed as

https://doi.org/10.1021/jacsau.4c01045
JACS Au 2024, 4, 4957−4967

Scheme 2. (a) Proposed Mechanism and Catalytic Steps for VHb-Catalyzed Cyclopropanation of Styrene and Ethyl Diazoacetate; (b) 3D Structure Model of the Carbenoid Intermediate (green) Docking with the Active Site of VHb (gray), the Four Amino Acid Residues Closest to the Intermediate Are 53, 54, 57, and 29 (Blue)



$$P(c|d) \propto P(d|c)P(c)$$

$P(c)$ is the prior probability of category c, which can be estimated based on the relative frequencies of each category in the training data. In this manner, each mutant is not simply classified as 0 or 1 but is assigned a probability value, indicating its likelihood of belonging to a particular category.

The prediction results of the independent validation set are shown in Figure 1. The confusion matrix is an error matrix commonly used to visually evaluate the performance of a supervised machine-learning algorithm. The ROC (Receiver

https://doi.org/10.1021/jacsau.4c01045
JACS Au 2024, 4, 4957−4967

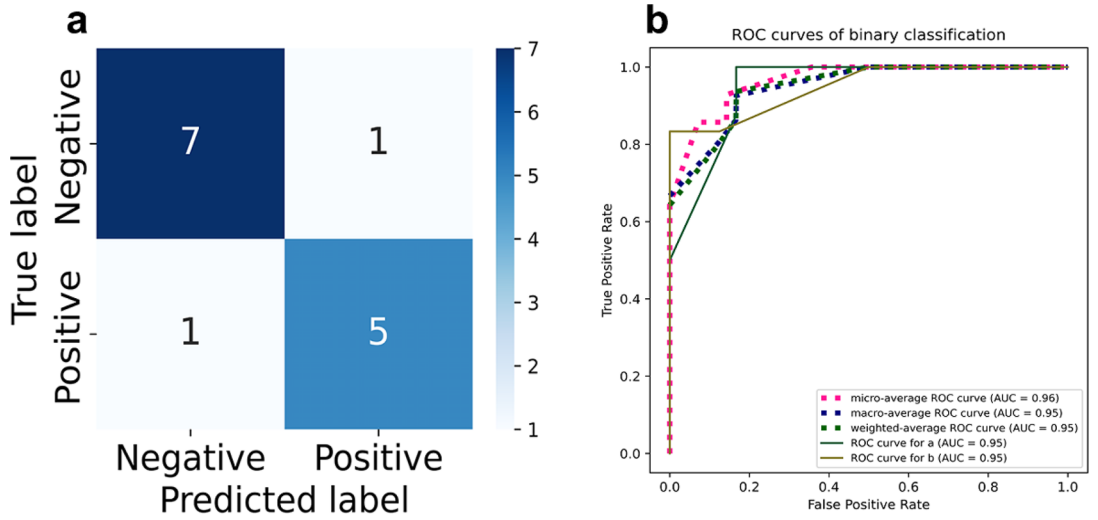


Figure 1. (a) Confusion matrix of the prediction model. (b) ROC (Receiver Operating Characteristic) curves of binary classification.

Table 2. Actual Enantioselectivity of the Predicted Mutants[a]

entry	Y29	Q53	P54	L57	score	*ee* %
1	Y29N		P54C		0.968	72
2	Y29W		P54C		0.944	81
3	Y29R		P54C		0.937	39
4	Y29W	Q53L	P54S		0.933	84
5	Y29W		P54R		0.928	86
6	Y29W		P54K		0.919	99
7	Y29W	Q53M	P54A		0.907	75
8	Y29W		P54S		0.907	91
9	Y29W	Q53M	P54N		0.882	90
10	Y29F	Q53L	P54D		0.848	81
11	Y29F	Q53M	P54D	L57H	0.848	63
12	Y29W		P54D	L57H	0.840	80
13	Y29F	Q53E	P54S	L57M	0.832	88
14	Y29F	Q53H	P54S	L57H	0.832	94
15	Y29F		P54F	L57D	0.832	71

[a]The ratio of enantiomers on the reaction mixture was determined by HPLC.

Operating Characteristic) curve plots the true positive rate against the false positive rate at various threshold settings, and the AUC (Area Under the Curve) curve provides a single value summarizing the model's ability to discriminate between classes across all thresholds.

Among the 14 samples, 12 were correctly predicted and the AUC value reached 0.95. The key performance indicator of our binary prediction is shown below, with an accuracy of 0.85. The formula for calculating the metrics is as follows, TP (True Positives) and TN (True Negatives) represent correct predictions for positive and negative classes, respectively. FP occurs when the negative class is incorrectly predicted as positive, and FN (False Negatives) occurs when the positive class is incorrectly predicted as negative.

$$\text{precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} = 0.833$$

$$\text{recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} = 0.833$$

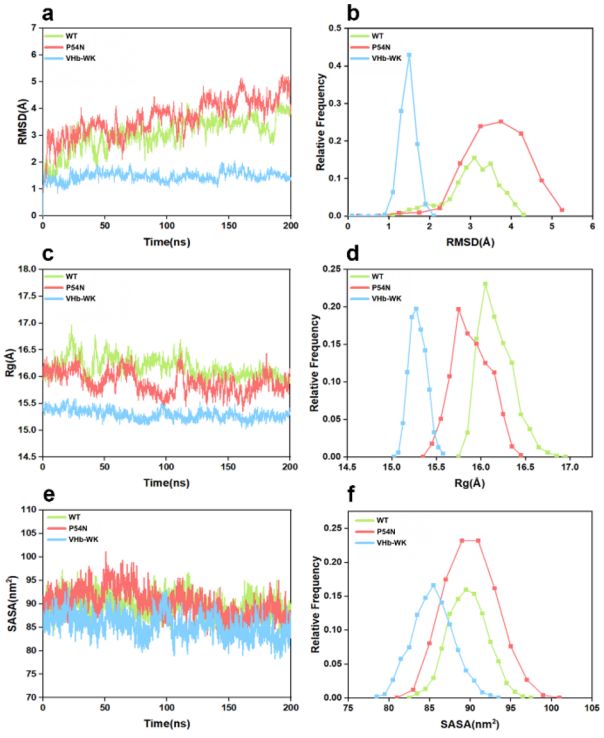


Figure 2. Analysis of structure stability. (a) Temporal evolution of the RMSDs from their initial structure of the three systems. (b) Relative Frequencies of RMSDs. (c) Radius of gyration over 200 ns MD for the three systems. (d) Relative Frequencies of radius gyration. (e) SASA over 200 ns MD. (f) Relative frequencies of SASA.

$$\text{F1s} = \frac{2 \times \text{precision} \times \text{recall}}{\text{precision} + \text{recall}} = 0.833$$

$$\text{accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} = 0.85$$

Subsequently, we predicted 160,000 proteins with 20 × 20 × 20 × 20 combinations of amino acid residues 29, 53, 54, and 57 and selected 15 mutants with the highest scores for experimental validation. The enantioselectivity result of 14 mutants aligns with the expected predictions from the model, with a maximum enantioselectivity of 99% *ee* (Table2. entry 6). The results

https://doi.org/10.1021/jacsau.4c01045
JACS Au 2024, 4, 4957−4967

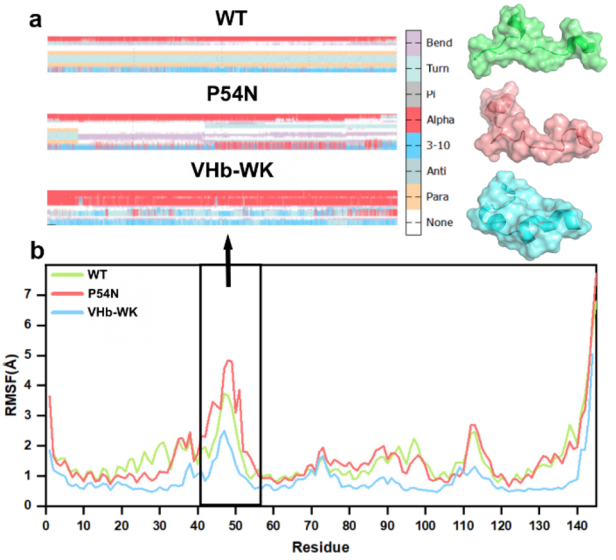


Figure 3. (a) DSSP (Dictionary of Secondary Structure of Protein) and structures comparison of region 40−60. (b) RMSFs of CA atoms in the three systems.

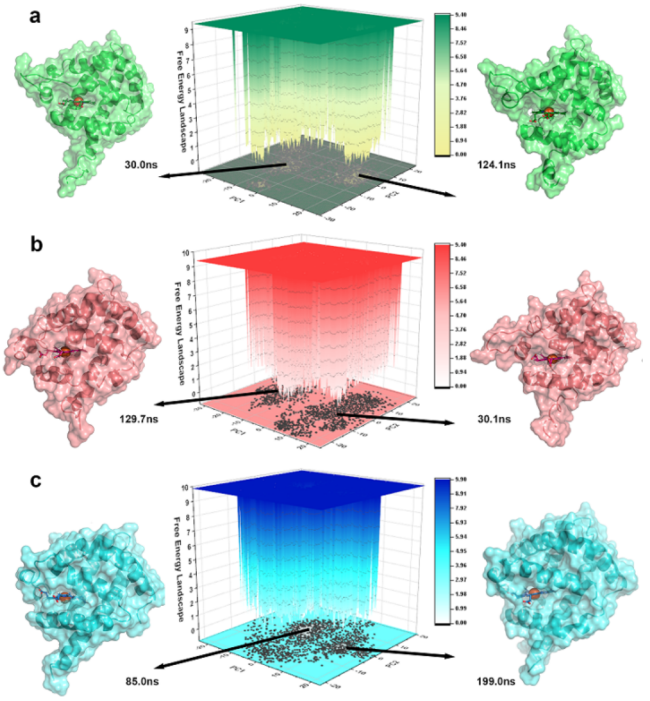


Figure 5. Free energy landscape for the following three systems: (a) WT, (b) P54N, and (c) VHb-WK. Representative conformations of the low-energy regions are displayed.

predicted by EESP offered a means to reduce redundancy in protein mutations and promote the development of the intended engineered protein. By this way, we made only 92 (77 + 15) mutants to obtain the ideal engineered carbene transferase. The machine-learning-assisted methodology achieved prediction of the partial performance of engineering hemoglobin, exhibiting comparable results to those previously reported by prominent researchers in this field.[9,44,45]

Molecular Dynamics Simulation

To gain a deeper understanding of how mutations enhance the stereoselectivity of VHb, we conducted a 200 ns molecular dynamics simulation on two selected mutants with the highest *ee*, as well as the WT. Specifically, we chose the P54N mutant, which exhibited the highest *ee* (89%) for single mutation, and the VHb-WK (the best predicted mutant, which contains the Y29W and P54K mutations) with the highest *ee* (99%), which was obtained from prediction.

Following the generation of MD trajectories, the stability of simulations was assessed by calculating the root-mean-square deviation (RMSD) of CA atoms (Figure 2a,b). In the VHb-WK system, a narrower attribution was noted, hinting at its higher stability compared to those of the other systems and less significant structural changes. Additionally, the RMSD fluctuations of the WT and P54N indicated significant conformational changes in these systems. In summary, the equilibrated 200 ns trajectories were deemed to be suitable for further analysis. The radius of gyration (R_g) value was an indicator of the overall size and compactness of the protein conformation. We found it to be lower for the VHb-WK system than for the other systems. This finding suggested a smaller volume (Figure 2c,d). The narrower attribution in the system also indicated its higher stability than that of the other systems.

The solvent-accessible surface area (SASA) was used to estimate the number of residues present in the surface regions of the protein and the number of residues that were in the hydrophobic core, which were buried. As shown in Figure 2e,f, SASA values were also found to be slightly lower for the VHb-WK system than for the others, consistent with that of the R_g

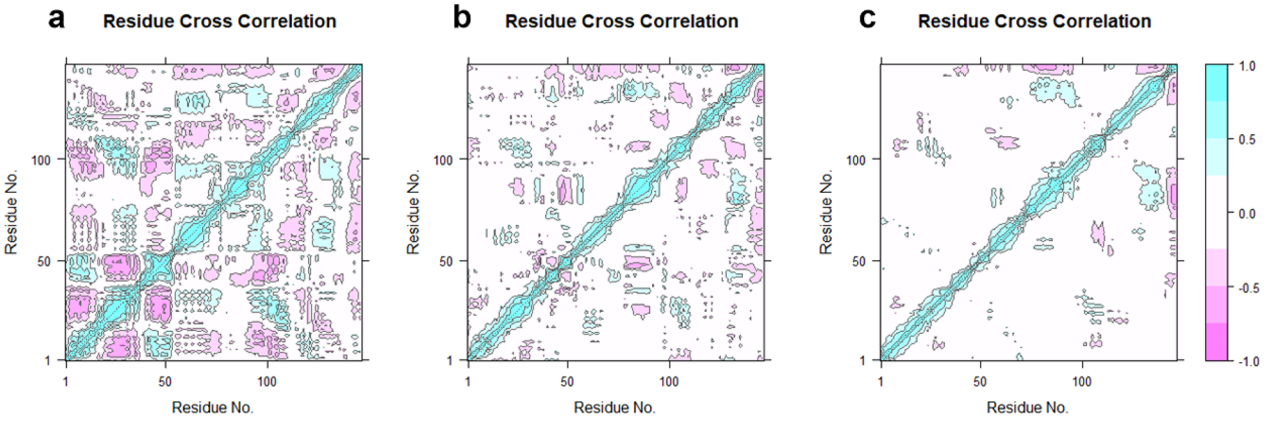


Figure 4. Dynamic cross-correlation map for the 200 ns MD simulation trajectories of the three systems: (a) WT, (b) P54N, and (c) VHb-WK.

https://doi.org/10.1021/jacsau.4c01045
JACS Au 2024, 4, 4957−4967

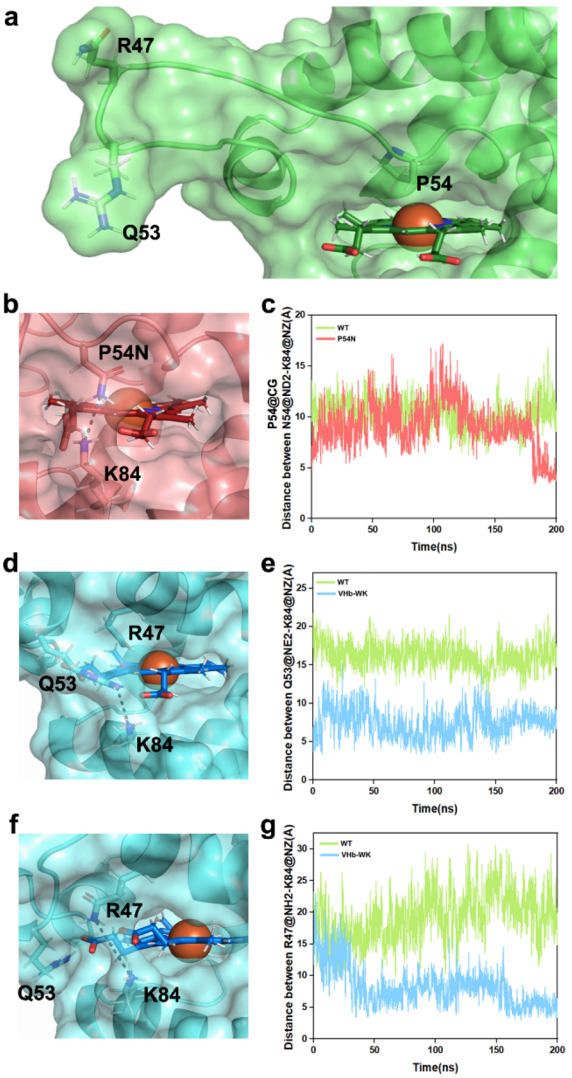


Figure 6. (a) Positions of R47, Q53, and P54 and the catalytic pocket in WT VHb. (b) Schematic diagram of the P54N pocket. The dashed line shows the distance between residues 54N and K84. (c) Distance between 54N and K84 in P54N mutant, comparing with distance between P54 and K84 in WT. (d) Schematic diagram of the VHb-WK pocket. The dashed line shows the distance between residues Q53 and K84. (e) Distance between Q53 and K84 in VHb-WK mutant, comparing with the distance in WT. (f) Schematic diagram of the VHb-WK pocket. The dashed line shows the distance between residues R47 and K84. (g) Distance between R47 and K84 in VHb-WK mutant, comparing with the distance in WT.

values. This finding also indicated that some regions of the protein were folded.

Then, we calculated the RMSF values of CA atoms for the three systems (Figure 3b) and the secondary structures of residues 40−60 (Figure 3a). VHb-WK (Y29W/P54K) exhibited a lower RMSF (root-mean-square fluctuation) than both WT and P54N especially in the 40−60 region, indicating better rigidity. In VHb-WN, region 40−60 formed stable helices, rather than turn and para formed in WT, and random coil formed in P54N. Residues 40−60 region were the upper part of the catalytic pocket, and its conformational change may affect the opening size of the catalytic pockets.

The dynamic cross-correlation map for the 200 ns MD simulation trajectories is shown in Figure 4. The positive regions are shown in cyan, and the negative regions are in pink, representing correlated and anticorrelated motions among residue CA atoms. VHb-WK had fewer internal interactions, and the protein was more stable. The RMSFs of CA atoms in the three systems.

A principal component analysis (PCA) of the CA atoms for the three systems is shown in Figure 5. This analysis revealed protein conformations in different energy wells and allowed us to observe distinct pocket inlet shapes. In the WT VHb, the entrances to the active pockets were wide open in both representative conformations. In the P54N mutant, one representative conformation had an open entrance, whereas the other was partially closed. In the VHb-WK mutant, the entrances to the active pockets in both of the representative conformations decreased.

Upon further analysis of the pockets (Figure 6), we discovered that in the P54N, 54N can interact with K84, bridging the upper and lower parts of the pocket and thus reducing the size of the opening. In the VHb-WK mutant, the two residues, R47 and Q53, can also interact with K84, bridging the upper and lower parts of the pocket, which resulted in a smaller opening. However, distance analysis revealed that 54N in the P54N mutant cannot stably bridge with K84 whereas R47 and Q53 in the VHb-WK mutant were stably close to K84 and exhibited alternating states. This finding suggested that among the two residues, at least one always interacted with K84 to reduce the size of the pocket entrance. Analysis of the pockets may explain the relatively high spatial enantioselectivity observed in both mutants. The connection between the upper and lower parts of the catalytic pocket prevented carbenes from directly accessing the pocket, thereby limiting the direction that the carbenes can approach. This finding results in the creation of carbene intermediates in a specific orientation, contributing to the *ee* of

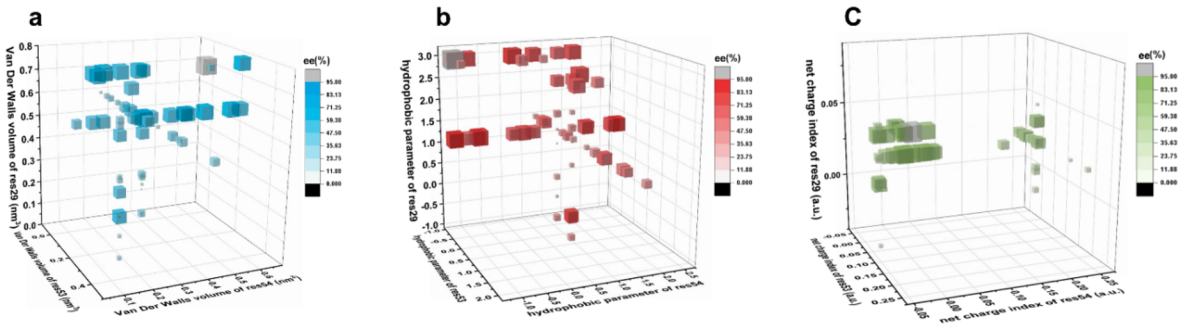
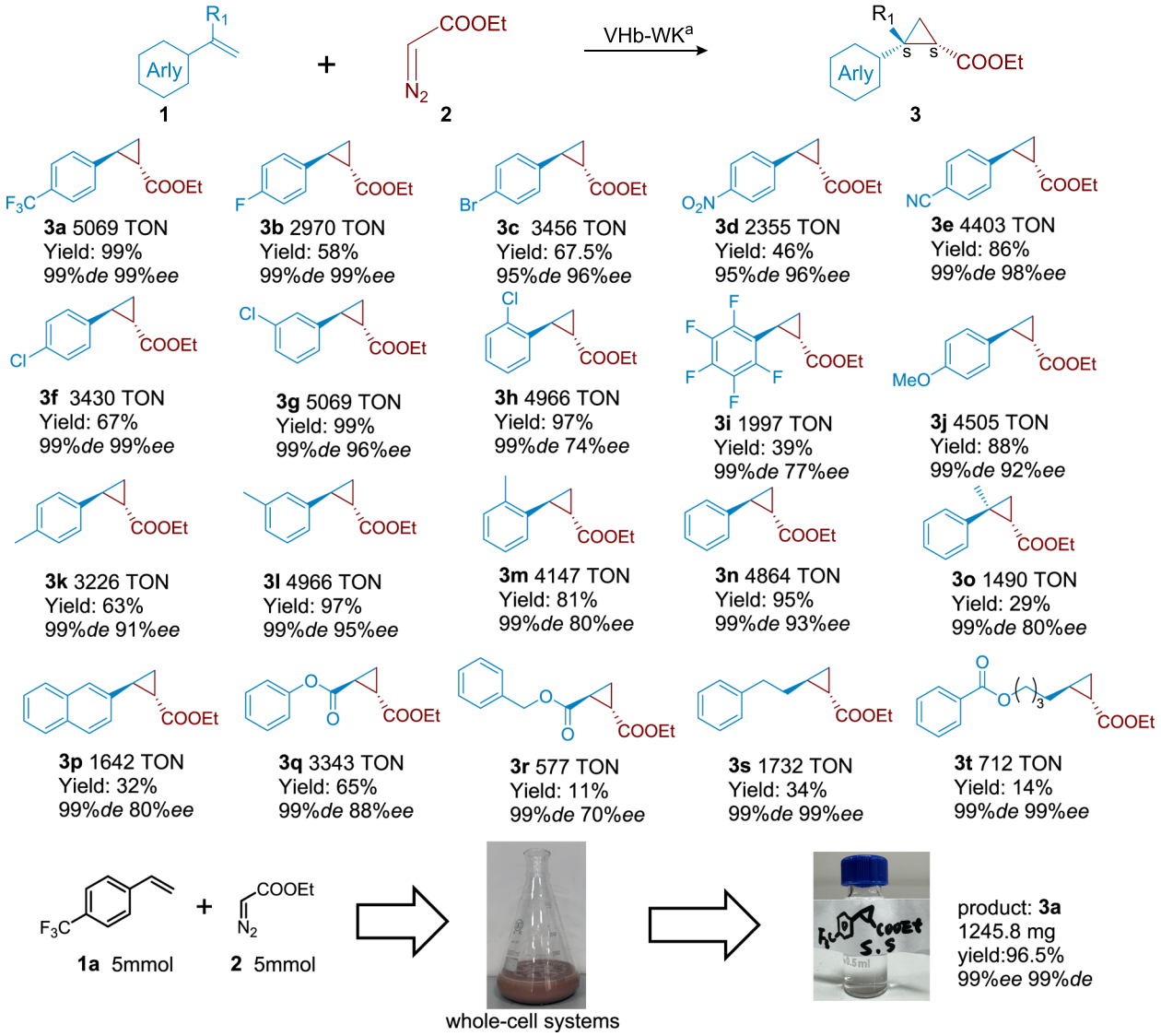


Figure 7. Analysis of the impact of three residue characteristics on the *ee*. (a) van der Waals volume. (b) Net charge index. (c) Hydrophobic parameters.

https://doi.org/10.1021/jacsau.4c01045
JACS Au 2024, 4, 4957−4967

Table 3. Cyclopropanation of EDA with Olefin-Derived Reagents to Explore the Substrate Scope[a]



[a]Reaction conditions: 5 mM olefin-derived, 5 mM EDA (**2**), 5 mM sodium dithionite, promoting solvent: MeOH (30%), VHb-WK (0.025% mol) in 4 mL PBS (pH7.4), rt, 5 h. The ratio of enantiomers on the reaction mixture and yield of products was determined by HPLC.

the reaction products. The reason for the VHb-WK mutant exhibiting higher enantioselectivity than the P54N mutant may be the more stable connection in the VHb-WK mutant, which more effectively limits the carbenes in approaching from predetermined directions.

van der Waals volume, net charge index, and hydrophobic parameters of residues were selected from many characteristics of amino acid residues that have been proven to be key factors affecting interactions among such residues.[43,42] Due to the relatively small contribution of the mutation at residue 57 to the change in *ee* and the lack of significant differences between different mutations, residues 29, 53, and 54 were chosen for analysis. As shown in Figure 7, the spatial coordinate of each mutant was the property mapping of three sites, and the size, color of the squares was mapped by *ee*. The increased van der Waals volume and hydrophobicity for the side chain of residue 29 were beneficial for improving enantioselectivity. This finding may have been due to the larger steric hindrance and stronger hydrophobicity of residue 29, which favored the compression of helix 29 above the catalytic pocket and facilitated the closure of the pocket. The side chain of residue 53 was not conducive to enantioselectivity if it was too large or too small, and the optimal value was around 0.32 nm^3. The larger side chain of residue 54 was beneficial for it. The same charge of residues 53 and 54 was not beneficial for enantioselectivity possibly due to mutual repulsion affecting bridging above and below the pocket, whereas their hydrophilicity may be beneficial.

Substrate Scope of Olefin

The effects of various factors such as the reaction time and substrate ratio were explored (Supporting Information). Under optimal conditions, the substrate scope of the VHb variant was investigated by subjecting various styrene derivatives and other olefin substrates to VHb-WK catalyzed cyclopropanation in the presence of EDA, successfully achieving the synthesis of desired products **3** with synthetically valuable yields and exceptional diastereo- and enantioselectivities (up to 99% *de* and 99% *ee*; Table 3). Remarkably, utilizing this mutant variant enabled various styrene derivatives including ortho-, meta-, and para-substituted analogs to be efficiently converted into the desired

https://doi.org/10.1021/jacsau.4c01045
JACS Au 2024, 4, 4957−4967

cyclopropanation products (**3a**−**3o**) with notable turnover numbers (TONs) (1490−5069) and high levels of diastereoselectivity and enantioselectivity (95−99% *de* and 74−99% *ee*) in closed-vessel reactions. Notably, the biocatalyst exhibited tolerance toward both electron-withdrawing (**1a**−**1i**) and electron-donating (**1j**−**1m**) groups on the phenyl ring. Moreover, α-methylstyrene (**1o**) underwent conversion and exhibited excellent diastereoselectivity and enantioselectivity (99% *de* and 80% *ee*). Intriguingly, using the fused alkene naphthalene ethylene (**1p**) as a substrate resulted in significant enantioselectivity (80% *ee*). VHb also demonstrated outstanding enantioselectivity (70−99% *ee*) when unactivated alkenes (**1q**−**1t**) were tested. Although relatively lower TONs values were obtained with **3r** and **3t**, these findings underscored the broad substrate scope of VHb in the context of cyclopropanation reactions. We subsequently assessed the compatibility of the whole-cell system for this reaction. Under the whole-cell system (1L, OD_{600} = 10, corresponding to the VHb-WK of 40 mg), the conversion of substrate 1a to product 3a was nearly quantitative, occurring within 5−6 h. Moreover, the production of the trans-(1*S*,2*S*)-configured cyclopropanation product had excellent diastereomeric excess (99% *de*) and enantiomeric excess (99% *ee*) (Table 3).

CONCLUSIONS

This study introduced a new strategy using engineered VHb to synthesize chiral cyclopropane products. Through the construction of simplified mutant library and the training of the NLP model, the efficient mutant VHb-WK with a remarkable 99% *ee* was screened out and identified successfully. MD simulations revealed the molecular mechanisms behind the enhanced enantioselectivity of VHb-WK, shedding light on factors contributing to its superior performance by narrowing the entrance of the enzyme's pocket, favoring the restriction of the formation of reaction intermediates.

The results of this study have important implications for engineering enzymes with enhanced enantioselectivity. Using NLP for mutant prediction and understanding the factors that influence the enantioselectivity of these biocatalysts enable the creation of enzymes with tailored properties to serve various needs in synthetic chemistry, pharmaceuticals, and other industries. More efficient and enantioselective biocatalysts can be developed. Overall, this work expands the catalytic repertoire of hemoglobin-catalyzed compounds and provides insights into the catalytic mechanism of residues near the active site, advancing our comprehension of the intricate relationship between the enzyme structure and function. Meanwhile, we set the stage for the development of carbene-transfer biocatalysts in the future by integrating machine learning and protein engineering.

METHODS

All of the chemicals and reagents were purchased from commercial suppliers (Sigma-Aldrich, Bide Pharmatech, Aladdin, Energy Chemical, TCI) and used without any further purification, unless otherwise stated.

General Procedure for the Biocatalytic Cyclopropanation Reactions

In a typical procedure, the olefin (5 mM, 0.02 mmol in 600 uL of methanol) was added to a 10 mL three-necked flask containing 2.8 mL of PBS solutions of hemoproteins (0.025% mol) and sodium dithionite (5 mM, 0.02 mmol), equipped with a magnetic stir bar and sealed with a rubber septum. A solution of EDA (5 mM, 0.02 mmol in 600 uL of Methanol) was injected into the three-necked flask slowly in 30 min at rt. The reaction mixture was stirred 5 h under a nitrogen atmosphere. For product analysis, the reaction mixtures were extracted with dichloromethane (4 mL x 3) and the combined organic layers were dried over $MgSO_4$ and concentrated under reduced pressure. The crude product was purified by flash column chromatography using silica gel and ethyl acetate/petroleum ether as the eluent to isolate the cyclopropanation product. The purified product was characterized by NMR and chiral HPLC for stereoselectivity determination, and they were used as authentic standards for the construction of calibration curves for determination of TON and yield values.

ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/jacsau.4c01045.

The Supporting Information contains detailed descriptions of the experimental procedures, product characterization data, NMR spectra, screening data, and data related to the computational studies (PDF)

AUTHOR INFORMATION

Corresponding Authors

Fengxi Li − *Key Laboratory of Molecular Enzymology and Engineering of Ministry of Education, School of Life Sciences, Jilin University, Changchun 130023, P. R. China*; Email: lifengxi@jlu.edu.cn

Weiwei Han − *Key Laboratory of Molecular Enzymology and Engineering of Ministry of Education, School of Life Sciences, Jilin University, Changchun 130023, P. R. China*; orcid.org/0000-0002-1931-9316; Email: weiweihan@jlu.edu.cn

Lei Wang − *Key Laboratory of Molecular Enzymology and Engineering of Ministry of Education, School of Life Sciences, Jilin University, Changchun 130023, P. R. China*; orcid.org/0000-0002-9728-0613; Email: w_lei@jlu.edu.cn

Authors

Hanqing Xie − *Key Laboratory of Molecular Enzymology and Engineering of Ministry of Education, School of Life Sciences, Jilin University, Changchun 130023, P. R. China*

Kaifeng Liu − *Key Laboratory of Molecular Enzymology and Engineering of Ministry of Education, School of Life Sciences, Jilin University, Changchun 130023, P. R. China*

Zhengqiang Li − *Key Laboratory of Molecular Enzymology and Engineering of Ministry of Education, School of Life Sciences, Jilin University, Changchun 130023, P. R. China*

Zhi Wang − *Key Laboratory of Molecular Enzymology and Engineering of Ministry of Education, School of Life Sciences, Jilin University, Changchun 130023, P. R. China*

Chunyu Wang − *State Key Laboratory of Supramolecular Structure and Materials, Jilin University, Changchun 130023, P. R. China*

Complete contact information is available at: https://pubs.acs.org/10.1021/jacsau.4c01045

Author Contributions

[#]Xie, H. and Liu, K. contributed equally to this paper. Xie, H. performed the study and analyzed the experimental data; Liu, K. performed the machine learning and NLP building components; Li, F. and Wang, C. performed the work related to product characterization; Han, W. directed the research related to

https://doi.org/10.1021/jacsau.4c01045
JACS Au 2024, 4, 4957−4967

molecular dynamics modeling; Li, Z. directed the culture of the strain and the process of protein modification; Wang, L. and Wang, Z. designed and directed the whole project. All authors contributed to the manuscript writing and reviewing.

Funding

Supported by the Science and Technology Development Program of Jilin Province (no. 20230101135JC). Supported by Graduate Innovation Fund of Jilin University.

Notes

The authors declare no competing financial interest.

REFERENCES

(1) Devine, P. N.; Howard, R. M.; Kumar, R.; Thompson, M. P.; Truppo, M. D.; Turner, N. J. Extending the application of biocatalysis to meet the challenges of drug development. *Nat. Rev. Chem.* **2018**, *2* (12), 409−421.
(2) Drienovská, I.; Mayer, C.; Dulson, C.; Roelfes, G. A designer enzyme for hydrazone and oxime formation featuring an unnatural catalytic aniline residue. *Chem. Rev.* **2018**, *10* (9), 946−952.
(3) Liu, Y.; Ma, T.; Guo, Z.; Zhou, L.; Liu, G.; He, Y.; Ma, L.; Gao, J.; Bai, J.; Hollmann, F.; Jiang, Y. Asymmetric α-benzylation of cyclic ketones enabled by concurrent chemical aldol condensation and biocatalytic reduction. *Nat. Commun.* **2024**, *15* (1), 71.
(4) Li, J.-K.; Qu, G.; Li, X.; Tian, Y.; Cui, C.; Zhang, F.-G.; Zhang, W.; Ma, J.-A.; Reetz, M. T.; Sun, Z. Rational enzyme design for enabling biocatalytic Baldwin cyclization and asymmetric synthesis of chiral heterocycles. *Nat. Commun.* **2022**, *13* (1), 7813.
(5) Carminati, D. M.; Decaens, J.; Couve-Bonnaire, S.; Jubault, P.; Fasan, R. Biocatalytic Strategy for the Highly Stereoselective Synthesis of CHF2-Containing Trisubstituted Cyclopropanes. *Angew. Chem., Int. Ed.* **2021**, *60* (13), 7072−7076.
(6) Fan, S.; Cong, Z. Emerging Strategies for Modifying Cytochrome P450 Monooxygenases into Peroxizymes. *Acc. Chem. Res.* **2024**, *57* (4), 613−624.
(7) Brandenberg, O. F.; Chen, K.; Arnold, F. H. Directed Evolution of a Cytochrome P450 Carbene Transferase for Selective Functionalization of Cyclic Compounds. *JACS.* **2019**, *141* (22), 8989−8995.
(8) Coelho, P. S.; Brustad, E. M.; Kannan, A.; Arnold, F. H. Olefin Cyclopropanation via Carbene Transfer Catalyzed by Engineered Cytochrome P450 Enzymes. *Science* **2013**, *339* (6117), 307−310.
(9) Tinoco, A.; Wei, Y.; Bacik, J.-P.; Carminati, D. M.; Moore, E. J.; Ando, N.; Zhang, Y.; Fasan, R. Origin of High Stereocontrol in Olefin Cyclopropanation Catalyzed by an Engineered Carbene Transferase. *ACS Catal.* **2019**, *9* (2), 1514−1524.
(10) Huang, S.; Deng, W.; Liao, R.; He, C. Repurposing a Nitric Oxide Transport Hemoprotein Nitrophorin 2 for Olefin Cyclopropanation. *ACS Catal.* **2022**, *12* (21), 13725−13731.
(11) Xie, H.; Li, F.; Xu, Y.; Wang, C.; Xu, Y.; Wu, J.; Li, Z.; Wang, Z.; Wang, L. *Vitreoscilla* hemoglobin: a natural carbene transfer catalyst for diastereo- and enantioselective synthesis of nitrile-substituted cyclopropanes. *Green Chem.* **2023**, *25* (17), 6853−6858.
(12) Brandenberg, O. F.; Prier, C. K.; Chen, K.; Knight, A. M.; Wu, Z.; Arnold, F. H. Stereoselective Enzymatic Synthesis of Heteroatom-Substituted Cyclopropanes. *ACS Catal.* **2018**, *8* (4), 2629−2634.
(13) Mao, R.; Wackelin, D. J.; Jamieson, C. S.; Rogge, T.; Gao, S.; Das, A.; Taylor, D. M.; Houk, K. N.; Arnold, F. H. Enantio- and Diastereoenriched Enzymatic Synthesis of 1,2,3-Polysubstituted Cyclopropanes from (Z/E)-Trisubstituted Enol Acetates. *JACS.* **2023**, *145* (29), 16176−16185.
(14) Mahajan, M.; Mondal, B. How Axial Coordination Regulates the Electronic Structure and C−H Amination Reactivity of Fe−Porphyrin−Nitrene? *JACS Au.* **2023**, *3* (12), 3494−3505.
(15) Mao, R.; Taylor, D. M.; Wackelin, D. J.; Rogge, T.; Wu, S. J.; Sicinski, K. M.; Houk, K. N.; Arnold, F. H. Biocatalytic, stereoconvergent alkylation of (Z/E)-trisubstituted silyl enol ethers. *Nat. Synth.* **2023**, *3* (2), 256−264.
(16) Zhang, J.; Maggiolo, A. O.; Alfonzo, E.; Mao, R.; Porter, N. J.; Abney, N. M.; Arnold, F. H. Chemodivergent C(sp3)−H and C(sp2)−H cyanomethylation using engineered carbene transferases. *Nat. Catal* **2023**, *6* (2), 152−160.
(17) Ren, X.; Couture, B. M.; Liu, N.; Lall, M. S.; Kohrt, J. T.; Fasan, R. Enantioselective Single and Dual α-C−H Bond Functionalization of Cyclic Amines via Enzymatic Carbene Transfer. *J. Am. Chem. Soc.* **2023**, *145* (1), 537−550.
(18) Nam, D.; Tinoco, A.; Shen, Z.; Adukure, R. D.; Sreenilayam, G.; Khare, S. D.; Fasan, R. Enantioselective Synthesis of α-Trifluoromethyl Amines via Biocatalytic N−H Bond Insertion with Acceptor-Acceptor Carbene Donors. *J. Am. Chem. Soc.* **2022**, *144* (6), 2590−2602.
(19) McIntosh, J. A.; Coelho, P. S.; Farwell, C. C.; Wang, Z. J.; Lewis, J. C.; Brown, T. R.; Arnold, F. H. Enantioselective Intramolecular C-H Amination Catalyzed by Engineered Cytochrome P450 Enzymes In Vitro and In Vivo. *Angew. Chem., Int. Ed.* **2013**, *52* (35), 9309−9312.
(20) Hamaker, C. G.; Djukic, J.-P.; Smith, D. A.; Woo, L. K. Mechanism of Cyclopropanation Reactions Mediated by (5,10,15,20-Tetra-p-tolylporphyrinato) osmium (II) Complexes. *Organometallics* **2001**, *20* (24), 5189−5199.
(21) Jana, S.; Guo, Y.; Koenigs, R. M. Recent Perspectives on Rearrangement Reactions of Ylides via Carbene Transfer Reactions. *CHEM-EUR J.* **2021**, *27* (4), 1270−1281.
(22) Epping, R. F. J.; Hoeksma, M. M.; Bobylev, E. O.; Mathew, S.; de Bruin, B. Cobalt (II)−tetraphenylporphyrin-catalysed carbene transfer from acceptor−acceptor iodonium ylides via N-enolate−carbene radicals. *Nat. Chem.* **2022**, *14* (5), 550−557.
(23) Wong, H. N. C.; Hon, M.-Y.; Tse, C.-W.; Yip, Y.-C.; Tanko, J.; Hudlicky, T. Use of cyclopropanes and their derivatives in organic synthesis. *Chem. Rev.* **1989**, *89*, 165−198.
(24) Chen, D. Y.-K.; Pouwer, R. H.; Richard, J.-A. Recent advances in the total synthesis of cyclopropane-containing natural products. *Chem. Soc. Rev.* **2012**, *41*, 4631−4642.
(25) Talele, T. T. The "Cyclopropyl Fragment" is a Versatile Player that Frequently Appears in Preclinical/Clinical Drug Molecules. *J. Med. Chem.* **2016**, *59* (19), 8712−8756.
(26) Abu-Elfotoh, A.-M.; Phomkeona, K.; Shibatomi, K.; Iwasa, S. Asymmetric Inter- and Intramolecular Cyclopropanation Reactions Catalyzed by a Reusable Macroporous-Polymer-Supported Chiral Ruthenium (II)/Phenyloxazoline Complex. *Angew. Chem., Int. Ed.* **2010**, *49* (45), 8439−8443.
(27) Carreiro, E. P.; Burke, A. J.; Ramalho, J. P. P.; Rodrigues, A. I. Arylid-OX and Arylid-BOX derived catalysts: applications in catalytic asymmetric cyclopropanation. *Tetrahedronasymmetryt.* **2009**, *20* (11), 1272−1278.
(28) Kanchiku, S.; Suematsu, H.; Matsumoto, K.; Uchida, T.; Katsuki, T. Construction of an Aryliridium−Salen Complex for Highly cis- and Enantioselective Cyclopropanations. *Angew. Chem., Int. Ed.* **2007**, *46* (21), 3889−3891.
(29) Mix, K. A.; Aronoff, M. R.; Raines, R. T. Diazo Compounds: Versatile Tools for Chemical Biology. *ACS Chem. Biol.* **2016**, *11* (12), 3233−3244.
(30) Green, S. P.; Wheelhouse, K. M.; Payne, A. D.; Hallett, J. P.; Miller, P. W.; Bull, J. A. Thermal Stability and Explosive Hazard Assessment of Diazo Compounds and Diazo Transfer Reagents. *Org. Process Res. Dev.* **2020**, *24* (1), 67−84.
(31) Durka, J.; Turkowska, J.; Gryko, D. Lightening Diazo Compounds? *ACS Sustainable Chem. Eng.* **2021**, *9* (27), 8895−8918.
(32) Marshall, L. R.; Bhattacharya, S.; Korendovych, I. V. Fishing for Catalysis: Experimental Approaches to Narrowing Search Space in Directed Evolution of Enzymes. *JACS Au.* **2023**, *3* (9), 2402−2412.
(33) García-García, J. D.; Van Gelder, K.; Joshi, J.; Bathe, U.; Leong, B. J.; Bruner, S. D.; Liu, C. C.; Hanson, A. D. Using continuous directed evolution to improve enzymes for plant applications. *Plant Physiol.* **2022**, *188* (2), 971−983.
(34) Voskarides, K. Directed Evolution. The Legacy of a Nobel Prize. *J. Mol. Evol.* **2021**, *89* (3), 189−191.
(35) Zeymer, C.; Hilvert, D. Directed Evolution of Protein Catalysts. *Annu. Rev. Biochem.* **2018**, *87*, 131−157.

https://doi.org/10.1021/jacsau.4c01045
JACS Au 2024, 4, 4957−4967

(36) Kunka, A.; Marques, S. M.; Havlasek, M.; Vasina, M.; Velatova, N.; Cengelova, L.; Kovar, D.; Damborsky, J.; Marek, M.; Bednar, D.; Prokop, Z. Advancing Enzyme's Stability and Catalytic Efficiency through Synergy of Force-Field Calculations, Evolutionary Analysis, and Machine Learning. *ACS Catal.* **2023**, *13* (19), 12506−12518.
(37) Kroll, A.; Ranjan, S.; Engqvist, M. K. M.; Lercher, M. J. A general model to predict small molecule substrates of enzymes based on machine and deep learning. *Nat. Commun.* **2023**, *14* (1), 2787.
(38) Madani, A.; Krause, B.; Greene, E. R.; Subramanian, S.; Mohr, B. P.; Holton, J. M.; Olmos, J. L.; Xiong, C.; Sun, Z. Z.; Socher, R.; Fraser, J. S.; Naik, N. Large language models generate functional protein sequences across diverse families. *Nat. Biotechnol.* **2023**, *41* (8), 1099−1106.
(39) Lu, H.; Diaz, D. J.; Czarnecki, N. J.; Zhu, C.; Kim, W.; Shroff, R.; Acosta, D. J.; Alexander, B. R.; Cole, H. O.; Zhang, Y.; Lynd, N. A.; Ellington, A. D.; Alper, H. S. Machine learning-aided engineering of hydrolases for PET depolymerization. *Nature* **2022**, *604* (7907), 662−667.
(40) Zhang, Q.; Zheng, W.; Song, Z.; Zhang, Q.; Yang, L.; Wu, J.; Lin, J.; Xu, G.; Yu, H. Machine Learning Enables Prediction of Pyrrolysyl-tRNA Synthetase Substrate Specificity. *ACS Synth.* **2023**, *12* (8), 2403−2417.
(41) Ferruz, N.; Schmidt, S.; Höcker, B. ProtGPT2 is a deep unsupervised language model for protein design. *Nat. Commun.* **2022**, *13* (1), 4348.
(42) Inoue, M.; Sumii, Y.; Shibata, N. Contribution of Organofluorine Compounds to Pharmaceuticals. *ACS Omega* **2020**, *5* (19), 10633−10640.
(43) Lin, Z.; Long, H.; Bo, Z.; Wang, Y.; Wu, Y. New descriptors of amino acids and their application to peptide QSAR study. *Peptides.* **2008**, *29* (10), 1798−1805.
(44) Bordeaux, M.; Tyagi, V.; Fasan, R. Highly Diastereoselective and Enantioselective Olefin Cyclopropanation Using Engineered Myoglobin-Based Catalysts. *Angew. Chem., Int. Ed.* **2015**, *54* (6), 1744−1748.
(45) Zhu, T.; Sun, J.; Pang, H.; Wu, B. Computational Enzyme Redesign Enhances Tolerance to Denaturants for Peptide C-Terminal Amidation. *JACS Au* **2024**, *4* (2), 788−797.

https://doi.org/10.1021/jacsau.4c01045
JACS Au 2024, 4, 4957−4967