Angewandte Chemie International Edition

Angewandte International Edition Chemie

GDCh German Chemical Society

How to cite: *Angew. Chem. Int. Ed.* **2026**, e26025
doi.org/10.1002/anie.202526025

RESEARCH ARTICLE OPEN ACCESS Hot Paper

PCA-Based Database Mining Enables the Discovery of Bacterial Carbene Transferases for Stereodivergent Cyclopropanation

Shunsuke Kato[1,2] | Koki Takeuchi[3] | Kohei Umeda[3] | Hisashi Kudo[1] | Tomohisa Hasunuma[1,2,4] | Takashi Hayashi[3]

[1]Engineering Biology Research Center, Kobe University, Kobe, Japan | [2]Graduate School of Science, Technology, and Innovation, Kobe University, Kobe, Japan | [3]Department of Applied Chemistry, Graduate School of Engineering, The University of Osaka, Suita, Osaka, Japan | [4]RIKEN Center for Sustainable Resource Science, Yokohama, Japan

Correspondence: Shunsuke Kato (s_kato@port.kobe-u.ac.jp) | Takashi Hayashi (thayashi@chem.eng.osaka-u.ac.jp)

Received: 24 November 2025 | **Revised:** 17 January 2026 | **Accepted:** 20 January 2026

Keywords: biocatalysis | database mining | enzyme discovery | hemoproteins | stereodivergent synthesis

ABSTRACT

Protein engineering is a practical approach to providing enzymes with an "abiotic" catalytic activity. However, it remains difficult to explore the full diversity of natural sequence space through the engineering of a single specific protein. As an alternative to these protein engineering approaches, we here demonstrate a database mining approach using a principal component analysis (PCA)-based clustering method to facilitate the identification of promising enzyme candidates. As a proof of concept, we applied this method to the cyclopropanation of styrene, and the sequence space of bacterial globins in the database was extensively investigated. By screening 275 globins from 171 different organisms, we successfully discovered enzymes capable of catalyzing stereodivergent carbene transfer reactions. Furthermore, statistical analyses of sequence data allowed us to detect characteristic structural properties of these globins, which determine the unique stereoselectivity of cyclopropanation. While these bioinformatics tools have primarily been applied to predict enzymes' natural biological functions, this study demonstrates their applicability to exploring enzyme candidates for abiotic reactions unrelated to their native biological activity. Given the increasing interest in biocatalytic applications beyond natural reactivity, this PCA-based mining approach provides a promising direction for expanding the functional diversity of biocatalysts.

1 | Introduction

With the rapid progress in biotechnology experienced in recent years, biocatalysis has emerged as a powerful synthetic tool that provides an efficient way to produce chemical feedstocks [1]. One of the main challenges associated with biocatalysis is to expand the catalytic repertoire of natural enzymes to meet the demands of synthetic chemistry. In this aspect, protein engineering approaches such as directed evolution are quite useful in efforts to provide enzymes with an "abiotic" catalytic activity [2–4]. Through iterative cycles of mutagenesis and screening, it has been demonstrated that the engineered enzymes acquire abiotic reactivities distinct from the natural biological functions of the parent enzymes [5–16]. Focusing on hemoproteins, a number of excellent catalytic systems have been reported over the past decade, as exemplified by the pioneering work on engineering of cytochrome P450s for carbene [17–20] and nitrene [21–23] transfer reactions. However, it remains difficult to fully derive a targeted non-natural function from the sequence space resulting from the engineering of a single specific protein. For example,

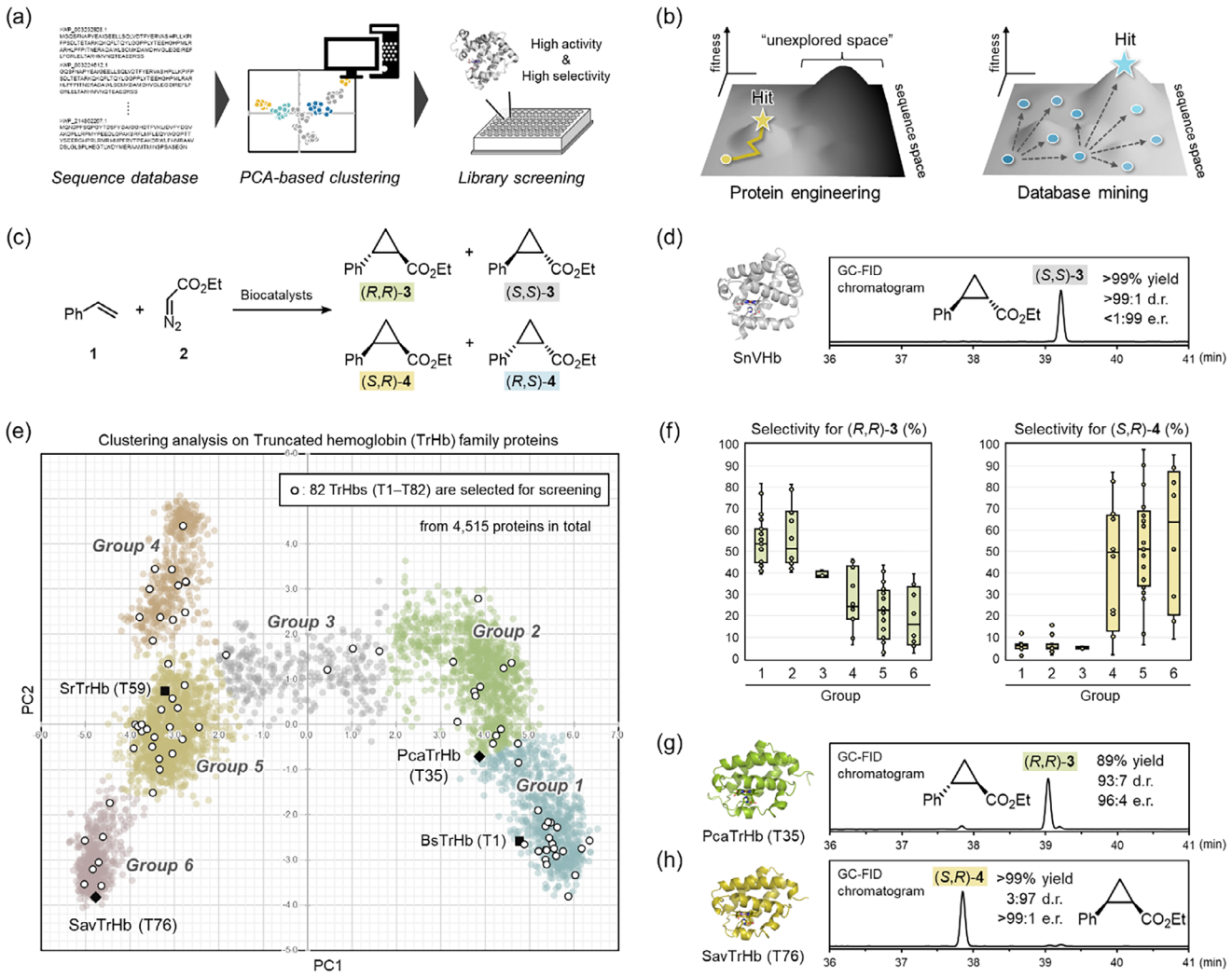

FIGURE 1 | Database mining for the discovery of stereoselective carbene transferases. (a) Schematic workflow for database mining using the PCA-based clustering method. (b) Schematic comparison between protein engineering and database mining for exploration of the sequence space. (c) Biocatalytic cyclopropanation of styrene (**1**) with ethyl diazoacetate (**2**). (d) GC-FID chromatogram for the reaction mixture (**1** + **2**) catalyzed by SnVHb. (e) PCA-based cluster analysis of the TrHb family proteins. The 4515 sequences of TrHbs were collected by BLAST using BsTrHb and SrTrHb (highlighted in a closed square) as queries and analyzed based on the PCA-based clustering method. The data points for the 82 TrHbs (T1–T82) selected for the screening are shown as open circles. The data points for PcaTrHb (T35) and SavTrHb (T76) are shown as solid diamonds. (f) Comparison of product selectivities between Groups 1 and 6 in the TrHb family. Product selectivities of each TrHb group for (*R*,*R*)-**3** and (*S*,*R*)-**4** are plotted in the box-plot graphs. Product selectivity is defined as the ratio (%) of the yield of each stereoisomer to the total yield of **3** and **4**. Boxes enclose the interquartiles (25%–75%), horizontal lines represent medians, and range bars show the maximum and minimum values excluding outliers. (g and h) GC-FID chromatograms for the reaction mixture (**1** + **2**) catalyzed by PcaTrHb and SavTrHb under the optimized reaction conditions.

conventional mutation methods, such as error-prone PCR and site-saturation mutagenesis, do not allow for dramatic diversification of protein sequences. This sometimes leads to misdirecting a given evolutionary effort. To overcome the shortcomings of such protein engineering approaches, our group is exploring an alternative approach to search for uncharacterized biocatalysts, which are best suited for the target reaction among the numerous proteins found in nature. We demonstrate herein a database mining approach using a principal component analysis (PCA)-based clustering method to fully investigate the sequence space for the generation of biocatalysts (Figure 1a).

Biological sequence databases such as UniProt, NCBI, and KEGG provide an enormous and growing number of protein sequences. Only a small fraction of proteins in these databases have been experimentally investigated in connection with their original functions, while the functions of most other proteins have not yet been characterized. For example, among 1 162 214 bacterial heme-binding proteins included in UniProtKB, the functions of only 2309 proteins (less than 0.2 %) have been experimentally clarified to date. Therefore, database mining of "uncharacterized" proteins should provide a more expansive sequence space for discovering novel enzymes with high catalytic performance as an alternative to conventional engineering approaches for one specific "well-known" protein (Figure 1b) [24–28]. In addition, to efficiently explore the full diversity of these expansive sequence spaces, we decided to investigate a PCA-based clustering method. PCA is an unsupervised multivariate technique, which can be used

to simplify multidimensional data [29]. By converting multiple alignment data of protein sequences into a vector of binary variables, the protein sequences can be plotted in a low-dimensional space, which may reflect functional specificity within protein families. Although these multivariate statistical techniques have been applied for many years primarily to predict natural biological functions of the enzymes (e.g., substrate specificity) [30–34], we postulate that the PCA-based clustering method may also provide a new discipline for discovery of enzymes that promote abiotic chemical transformations.

Based on this concept, we perform database mining using the PCA-based clustering method to discover enzymes that exhibit promising catalytic activities for non-natural chemical transformations. As a proof of concept, the cyclopropanation reaction of styrene (**1**) with ethyl diazoacetate (**2**) is investigated (Figure 1c). This is one of the most representative benchmark reactions that hemoproteins are known to promote apart from their natural biological functions [35]. In previous studies, this carbene transfer reaction has been extensively studied based on the protein engineering approaches for specific hemoproteins, such as cytochrome P450s [17, 36–40], myoglobins [41–44], and several artificial enzymes [45–49]. In spite of these substantial advances, stereodivergent synthesis of all four possible stereoisomers (*R*,*R*)-**3**, (*S*,*S*)-**3**, (*S*,*R*)-**4**, and (*R*,*S*)-**4** with complete control over the stereoselectivities remains challenging (Figure S1) [17, 38–42, 47–50]. To demonstrate the utility of the PCA-based database mining approach, this study investigated the sequence space of hemoproteins to identify four heme-dependent enzymes (hereafter referred to as carbene transferases), which produce each of the four stereoisomers with improved stereoselectivity.

2 | Results and Discussion

2.1 | Phylogenetic Tree Analysis of Carbene Transferases

First, we began data analysis of previous studies to set a starting point for the PCA-based database mining. Our group has previously performed a non-targeted screening of bacterial hemoproteins and identified several carbene transferases that exhibit catalytic activities for stereoselective cyclopropanation (Supporting Data 1) [51]. For example, bacterial hemoglobin from *Starkeya novella* (SnVHb, UniProt ID: D7A317) was found to produce the *trans*-isomer (*S*,*S*)-**3** with high activity (k_{cat} = 4.9 × 10^4 min^{-1}) and excellent stereoselectivities (>99:1 d.r., <1:99 e.r.) (Figure 1d). In contrast, truncated hemoglobin from *Streptosporangium roseum* (SrTrHb, UniProt ID: D2BDZ5) afforded the *cis*-isomer (*S*,*R*)-**4** with moderate activity (k_{cat} = 68 min^{-1}) and high stereoselectivities (2:98 d.r., >99:1 e.r.). Although the screening has targeted various types of hemoproteins (e.g., cytochrome P450s, peroxidases, catalases, tryptophan 2,3-dioxygenases, and nitric oxide synthases), both of these identified enzymes belong to the same globin superfamily [52], suggesting that the arrangement of globin folds will provide an appropriate basis for the stereoselective cyclopropanation reactions. This tendency aligns with previous efforts that have mainly focused on engineering myoglobins for stereoselective carbene-transfer reactions [41–44]. Furthermore, phylogenetic analysis of previous screening data reveals that these bacterial globins have a characteristic tendency toward precision of stereoselectivity, depending on the protein families (Figure S2). The nitric oxide dioxygenase family [53, 54], which includes SnVHb, shows a strong preference for (*S*,*S*)-**3**, whereas the globin-coupled sensor (GCS) protein family [55] shows a moderate preference for the thermodynamically unfavored *cis*-isomers (*S*,*R*)-**4** and (*R*,*S*)-**4**. In addition, several truncated hemoglobins (TrHbs), such as TrHb from *Bacillus subtilis* (BsTrHb, UniProt ID: O31607) [56] and TrHb from *Geobacillus stearothermophilus* (GsTrHb, Uniprot ID: A0A150NBF4) [57], exhibit slight selectivity for the unexplored *trans*-isomer (*R*,*R*)-**3**, whereas more than half of TrHbs, as represented by SrTrHb, provide (*S*,*R*)-**4** selectivity. Given these initial screening results, we performed PCA-based protein database mining on these bacterial globin families, particularly the TrHb and GCS families, to identify carbene transferases for the stereodivergent cyclopropanation.

2.2 | PCA-Based Exploration of Bacterial TrHb Family for Synthesis of (*R*,*R*)-3 and (*S*,*R*)-4

We first explored the sequence space of the TrHb family for the selective synthesis of (*R*,*R*)-**3** and (*S*,*R*)-**4** using the PCA-based clustering method. The sequence of SrTrHb with high selectivity for (*S*,*R*)-**4** (2:98 d.r., >99:1 e.r.) and the sequence of BsTrHb with moderate stereoselectivity for (*R*,*R*)-**3** (89:11 d.r., 76:24 e.r.) were subjected to Basic Local Alignment Search Tool (BLAST) [58, 59] to collect the homologous sequences. The obtained sequences mainly consisted of TrHbs from *Actinomycetota* and *Bacillota* but also included some TrHbs from *Deinococcota*, *Pseudomonadota*, *Bacteroidota*, and *Chloroflexota*, indicating widespread distribution of the TrHb family in nature. After combining the output of the BLAST search with deletion of the overlapped sequences, the dataset containing 4515 sequences of TrHbs was analyzed by the PCA-based clustering method. [34] In short, the sequence dataset was first aligned using the MAFFT program, and the resulting alignment data was converted into numerical representation schemes based on a binary vector matrix (Table S1) to maximize the variance of the mean-centered variance/covariance matrix. PCA was next performed based on the covariance matrix to reduce the data dimension of the matrix data. The PCA score plot (PC1 vs. PC2) was then created, and the clustering analysis was performed based on the X-means clustering algorithm. Consequently, the collected sequences of TrHb family were found to be classified into six groups (Groups 1–6) (Figure 1e). BsTrHb, which was used as a BLAST query, was placed in Group 1, which occupies the bottom right part of the graph, and SrTrHb was placed in Group 5, which is positioned in the middle left part of the graph. According to these clustering results, we selected 82 TrHbs (T1–T82), which are widely distributed throughout the clustering graph and screened them to identify new carbene transferases with enhanced stereoselectivities (Figure S3, Supporting Data 2).

The screening was performed in vitro using purified enzymes as follows. The genes of the selected TrHbs (T1–T82) were obtained from the genomic DNA of the corresponding source bacterium and cloned into a pET-21 vector as a fusion with *Strep*-tag II [60]. TrHbs were then recombinantly expressed by the *E. coli* BL21-Gold(DE3) strain and purified using a chitin- and streptavidin-mediated affinity purification (CSAP) system that we developed previously [51]. Essentially all TrHbs (80 out of

82 TrHbs) were efficiently obtained as soluble proteins in the purified fractions and directly investigated for catalysis of the cyclopropanation reaction of **1** with **2** to assess the stereoselectivity of each TrHb. Interestingly, the screening revealed that each TrHb group has a clear tendency toward the stereoselectivity of the cyclopropanation as indicated in box plots (Figures 1f and S4). TrHbs in Group 1 and Group 2, which scored positive values in PC1, exhibit relatively high stereoselectivity for formation of the *trans*-isomer (*R*,*R*)-**3**. In contrast, TrHbs in Groups 4–6, which have negative PC1 scores, have a strong preference for formation of the *cis*-isomer (*S*,*R*)-**4**. Although the natural biological functions of TrHbs in living organisms are not related to the carbene transfer reactions [61, 62], it should be noted that the PCA-based clustering clearly reflects the stereoselectivity tendency for this abiotic chemical transformation. This is the first example demonstrating the applicability of PCA-based clustering for the classification of stereoselectivities in the enzymatic transformations, including not only abiotic chemical transformations but also native reactions of enzymes. In addition, when these TrHbs were classified into two groups based on the optimal growth temperature of source microorganisms, the TrHbs of thermophiles were found to generate, on average, higher stereoselectivity than TrHbs from mesophiles (Figure S5). This suggests that the rigidity of the overall protein structure as well as the partial structure of the TrHb's active site may contribute to the stereoselectivity of these enzymes.

Through the screening of T1–T82, we succeeded in identifying two best-performing carbene transferases, which promote diastereodivergent synthesis of (*R*,*R*)-**3** and (*S*,*R*)-**4** (Figure 1g,h). Truncated hemoglobin from *Polycladomyces abyssicola* (Pca-TrHb, T35, UniProt ID: A0A8D5UG41), which is located at the boundary between Group 1 and Group 2, selectively produced the *trans*-isomer (*R*,*R*)-**3** in 89% yield under an optimized condition (Figures 1g and S6a). Starting from BsTrHb with moderate selectivity (89:11 d.r., 76:24 e.r.), the stereoselectivity for (*R*,*R*)-**3** is dramatically increased up to 93:7 d.r. and 96:4 e.r. through the database mining procedure. Furthermore, truncated hemoglobin from *Streptomyces avermitilis* (SavTrHb, T76, UniProt ID: Q82CH9), which belongs to Group 5 in the clustering graph, quantitatively produces the *cis*-isomer (*S*,*R*)-**4** with excellent stereoselectivity (3:97 d.r., >99:1 e.r.) (Figures 1h and S6b). Notably, SavTrHb was found to show much higher catalytic activity than SrTrHb, which was identified in the previous study [51]. The k_{cat} value of SavTrHb was determined to be 2.2×10^2 min^{-1}, which is >3-fold higher than that of SrTrHb (Figure S7 and Table S2). These results clearly demonstrate the applicability of the PCA-based clustering methods to perform protein database mining for this abiotic chemical transformation.

2.3 | PCA-Based Exploration of Bacterial GCS Family for Stereoselective Synthesis of (*R*,*S*)-4

In addition to the TrHb family, we also investigated the sequence space of the globin-coupled sensor (GCS) family to identify another enzyme capable of producing the unacquired *cis*-isomer (*R*,*S*)-**4**. Among the globins investigated in our previous study, the GCS proteins were found to show the highest selectivity for (*R*,*S*)-**4**, although there is much room for improvement (Figure S2). Setting this finding as a starting point, the PCA-based clustering analysis was performed to explore the full diversity of this GCS family. The sequence collection was performed by restricting the BLAST output to the globin domains because the whole sequences of the obtained GCS proteins predominantly contain transmembrane domains or other functional domains [55]. The PCA-based cluster analysis was then carried out on the collected 3697 sequences of these globin domains as described above. Consequently, the collected sequences were found to be classified into nine groups (Groups 1–9) (Figure 2a). Groups 1–7, which constitute one large cluster in the center of the graph, are mainly derived from the globin domains of membrane sensor proteins, such as methyl accepting chemotaxis proteins [63] and diguanylate cyclases [64]. Groups 8 and 9, which form small clusters at a distant location, were likely to be cytosolic sensor proteins, which do not possess the other functional domains [65]. Based on this clustering result, 54 proteins (G1–G54) were selected throughout the clustering graph, prepared as soluble proteins after removing the transmembrane domains and other functional domains, and evaluated for their stereoselectivities toward (*R*,*S*)-**4** as described above (Figure S8, Supporting Data 3).

Through the screening of G1–G54, we revealed that the GCS proteins in Group 8, which mostly originate from thermophilic bacteria of the *Thermaceae* family, have higher selectivities for the targeted stereoisomer (*R*,*S*)-**4** than the other groups (Figures 2b and S9). In particular, GCS from *Meiothermus granaticius* (MgGCS, G42, UniProt: A0A399FF38) was identified as having the highest selectivity among all of the proteins (Figure 2c). Under optimized reaction conditions, MgGCS affords the desired stereoisomer (*R*,*S*)-**4** in 96% yield with high stereoselectivity (12:88 d.r., 9:91 e.r.) (Figures 2d and S10), ultimately leading to the stereodivergent synthesis of ethyl 2-phenylcyclopropanecarboxylates (**3**, **4**) in combination with the other carbene transferases identified so far (Figure S11). To clarify the practical advantages of our PCA-based clustering method, we compared the results with those of a sequence similarity network (SSN), which is widely used to visualize relationships between protein sequences. SSN was constructed based on the same dataset of 3697 GCS sequences using the EFI-EST web tool [66–68]. When the similarity threshold is increased from 30% to 55% identity, (*R*,*S*)-**4**-selective GCS proteins in Group 8 were found to be assembled into a single cluster in the SSN (Figure S12). However, it may be difficult to experimentally distinguish this beneficial cluster because this SSN contains numerous non-selective clusters and more than 500 isolated nodes (Figure S13). Since SSN algorithm evaluates sequence similarity across the entire protein length with equal weighting via pairwise alignment [67], SSN may overestimate the phylogenetic relationships among proteins, thereby producing an excessive number of clusters and isolated nodes, which are heavily biased by the taxonomic origin of each protein. On the other hand, PCA-based method captures the most significant variations of multiple alignment data as principal components PC1 and PC2. As exemplified by the functional classification of alcohol dehydrogenases in a recent report [34], the principal components may provide a concise and intuitive classification by capturing sequence-based patterns that correlate with functionally relevant features of GCS proteins. Furthermore, the PCA-based clustering method can reflect the relationships among protein sequences as distances within a unified graph, even for sequences that appear as an isolated node in SSN. In contrast, the distances between nodes and

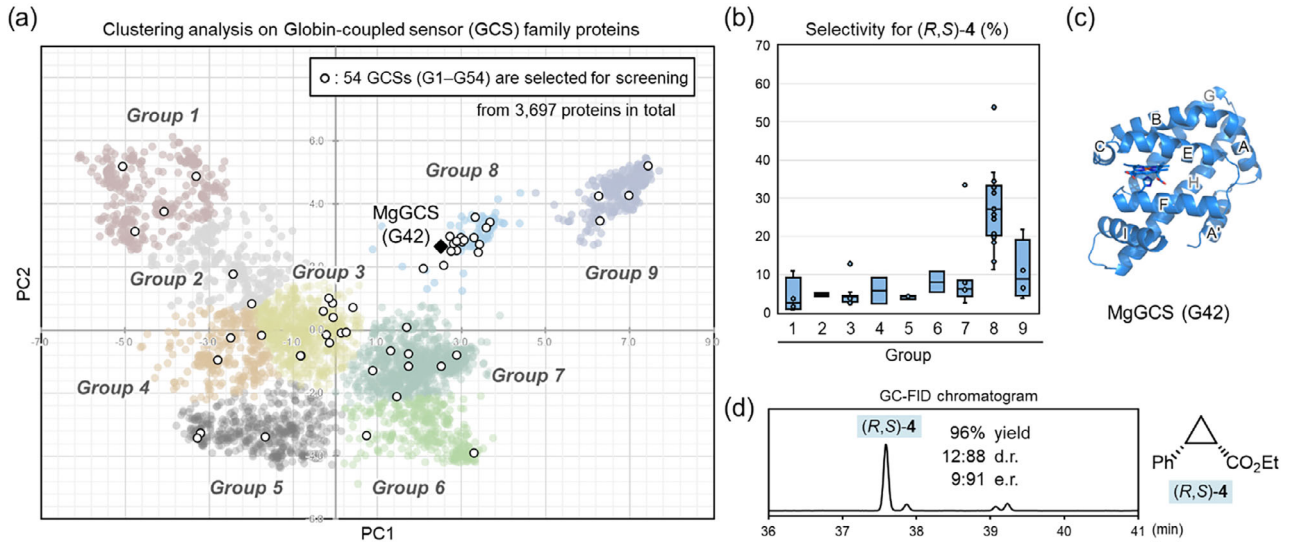

FIGURE 2 | Mining of GCS family proteins for the synthesis of (*R*,*S*)-**4**. (a) PCA-based cluster analysis of the GCS family proteins. The 3697 sequences of the heme domain of the GCS proteins were collected by BLAST and analyzed based on the PCA-based clustering method. The data points for the 54 proteins (G1–G54) selected for the screening are shown as open circles. MgGCS (G42) is shown as a solid diamond. (b) Comparison of product selectivities between Groups 1 and 9 in the GCS family. Product selectivities of each group for (*R*,*S*)-**4** are plotted in the box-plot graphs. Product selectivity is defined as the ratio (%) of the yield of each stereoisomer to the total yield of **3** and **4**. Boxes enclose the interquartiles (25%–75%), horizontal lines represent medians, and range bars show the maximum and minimum values excluding outliers. (c) Three-dimensional structure of MgGCS predicted by AlphaFold3. α-Helices are designated by letters A–I according to the common classification of the globin fold. (d) GC-FID chromatogram for the reaction mixture (**1** + **2**) catalyzed by MgGCS under the optimized reaction conditions.

clusters in SSN do not necessarily represent the homology among proteins (Figure S14). Therefore, SSN does not preserve global relationships between independent clusters, which makes it less suitable for exploring the overall sequence-function landscape. Considering these advantages, our PCA-based clustering method is expected to provide a more practical and insightful approach to guide the experimental validation of enzyme discovery.

The Michaelis–Menten parameters determined for MgGCS are $k_{cat} = 85\ min^{-1}$, $K_M = 3.9\ mM$, and $k_{cat}/K_M = 22\ min^{-1}\ mM^{-1}$. These values are similar to those of other identified carbene transferases (Figure S15 and Table S2). The large K_M values of these bacterial globins may reflect a lack of structural similarity between styrene and native substrates. Interestingly, MgGCS was found to have a unique globin fold in which an additional α-helix motif at *C*-terminus (hereafter referred to as "helix I" following common classification of the globin fold) is attached to the proximal heme-binding helix F as predicted by AlphaFold3 [69] (Figure 2c). This unique *C*-terminus motif is also conserved in the GCS proteins in Group 8 but is not present in the other groups (Groups 1–7 and 9). We assumed that these "group-specific" properties might be the main factor responsible for the stereoselectivity of MgGCS and other globins. To investigate this possibility, we next performed structural investigations and statistical analyses on the identified globin proteins.

2.4 | Statistical Analysis and Structural Investigation on the Stereoselectivity of Bacterial Carbene Transferases

Through the screening of 151 globin proteins from a total of 130 bacteria, we identified three bacterial globins (PcaTrHb, SavTrHb, and MgGCS), ultimately achieving the stereodivergent synthesis of **3** and **4** in combination with the previously identified SnVHb (Figure S11). In addition to these efforts, we also investigated 124 additional globin proteins, most of which are encoded as isoproteins in the genomes of the 130 bacteria above (Supporting Data 4). To clarify the "group-specific" properties of these diverse bacterial globins, the PCA-based clustering was performed on all globins investigated in this study (275 proteins in total) (Figure 3a). As a result, these globins were efficiently classified into nine groups (Groups A–I). It appears that this classification reflects the differences in the structural motifs of these globin proteins [52]. Globins, which have 3-on-3 α-helical sandwich motifs, such as the GCS family (Groups A–C) and the nitric oxide dioxygenase family (Group D), are clustered in an area with lower PC1 values. Globins, which have a 2-on-2 α-helical sandwich motif, the so-called 2/2 hemoglobins HbP (Group E), HbN (Group F), and HbO (Groups G–I) [70], are assembled in clusters with higher PC1 values. As also represented by the AlphaFold3 structures [69], the structural motifs of the identified enzymes SnVHb, PcaTrHb, SavTrHb, and MgGCS are quite different from each other, even though these enzymes are classified as having a globin fold (Figure 3b). SnVHb has the 3-on-3 motif, which is composed of helices A, B, E, F, G, and H, whereas PcaTrHb and SavTrHb have the 2-on-2 motif, which is composed of helices B, E, G, and H. In addition, MgGCS includes additional α-helices I and A,' which support the 3-on-3 motif of the GCS proteins. These structural differences in the globin folds were predicted to influence the distal environment of the heme active site. In particular, α-helices B and E, located above the heme cofactor, occupy quite different positions among these four globin proteins. This feature may play a crucial role in determining the stereoselectivity of the cyclopropanation (Figures S16 and S17). Since the dramatic changes in the main-chain structure cannot

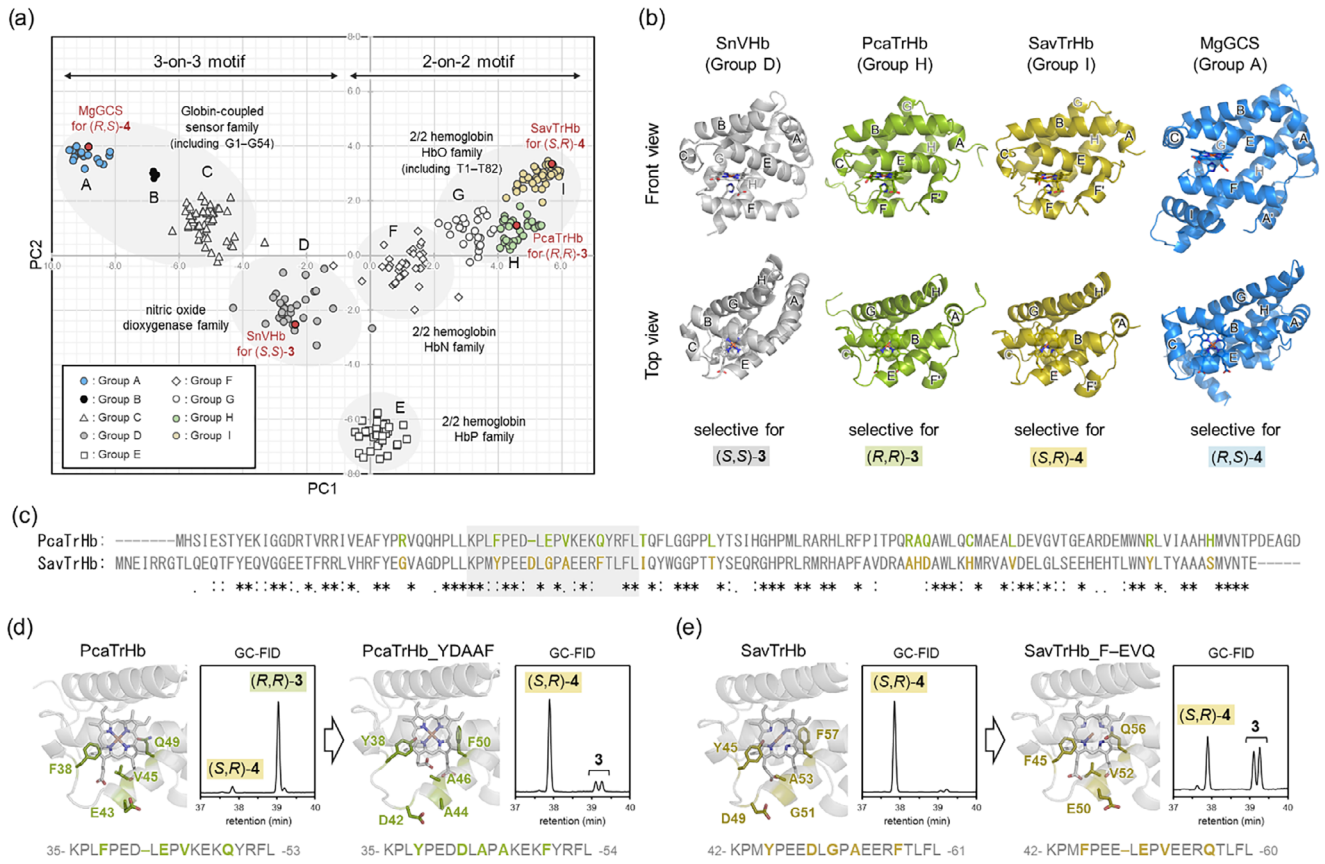

FIGURE 3 | (a) PCA-based cluster analysis for all globins investigated in this study (275 proteins in total). The globins are separated into nine groups (Groups A–I). The data points for SnVHb, PcaTrHb, SavTrHb, and MgGCS, which promote stereodivergent synthesis of cyclopropanes **3** and **4** are highlighted as red circles. (b) Comparison between protein structures of SnVHb, PcaTrHb, SavTrHb, and MgGCS predicted by AlphaFold3. The structures are aligned against the heme cofactor. α-Helices are designated by letters A–I according to the common classification of globin fold. (c) Aligned sequences of PcaTrHb and SavTrHb. The top 15 scoring group-specific amino acid residues in the output of multi-Harmony are highlighted in green and yellow. The regions subjected to the mutagenesis are highlighted with a gray rectangle. (d and e) Cyclopropanation of **1** with **2** catalyzed by PcaTrHb_YDAAF and SavTrHb_F–EVQ in which the "group-specific" residues in the distal heme environments of PcaTrHb and SavTrHb are designed to be exchanged with each other. In the design of PcaTrHb_YDAAF, E43 residue in PcaTrHb was replaced with Ala, which is also conserved in the Group I globin family (such as SrTrHb), although SavTrHb has a Gly residue at this position (Table S4).

be achieved by conventional point mutation methods, this finding may indicate a significant advantage in the larger sequence space offered by the database mining approach.

In addition to the structural information of the globin fold, the output of the cluster analysis also reflects the stereoselectivity tendency in the cyclopropanation reactions. Particularly, the globins in Group A, Group D, Group H, and Group I tend to exhibit relatively higher stereoselectivities for (*R*,*S*)-**4**, (*S*,*S*)-**3**, (*R*,*R*)-**3**, and (*S*,*R*)-**4**, respectively. This is consistent with the results of MgGCS, SnVHb, PcaTrHb, and SavTrHb identified from these groups (Figures 3a and S18). To gain deeper insight into these globin groups and their stereoselectivities, we next analyzed the sequence data using multi-Harmony [71], a web server designed to detect "group-specific" residues in proteins. In the analysis of the GCS family, the sequences in Group A (17 sequences) were aligned against the sequences in Groups B and C (53 sequences in total) using MAFFT program, and the resulting multiple alignment data were analyzed by multi-Harmony to determine the residues, which are characteristically conserved in the Group A globins (Table S3). Among the top 20 high-scoring residues in the output, the majority were found to be located at the interface of α-helices, which may affect the overall structure of the globin fold (Figure S19). In particular, nine amino acid residues are found within the characteristic *C*-terminus motifs of "helix I" among the top 20 residues. To examine the contribution of this *C*-terminus motif to the cyclopropanation reaction, we prepared an MgGCS(Δ160–185) variant in which the *C*-terminus residues (Gly160–Gln185) were truncated relative to wild-type MgGCS (Figures S20 and S21). Notably, the truncation causes a dramatic change in stereoselectivity from the *cis*-isomer (*R*,*S*)-**4** to the *trans*-isomer (*S*,*S*)-**3** (Figure S22). MgGCS(Δ160–185) produces (*S*,*S*)-**3** as a single product with high stereoselectivity (99:1 d.r., 2:98 e.r.), which clearly indicates that the "group A-specific" *C*-terminus motif contributes to the unique (*R*,*S*)-**4** selectivity of MgGCS.

Furthermore, we performed the multi-Harmony analysis on the TrHb families. Although the four carbene transferases identified in this study have diverse protein folding arrangements as described above, it was found that two enzymes, PcaTrHb in Group H and SavTrHb in Group I, which selectively produce the

trans-isomer (*R*,*R*)-**3** and *cis*-isomer (*S*,*R*)-**4**, share a similar 2-on-2 α-helical motif (Figure S23). In addition to the global influence of the globin folds, we speculated that the "group-specific" residues at the active site locally determine the stereoselectivities between (*R*,*R*)-**3** and (*S*,*R*)-**4** in these two globins. Based on this hypothesis, the multi-Harmony analysis was carried out using the sequence alignment data between Group H and Group I to identify the characteristic properties of each group. Interestingly, the outputs of multi-Harmony indicate five positions that are located close to the heme cofactor (Figures S24 and S25; Table S4). These five positions are assigned to F38, E43, V45, and Q49 and one gap in PcaTrHb and Y45, D49, G51, A53, and F57 in SavTrHb (Figure 3c–e). Notably, several of these residues correspond to the previously reported positions that can determine the stereoselectivity of myoglobin catalysis, but also multi-Harmony suggested additional "gap" sites that would be difficult to identify through a conventional rational engineering approach [41, 42]. Following this result, we prepared the PcaTrHb_YDAAF variant in which the five "Group H-specific" positions of PcaTrHb were mutated into "Group I-specific" residues (Figures 3d and S26a). Surprisingly, the PcaTrHb_YDAAF variant was found to exhibit swapped stereoselectivities for the cyclopropanation reaction compared to the wild-type. PcaTrHb_YDAAF produces the *cis*-isomer (*S*,*R*)-**4** with high selectivity (17:83 d.r. and 99:1 e.r.), which apparently reflects the tendency of the Group I globins (Figure S27). As suggested by the AlphaFold structures, these mutations might have caused the conformational changes of helix E (Figure S28) and generated a confined active site that is preferred for the formation of the thermodynamically unfavored *cis*-isomer (Figure S29). Furthermore, we also prepared the SavTrHb_F–EVQ variant, which possesses "Group H-specific" residues within the active site (Figure S26b). Although the products were obtained as a racemic mixture, the diastereoselectivities for the *trans*-isomer **3** have been increased by the mutations (Figures 3e and S30). These results clearly illustrate the advantages of statistical sequence analyses to rationalize the catalytic performance of the identified enzymes, and the information obtained on the "group-specific" properties will serve as a valuable resource to accelerate database mining and protein engineering in future studies.

3 | Conclusion

In summary, we have demonstrated the applicability of a database mining approach to discover promising enzymes capable of catalyzing stereodivergent carbene transfer reactions. Particularly, the PCA-based clustering method enables us to visualize the diverse sequence space of bacterial globins in the database and accelerates the discovery of promising enzymes with high activity and excellent stereoselectivities for cyclopropanation. Consequently, it was revealed that the PCA-based clustering clearly predicts the stereoselectivity of this non-natural carbene transfer reaction, which is the first example demonstrating the applicability of PCA-based clustering for the classification of stereoselectivities in the enzymatic transformations. Through the screening of 82 globins in the TrHb family and 54 globins in the GCS family, we succeeded in identifying three enzymes (PcaTrHb, SavTrHb, and MgGCS) which exhibit divergent stereoselectivity for the cyclopropanation. In combination with the previously identified SnVHb, the challenging stereodivergent syntheses of ethyl 2-phenylcyclopropanecarboxylates (*R*,*R*)-**3**, (*S*,*S*)-**3**, (*S*,*R*)-**4**, and (*R*,*S*)-**4** were achieved, thus indicating the utility of the database mining approach using PCA-based clustering. In addition, statistical analyses were performed on the sequence alignment data of each globin family to determine the "group-specific" properties of these carbene transferases. The characterized "group-specific" residues and structural motifs were revealed to rationalize the unique stereoselectivity of each carbene transferase. While these bioinformatics tools (e.g., protein database, PCA-based clustering, and multi-Harmony analysis) have been mainly applied to predict natural biological functions of enzymes, these results clearly indicate their applicability to exploring enzyme candidates for abiotic chemical transformations unrelated to their native reactivity. Furthermore, with the recent explosive growth of biological sequence databases and the statistical analysis tools, the PCA-based database mining approach has the potential to be widely applied to investigate diverse types of biocatalytic reactions. This includes not only hemoprotein-catalyzed carbene transfer reactions but also various other abiotic reactions catalyzed by a variety of enzymes. For example, the PCA-based clustering methods have been shown to capture trends in catalytic activity, as demonstrated in the radical ring-opening reaction by aldoxime dehydratases which were previously reported by our group (Figure S31) [72]. Given the recent increase in demand for biocatalysis, we are convinced that the PCA-based database mining approach presented in this study will provide powerful and practical methodologies for expanding the functional diversity of enzyme catalysts.

Acknowledgments

We wish to thank Prof. Shigeru Kitani of Aoyama Gakuin University and Prof. Kohsuke Honda of the University of Osaka for their constructive suggestions regarding the DNA cloning experiments. This work was supported by JSPS KAKENHI Grant Numbers 25H01579 (Forecast-Biosyn), JP24H01136 (Bottom-up Biotech), JP23H04554 (ForecastBiosyn), JP22H05421 (Bottom-up Biotech), JP24K01630, JP22K14783, JP21K20535, JP25H00887, JP22K21348, JST ACT-X Grant Number JPMJAX22B6 (Environments and Biotechnology), the Kaneko-Narita Research Fund (Protein Research Foundation), and the Daiichi-Sankyo "Habataku" Support Program for the Next Generation of Researchers (Naedoko Grant).

Conflicts of Interest

The authors declare no conflict of interest.

Data Availability Statement

The data that support the findings of this study are available in the Supporting Information of this article.

References

1. E. O. Romero, A. T. Saucedo, J. R. Hernández-Meléndez, D. Yang, S. Chakrabarty, and A. R. H. Narayan, "Enabling Broader Adoption of Biocatalysis in Organic Chemistry," *Journal of the American Chemical Society Gold* 3 (2023): 2073–2085, https://doi.org/10.1021/jacsau.3c00263.

2. R. B. Leveson-Gower, C. Mayer, and G. Roelfes, "The Importance of Catalytic Promiscuity for Enzyme Design and Evolution," *Nature Reviews Chemistry* 3 (2019): 687–705, https://doi.org/10.1038/s41570-019-0143-x.

3. K. Chen and F. H. Arnold, "Engineering New Catalytic Activities in Enzymes," *Nature Catalysis* 3 (2020): 203–213, https://doi.org/10.1038/s41929-019-0385-5.

4. S. Jain, F. Ospina, and S. C. Hammer, "A New Age of Biocatalysis Enabled by Generic Activation Modes," *Journal of the American Chemical Society Gold* 4 (2024): 2068–2080, https://doi.org/10.1021/jacsau.4c00247.

5. S. Studer, D. A. Hansen, Z. L. Pianowski, et al., "Evolution of a Highly Active and Enantiospecific Metalloenzyme From Short Peptides," *Science* 362 (2018): 1285–1288, https://doi.org/10.1126/science.aau3744.

6. K. F. Biegasiewicz, S. J. Cooper, X. Gao, et al., "Photoexcitation of Flavoenzymes Enables a Stereoselective Radical Cyclization," *Science* 364 (2019): 1166–1169, https://doi.org/10.1126/science.aaw1143.

7. Y. Ye, J. Cao, D. G. Oblinsky, et al., "Using Enzymes to Tame Nitrogen-Centred Radicals for Enantioselective Hydroamination," *Nature Chemistry* 15 (2023): 206–212, https://doi.org/10.1038/s41557-022-01083-z.

8. Q. Zhou, M. Chin, Y. Fu, P. Liu, and Y. Yang, "Stereodivergent Atom-Transfer Radical Cyclization by Engineered Cytochromes P450," *Science* 374 (2021): 1612–1616, https://doi.org/10.1126/science.abk1603.

9. X. Huang, J. Feng, J. Cui, et al., "Photoinduced Chemomimetic Biocatalysis for Enantioselective Intermolecular Radical Conjugate Addition," *Nature Catalysis* 5 (2022): 586–593, https://doi.org/10.1038/s41929-022-00777-4.

10. J. Rui, Q. Zhao, A. J. Huls, et al., "Directed Evolution of Nonheme Iron Enzymes to Access Abiological Radical-Relay C(sp3)–H Azidation," *Science* 376 (2022): 869–874, https://doi.org/10.1126/science.abj2830.

11. R. Crawshaw, A. E. Crossley, L. Johannissen, et al. "Engineering an Efficient and Enantioselective Enzyme for the Morita–Baylis–Hillman Reaction" *Nature Chemistry* 2022, 14, 313–320.

12. L. Cheng, D. Li, B. K. Mai, et al., "Stereoselective Amino Acid Synthesis by Synergistic Photoredox-Pyridoxal Radical Biocatalysis," *Science* 381 (2023): 444–451, https://doi.org/10.1126/science.adg2420.

13. S. Gergel, J. Soler, A. Klein, et al., "Engineered Cytochrome P450 for Direct Arylalkene-to-Ketone Oxidation via Highly Reactive Carbocation Intermediates," *Nature Catalysis* 6 (2023): 606–617, https://doi.org/10.1038/s41929-023-00979-4.

14. C. A. Gomez, D. Mondal, Q. Du, N. Chan, and J. C. Lewis, "Directed Evolution of an Iron(II)- and α-Ketoglutarate-Dependent Dioxygenase for Site-Selective Azidation of Unactivated Aliphatic C−H Bonds" *Angewandte Chemie International Edition* 62 (2023): e202301370.

15. Y. Xu, H. Chen, L. Yu, et al., "A Light-Driven Enzymatic Enantioselective Radical Acylation," *Nature* 625 (2024): 74–78, https://doi.org/10.1038/s41586-023-06822-x.

16. L. Longwitz, R. B. Leveson-Gower, H. J. Rozeboom, A.-M. W. H. Thunnissen, and G. Roelfes, "Boron Catalysis in a Designer Enzyme," *Nature* 629 (2024): 824–829, https://doi.org/10.1038/s41586-024-07391-3.

17. P. S. Coelho, E. M. Brustad, A. Kannan, and F. H. Arnold, "Olefin Cyclopropanation via Carbene Transfer Catalyzed by Engineered Cytochrome P450 Enzymes," *Science* 339 (2013): 307–310, https://doi.org/10.1126/science.1231434.

18. S. B. Kan, R. D. Lewis, K. Chen, and F. H. Arnold, "Directed Evolution of Cytochrome c for Carbon–silicon Bond Formation: Bringing Silicon to Life," *Science* 354 (2016): 1048–1051, https://doi.org/10.1126/science.aah6219.

19. S. B. J. Kan, X. Huang, Y. Gumulya, K. Chen, and F. H. Arnold, "Genetically Programmed Chiral Organoborane Synthesis," *Nature* 552 (2017): 132–136, https://doi.org/10.1038/nature24996.

20. R. K. Zhang, K. Chen, X. Huang, L. Wohlschlager, H. Renata, and F. H. Arnold, "Enzymatic Assembly of Carbon-Carbon Bonds via Iron-Catalysed sp3 C-H Functionalization," *Nature* 565 (2019): 67–72, https://doi.org/10.1038/s41586-018-0808-5.

21. J. A. McIntosh, P. S. Coelho, C. C. Farwell, et al., "Enantioselective Intramolecular C–H Amination Catalyzed by Engineered Cytochrome P450 Enzymes in Vitro and in Vivo," *Angewandte Chemie International Edition* 52 (2013): 9309–9312, https://doi.org/10.1002/anie.201304401.

22. T. K. Hyster, C. C. Farwell, A. R. Buller, J. A. McIntosh, and F. H. Arnold, "Enzyme-Controlled Nitrogen-Atom Transfer Enables Regiodivergent C–H Amination," *Journal of the American Chemical Society* 136 (2014): 15505–15508, https://doi.org/10.1021/ja509308v.

23. C. K. Prier, R. K. Zhang, A. R. Buller, S. Brinkmann-Chen, and F. H. Arnold, "Enantioselective, Intermolecular Benzylic C–H Amination Catalysed by an Engineered Iron-Haem Enzyme," *Nature Chemistry* 9 (2017): 629–634, https://doi.org/10.1038/nchem.2783.

24. B. F. Fisher, H. M. Snodgrass, K. A. Jones, M. C. Andorfer, and J. C. Lewis, "Site-Selective C–H Halogenation Using Flavin-Dependent Halogenases Identified via Family-Wide Activity Profiling," *ACS Central Science* 5 (2019): 1844–1856.

25. J. R. Marshall, P. Yao, S. L. Montgomery, et al., "Screening and Characterization of a Diverse Panel of Metagenomic Imine Reductases for Biocatalytic Reductive Amination," *Nature Chemistry* 13 (2021): 140–148, https://doi.org/10.1038/s41557-020-00606-w.

26. L. E. Zetzsche, J. A. Yazarians, S. Chakrabarty, et al., "Biocatalytic Oxidative Cross-Coupling Reactions for Biaryl Bond Formation," *Nature* 603 (2022): 79–85, https://doi.org/10.1038/s41586-021-04365-7.

27. M. Gajdoš, J. Wagner, F. Ospina, A. Köhler, M. K. M. Engqvist, and S. C. Hammer, "Chiral Alcohols From Alkenes and Water: Directed Evolution of a Styrene Hydratase," *Angewandte Chemie International Edition* 62 (2023): e202215093.

28. S. E. Champagne, C.-H. Chiang, P. M. Gemmel, C. L. Brooks, III, and A. R. H. Narayan, "Biocatalytic Stereoselective Oxidation of 2-Arylindoles," *Journal of the American Chemical Society* 146 (2024): 2728–2735, https://doi.org/10.1021/jacs.3c12393.

29. I. T. Jolliffe and J. Cadima, "Principal Component Analysis: A Review and Recent Developments," *Philosophical Transactions. Series A, Mathematical, Physical, and Engineering Sciences* 374 (2016): 20150202.

30. G. Casari, C. Sander, and A. Valencia, "A Method to Predict Functional Residues in Proteins," *Natural Structural Biology* 2 (1995): 171–178, https://doi.org/10.1038/nsb0295-171.

31. B. Wang and M. A. Kennedy, "Principal Components Analysis of Protein Sequence Clusters," *Journal of Structural and Functional Genomics* 15 (2014): 1–11, https://doi.org/10.1007/s10969-014-9173-2.

32. T. Shafee and M. A. Anderson, "A Quantitative Map of Protein Sequence Space for the Cis-Defensin Superfamily," *Bioinformatics* 35 (2019): 743–752, https://doi.org/10.1093/bioinformatics/bty697.

33. C. J. Vavricka, S. Takahashi, N. Watanabe, et al., "Machine Learning Discovery of Missing Links That Mediate Alternative Branches to Plant Alkaloids," *Nature Communications* 13 (2022): 1405, https://doi.org/10.1038/s41467-022-28883-8.

34. R. Hidese, K. Sakai, M. Takenaka, et al., "Identification of Subfamily Specific Residues Within Highly Active and Promiscuous Alcohol Dehydrogenases," *ACS Catalysis* 15 (2025): 11931–11943, https://doi.org/10.1021/acscatal.5c02764.

35. O. F. Brandenberg, R. Fasan, and F. H. Arnold, "Exploiting and Engineering Hemoproteins for Abiological Carbene and Nitrene Transfer Reactions," *Current Opinion in Biotechnology* 47 (2017): 102–111, https://doi.org/10.1016/j.copbio.2017.06.005.

36. P. S. Coelho, Z. J. Wang, M. E. Ener, et al., "A Serine-Substituted P450 Catalyzes Highly Efficient Carbene Transfer to Olefins In Vivo," *Nature Chemistry Biology* 9 (2013): 485–487, https://doi.org/10.1038/nchembio.1278.

37. T. Heel, J. A. McIntosh, S. C. Dodani, J. T. Meyerowitz, and F. H. Arnold, "Non-Natural Olefin Cyclopropanation Catalyzed by Diverse Cytochrome P450s and Other Hemoproteins," *Chembiochem* 15 (2014): 2556–2562, https://doi.org/10.1002/cbic.201402286.

38. J. G. Gober, A. E. Rydeen, E. J. Gibson-O'Grady, J. B. Leuthaeuser, J. S. Fetrow, and E. M. Brustad, "Mutating a Highly Conserved Residue in Diverse Cytochrome P450s Facilitates Diastereoselective Olefin Cyclo-

propanation," *Chembiochem* 17 (2016): 394–397, https://doi.org/10.1002/cbic.201500624.

39. K. Suzuki, Y. Shisaka, J. K. Stanfield, Y. Watanabe, and O. Shoji, "Enhanced Cis- and Enantioselective Cyclopropanation of Styrene Catalysed by Cytochrome P450BM3 Using Decoy Molecules," *Chemical Communications* 56 (2020): 11026–11029, https://doi.org/10.1039/D0CC04883F.

40. B. Wang, C. You, G. Xu, and Y. Ni, "Modulating Stereoselectivity and Catalytic Efficiency of Carbenoid Reactions Catalysed by Self-sufficient P450s," *Catalysis Science & Technology* 14 (2024): 835–839, https://doi.org/10.1039/D3CY01258A.

41. M. Bordeaux, V. Tyagi, and R. Fasan, "Highly Diastereoselective and Enantioselective Olefin Cyclopropanation Using Engineered Myoglobin-Based Catalysts," *Angewandte Chemie International Edition* 54 (2015): 1744–1748, https://doi.org/10.1002/anie.201409928.

42. P. Bajaj, G. Sreenilayam, V. Tyagi, and R. Fasan, "Gram-Scale Synthesis of Chiral Cyclopropane-Containing Drugs and Drug Precursors With Engineered Myoglobin Catalysts Featuring Complementary Stereoselectivity," *Angewandte Chemie International Edition* 55 (2016): 16110–16114, https://doi.org/10.1002/anie.201608680.

43. A. Tinoco, Y. Wei, J.-P. Bacik, et al., "Origin of High Stereocontrol in Olefin Cyclopropanation Catalyzed by an Engineered Carbene Transferase," *ACS Catalysis* 9 (2019): 1514–1524, https://doi.org/10.1021/acscatal.8b04073.

44. M. Pott, M. Tinzl, T. Hayashi, et al., "Noncanonical Heme Ligands Steer Carbene Transfer Reactivity in an Artificial Metalloenzyme**," *Angewandte Chemie International Edition* 60 (2021): 15063–15068, https://doi.org/10.1002/anie.202103437.

45. L. Villarino, K. E. Splan, E. Reddem, et al., "An Artificial Heme Enzyme for Cyclopropanation Reactions," *Angewandte Chemie International Edition* 57 (2018): 7785–7789, https://doi.org/10.1002/anie.201802946.

46. R. Stenner, J. W. Steventon, A. Seddon, and J. L. R. Anderson, "A De Novo Peroxidase Is Also a Promiscuous yet Stereoselective Carbene Transferase," *Proceedings National Academy of Science USA* 117 (2020): 1419–1428, https://doi.org/10.1073/pnas.1915054117.

47. I. Kalvet, M. Ortmayer, J. Zhao, et al., "Design of Heme Enzymes With a Tunable Substrate Binding Pocket Adjacent to an Open Metal Coordination Site," *Journal of the American Chemical Society* 145 (2023): 14307–14315, https://doi.org/10.1021/jacs.3c02742.

48. K. Hou, W. Huang, M. Qi, et al., "De Novo Design of Porphyrin-Containing Proteins as Efficient and Stereoselective Catalysts," *Science* 388 (2025): 665–670, https://doi.org/10.1126/science.adt7268.

49. G. Fittolani, D. A. Kutateladze, A. Loas, S. L. Buchwald, and B. L. Pentelute, "Automated Flow Synthesis of Artificial Heme Enzymes for Enantiodivergent Biocatalysis," *Journal of the American Chemical Society* 147 (2025): 4188–4197, https://doi.org/10.1021/jacs.4c13832.

50. A. M. Knight, S. B. J. Kan, R. D. Lewis, O. F. Brandenberg, K. Chen, and F. H. Arnold, "Diverse Engineered Heme Proteins Enable Stereodivergent Cyclopropanation of Unactivated Alkenes," *ACS Central Science* 4 (2018): 372–377, https://doi.org/10.1021/acscentsci.7b00548.

51. S. Kato, K. Takeuchi, M. Iwaki, K. Miyazaki, K. Honda, and T. Hayashi, "Chitin- and Streptavidin-Mediated Affinity Purification Systems: A Screening Platform for Enzyme Discovery," *Angewandte Chemie International Edition* 62 (2023): e202303764, https://doi.org/10.1002/anie.202303764.

52. H. Wajcman, L. Kiger, and M. C. Marden, "Structure and Function Evolution in the Superfamily of Globins," *C R Biology* 332 (2009): 273–282, https://doi.org/10.1016/j.crvi.2008.07.026.

53. A. D. Frey and P. T. Kallio, "Bacterial Hemoglobins and Flavohemoglobins: Versatile Proteins and Their Impact on Microbiology and Biotechnology," *FEMS Microbiology Review* 27 (2003): 525–545, https://doi.org/10.1016/S0168-6445(03)00056-1.

54. B. J. Wittmann, A. M. Knight, J. L. Hofstra, S. E. Reisman, S. B. Jennifer Kan, and F. H. Arnold, "Diversity-Oriented Enzymatic Synthesis of Cyclopropane Building Blocks," *ACS Catalysis* 10 (2020): 7112–7116, https://doi.org/10.1021/acscatal.0c01888.

55. T. A. K. Freitas, J. A. Saito, S. Hou, and M. Alam, "Globin-coupled Sensors, Protoglobins, and the Last Universal Common Ancestor," *Journal of Inorganic Biochemistry* 99 (2005): 23–33, https://doi.org/10.1016/j.jinorgbio.2004.10.024.

56. L. Giangiacomo, A. Ilari, A. Boffi, V. Morea, and E. Chiancone, "The Truncated Oxygen-Avid Hemoglobin From *Bacillus Subtilis*: X-Ray Structure and Ligand Binding Properties," *Journal of Biological Chemistry* 2005, 280, 9192–9202.

57. A. Ilari, P. Kjelgaard, C. v. Wachenfeldt, B. Catacchio, E. Chiancone, and A. Boffi, "Crystal Structure and Ligand Binding Properties of the Truncated Hemoglobin From *Geobacillus Stearothermophilus*," *Archives of Biochemistry and Biophysics* 457 (2007): 85–94, https://doi.org/10.1016/j.abb.2006.09.033.

58. S. F. Altschul, W. Gish, W. Miller, E. W. Myers, and D. J. Lipman, "Basic Local Alignment Search Tool," *Journal of Molecular Biology* 215 (1990): 403–410, https://doi.org/10.1016/S0022-2836(05)80360-2.

59. S. McGinnis and T. L. Madden, "BLAST: At the Core of a Powerful and Diverse Set of Sequence Analysis Tools," *Nucleic Acids Research* 32 (2004): W20–W25, https://doi.org/10.1093/nar/gkh435.

60. T. G. Schmidt and A. Skerra, "The Strep-Tag System for One-Step Purification and High-Affinity Detection or Capturing of Proteins," *Nature Protocols* 2 (2007): 1528–1535, https://doi.org/10.1038/nprot.2007.209.

61. J. B. Wittenberg, M. Bolognesi, B. A. Wittenberg, and M. Guertin, "Truncated Hemoglobins: A New Family of Hemoglobins Widely Distributed in Bacteria, Unicellular Eukaryotes, and Plants," *Journal of Biological Chemistry* 277 (2002): 871–874, https://doi.org/10.1074/jbc.R100058200.

62. H. Sarma, B. Sharma, S. Tiwari, and A. Mishra, "Truncated Hemoglobins: A Single Structural Motif With Versatile Functions in Bacteria, Plants and Unicellular Eukaryotes," *Symbiosis* 39 (2005): 151–158.

63. S. Hou, T. Freitas, R. W. Larsen, et al., "Globin-Coupled Sensors: A Class of Heme-Containing Sensors in Archaea and Bacteria," *Proceedings National Academy of Science USA* 98 (2001): 9353–9358, https://doi.org/10.1073/pnas.161185598.

64. M. Tarnawski, T. R. M. Barends, and I. Schlichting, "Structural Analysis of an Oxygen-Regulated Diguanylate Cyclase," *Acta Crystallographica Section D* 71 (2015): 2158–2177, https://doi.org/10.1107/S139900471501545X.

65. S. Mathur, S. K. Yadav, K. Yadav, S. Bhatt, and S. Kundu, "A Novel Single Sensor Hemoglobin Domain From the Thermophilic Cyanobacteria *Thermosynechococcus Elongatus* BP-1 Exhibits Higher pH but Lower Thermal Stability Compared to Globins From Mesophilic Organisms," *International Journal of Biological Macromolecules* 2023, 240, 124471.

66. J. A. Gerlt, J. T. Bouvier, D. B. Davidson, et al., "Enzyme Function Initiative-Enzyme Similarity Tool (EFI-EST): A Web Tool for Generating Protein Sequence Similarity Networks," *Biochimica et Biophysica Acta (BBA)—Proteins and Proteomics* 1854 (2015): 1019–1037, https://doi.org/10.1016/j.bbapap.2015.04.015.

67. R. Zallot, N. Oberg, and J. A. Gerlt, "The EFI Web Resource for Genomic Enzymology Tools: Leveraging Protein, Genome, and Metagenome Databases to Discover Novel Enzymes and Metabolic Pathways," *Biochemistry* 58 (2019): 4169–4182, https://doi.org/10.1021/acs.biochem.9b00735.

68. N. Oberg, R. Zallot, and J. A. Gerlt, "EFI-EST, EFI-GNT, and EFI-CGFP: Enzyme Function Initiative (EFI) Web Resource for Genomic Enzymology Tools," *Journal of Molecular Biology* 435 (2023): 168018, https://doi.org/10.1016/j.jmb.2023.168018.

69. J. Abramson, J. Adler, J. Dunger, et al., "Accurate Structure Prediction of Biomolecular Interactions With AlphaFold 3," *Nature* 2024, 630, 493–500.

70. A. Pesce, M. Bolognesi, and M. Nardini, in *Advances in Microbial Physiology, Vol.* 63, ed, R. K. Poole (Academic Press, 2013), 49–78, https://doi.org/10.1016/B978-0-12-407693-8.00002-9.

71. B. W. Brandt, K. A. Feenstra, and J. Heringa, "Multi-Harmony: Detecting Functional Specificity From Sequence Alignment," *Nucleic Acids Research* 38 (2010): W35–W40, https://doi.org/10.1093/nar/gkq415.

72. S. Kato, H. Nishiwaki, K. Endo, and T. Hayashi, "Radical Ring-Opening Reaction of Non-Activated Oximes Catalyzed by Aldoxime Dehydratases," *Angewandte Chemie International Edition* 64 (2025): e202511590, https://doi.org/10.1002/anie.202511590.

73. K. Katoh, K. Misawa, K. i. Kuma, and T. Miyata, "MAFFT: A Novel Method for Rapid Multiple Sequence Alignment Based on Fast Fourier Transform," *Nucleic Acids Research* 30 (2002): 3059–3066, https://doi.org/10.1093/nar/gkf436.

74. D. Pelleg and A. W. Moore, in *Proceedings of the Seventeenth International Conference on Machine Learning* (Morgan Kaufmann Publishers Inc., 2000), 727–734.

Supporting Information

Additional supporting information can be found online in the Supporting Information section.

Supporting File 1: General information, experimental details for sample preparation, Figures S1–S31, Tables S1–S4, Supporting Data 1–4, and amino acid sequences. The authors have cited additional references within the Supporting Information [73, 74].