Directed Evolution of Protoglobin Optimizes the Enzyme Electric Field

Shobhit S. Chaturvedi [a], Santiago Vargas [a], Pujan Ajmera [a], Anastassia N. Alexandrova [a,*]
[a] Department of Chemistry and Biochemistry, University of California, Los Angeles, California 90095, United States.
* Corresponding author email: ana@chem.ucla.edu

Abstract. To unravel why computational design fails in creating viable enzymes, while directed evolution (DE) succeeds, our research delves into the laboratory evolution of Protoglobin. DE has adapted this protein to efficiently catalyze carbene transfer reactions. We show that the previously proposed enhanced substrate access and binding alone cannot account for increased yields during DE. The 3D electric field in the entire active site is tracked through protein dynamics, clustered using the affinity propagation algorithm, and subjected to principal component analysis. This analysis reveals notable changes in the electric field with DE, where distinct field topologies influence transition state energetics and mechanism. A chemically meaningful field component emerges and takes the lead during DE and facilitates crossing the barrier to carbene transfer. Our findings underscore intrinsic electric field dynamic's influence on enzyme function, the ability of the field to switch mechanisms within the same protein, and the crucial role of the field in enzyme design.

Main

Nature has evolved enzymes as remarkably proficient biocatalysts to facilitate a vast array of chemical transformations.[1] Through billions of years of evolutionary fine-tuning, natural enzymes have unlocked extraordinary catalytic power, selectivity, and efficiency.[2–4] The drive to push beyond nature's set of catalyzed reactions, and achieve similarly efficient catalysis for other transformations has led to innovative approaches in modifying enzymes,[5–8] and designing them *de novo*.[9] Indeed, enzyme design has become a frontier of innovation, with the goal of customizing enzymes for the sustainable production of a variety of chemicals, pharmaceuticals, and materials.

Creating highly active enzymes from scratch remains an unsolved task, despite the potential.[10] The initially designed enzymes often need more catalytic vigor, and are subjected to subsequent rounds of directed evolution (DE) to reach appreciable activity levels.[11] DE serves as an optimization step that provides designed enzymes with properties absent from initial designs, from improving enantioselectivity in rhodium-catalyzed artificial metalloenzymes,[12] to dramatic boosts in the activity of computationally designed retro-aldolases,[13,14] and Kemp eliminases,[15,16] to name just a few examples. DE produces stunning enhancements of k_{cat}/K_M of over 4400-fold.[12–17] The need to evolve designed enzymes to attain catalytic viability underscores significant gaps in *de novo* design protocols. Understanding what DE contributes to enzyme design is crucial, as DE appears to provide essential elements that are missing in initial designs, potentially unlocking key strategies for efficient enzyme design in the future.

We study the directed evolution of *Aeropyrum pernix* Protoglobin, a Fe-heme protein, which was evolved to perform a new-to-nature selective carbene transfer to catalyze cyclopropanation of benzyl acrylate (Figure 1).[18,19] Mutations introduced by DE are dispersed throughout the protein structure, located both close to the Fe-center (F145Q, I149L, Y60A, W59L), and as far as >15 Å away from it (F175L, C102S, V63R), and include both hydrophobic and

hydrophilic residues. We use this rich evolutionary journey to gain an understanding of how DE can imbue new catalytic functions into an enzyme. We perform and analyze replica molecular dynamics simulations of wild-type (WT) Protoglobin and four evolved variants (LVRQ, LVRQL, GLVRSQL, GLAVRSQLL), initially focusing on substrate access and binding improvements at the active site. Upon indications of changes in electrostatic preorganization at the active site along the evolutionary pathway, we develop and utilize a novel framework to study the dynamics of the heterogeneous electric field in the active site, combining electric field topological analysis, high-throughput computation, and graph compression algorithms for a comprehensive picture. Finally, we correlate changes in electrostatic preorganization with experimental yield through QM/MM reaction mechanism calculations. This workflow illuminates the critical factors DE exploits to enhance enzyme catalysis—insights crucial for refining enzyme design protocols.

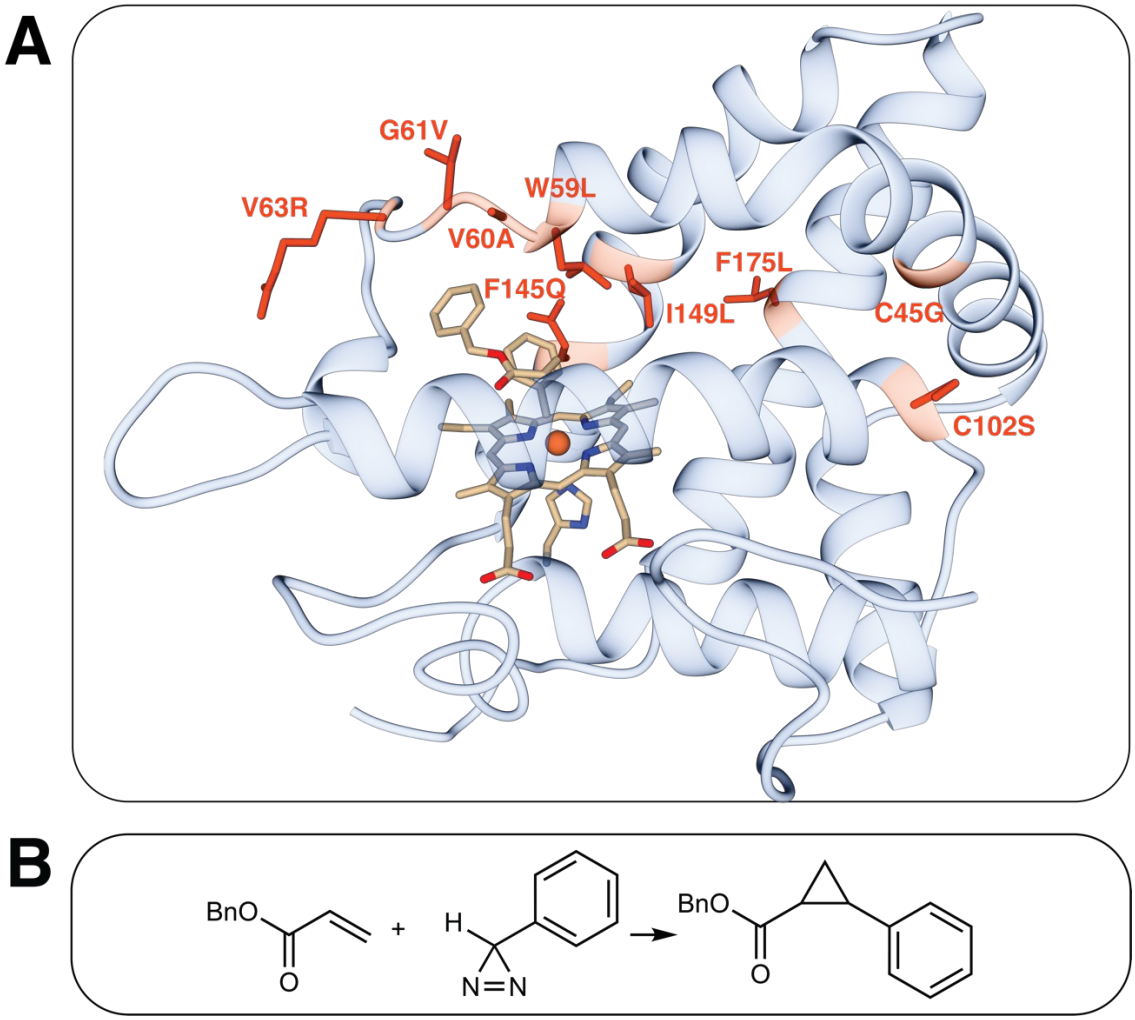

Figure 1. (A) Protoglobin with directed evolution mutation sites highlighted in red and labeled with the bound substrate. (B) the carbene transfer reaction being optimized along the directed evolution path.

Results and Discussions

Can substrate binding explain the yield increase?

Based on the microcrystal electron diffraction structure of the GLVRSQL Protoglobin variant, it was proposed that DE facilitates the new-to-nature catalysis by enhancing substrate access to the active site.[18,19] We analyzed the substrate access to the active site of Protoglobin, by measuring the distance between the terminal C1 atom of the benzyl acrylate substrate and the reactive CC atom of the carbene across the five replica molecular dynamics (MD) simulations of 100 ns each, for all variants (**Figure 2A**). In agreement with experiments, the mean distance of the substrate to the active site was high in the WT enzyme, measuring 17.10 ± 9.25 Å, suggesting the benzyl acrylate substrate stays away from the active site and has a very low chance of undergoing catalysis. However, during DE, the mean distance reduced in LVRQ (7.78 ± 3.53 Å) and LVRQL (8.87 ± 8.72 Å), signifying an improvement in substrate accessibility to the active site. The large standard deviation observed for LVRQ can be attributed to the divergent behavior of the benzyl acrylate substrate in the active site (see SI **Figure S1**). The reduction in substrate distance from the active site was further pronounced in the GLVRSQL variant (4.44 ± 0.90 Å), suggesting a significant enhancement in substrate entry and stabilization within the active site. However, in the GLAVRSQLL variant, the distance of the substrate to the active site remained comparable to GLVRSQL (4.53 ± 1.09 Å), while the experimental yield dramatically increased. This observation challenges the notion that solely substrate access to the active site dictates yield enhancements. While enhanced substrate access might be a major contributor to yield improvement from WT to GLVRSQL (0-6%), it could not explain the drastic increase in yield going from GLVRSQL to GLAVRSQLL (6-28%).[18]

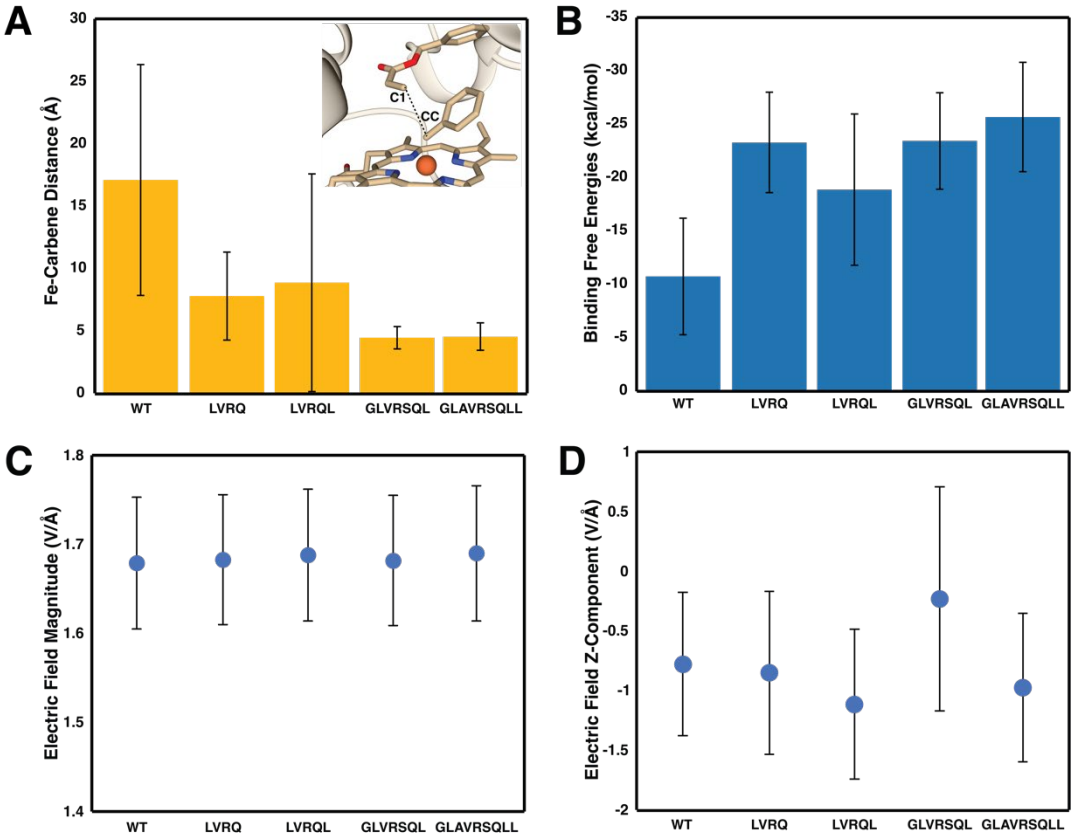

Figure 2. Initial parameters investigated as the cause of higher reactivity along DE path. (A) The mean and standard deviation of Fe-Carbene distance for all MD trajectories across all variants. (B) The mean and standard deviation of substrate-protein binding free energies ($\Delta G_{binding}$). (C) The total electric field magnitude computed on the Fe-Carbene bond of IPC for all systems across replica molecular dynamics. (D) The z-component of the electric field computed on the Fe-Carbene bond of IPC for all systems across replica molecular dynamics.

Further, we sought to explore if enhanced benzyl acrylate substrate binding is the reason for the observed yield increase (**Figure 2B**). MMGBSA binding free energy calculations were performed, across the five replica MDs for each variant. Notably, the substrate binding in the evolved variants (LVRQ: -23.30 ± 4.71, LVRQL: -18.90 ± 7.08, GLVRSQL: -23.46 ± 4.52, GLAVRSQLL: -25.69 ± 5.12 kcal/mol) was consistently stronger than in the WT (-10.76 ± 5.46 kcal/mol). However, the free energy of substrate binding did not reveal a discernible trend across the DE path. Specifically, LVRQ exhibited a higher binding energy, albeit at a larger distance from the active site, indicating strong substrate binding at non-active site regions and potentially contributing to the lower yield of carbene transfer (**Figure S2**). The mean binding free energies for GLVRSQL (-23.46 ± 4.52 kcal/mol) and GLAVRSQLL (-25.69 ± 5.12 kcal/mol) were within their respective standard deviations, and thus again, failing to provide a definitive explanation for the substantial yield increase from GLVRSQL to GLAVRSQLL. These findings suggest that while substrate binding energy is an important factor, it also does not adequately justify the enhancement in yield during DE of Protoglobin.

Electric field evolution during directed evolution

Now, we pivot towards analyzing if the electric fields generated by the studied enzyme changes, and its link to Protoglobin reactivity. Enzyme catalysis is often attributed to electrostatic preorganization and dynamics,[20–24] occasionally put in contradiction with each other.[25] We have previously observed that the reactivity of Fe-heme oxidoreductases is strongly regulated by the electric field form the protein scaffold, in addition to the regulation by the axial ligand to Fe.[26] Electrostatic preorganization has also been cited previously as a compass of directed evolution of Kemp eliminases.[27] To comprehensively address both electric fields and dynamics, we performed an electric field analysis over several replica MD trajectories to sample and compare the electrostatic behavior of the enzyme in a dynamic fashion.

We performed point electric fields calculations at the center of the Fe-carbene bond, for the carbene-substrate intermediate replica MDs of all systems. The mean electric field magnitude is evidently seen to not change meaningfully along DE, with only a very small decrease in the GLVRSQL variant (**Figure 2C**). A more noteworthy observation emerged when examining the projection of the electric field on the Fe-carbene bond (in the direction from Fe to the carbene CC atom). The projection shows larger variation across the mutants, especially in the field directionality (**Figure 2D**). This suggests that a point electric field-based analysis is not enough to capture the changes in the heterogeneous 3-dimensional (3D) electric field of the enzyme, therefore requiring a more comprehensive approach (introduced in **Figure 3**).

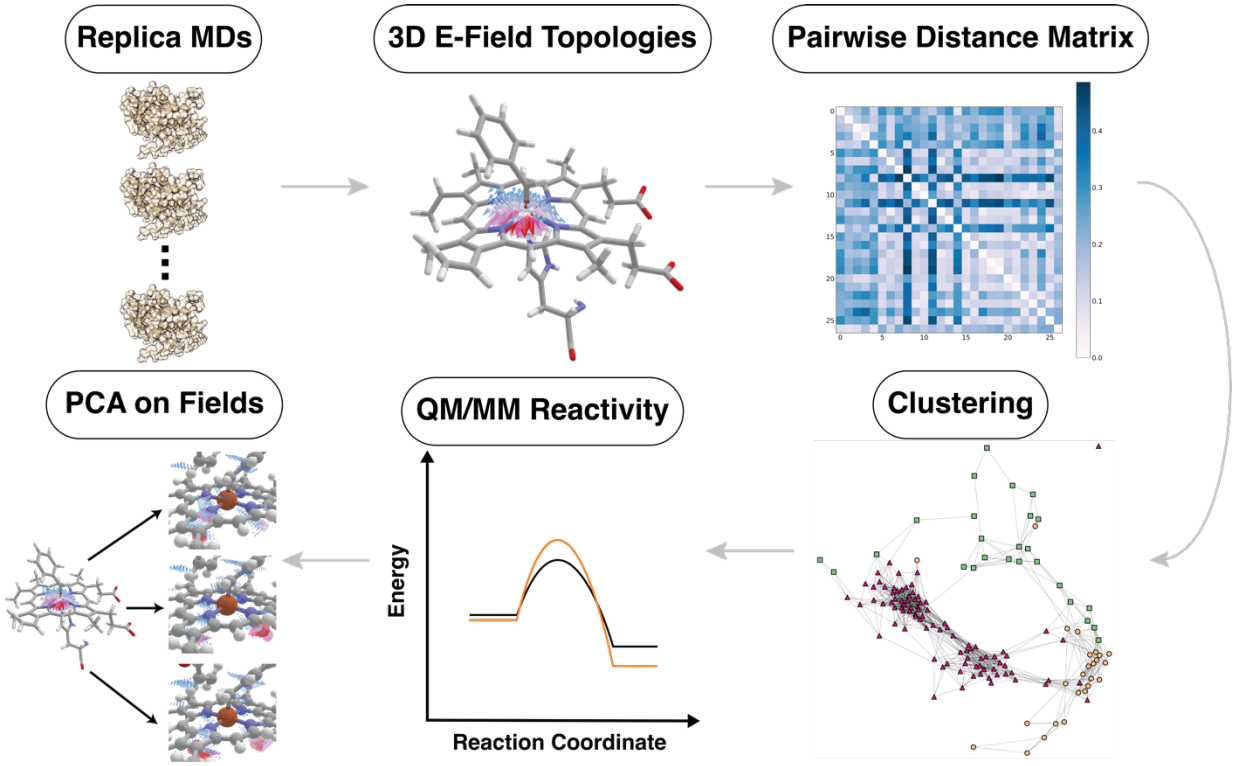

Figure 3. This study's approach measures electrostatic preorganization by analyzing the heterogeneous electric field topology across replica MD simulations. It further involves comparing these topologies using a pairwise distance matrix, clustering based on similarity, and then

quantifying reactivity through QM/MM methods. The reactivity difference is chemically elucidated using Principal Component Analysis.

We have previously developed a method to quantify the heterogeneous 3D electric field topology of an enzyme active site, and showed that 3D fields correlate with reactivity better than fields at a point or along a particular bond for ketosteroid isomerases.[28] This approach involves defining a volume of interest for electric field topology calculations, which, in this study, is a cubic box centered on the Fe-carbene bond (**Figure 4A**). The heterogeneous electric field (**Figure 4B**) was calculated for a total of 5,000 frames derived from 5 x 100 ns replica MD runs for each variant. To analyze this vast dataset, we employed an affinity propagation algorithm to cluster similar electric field topologies within a dynamical trajectory. Affinity propagation provides a distinct advantage by eliminating *a priori* knowledge of the number of clusters. This flexibility allows us to track the changes in the distribution and the number of clusters along the DE of Protoglobin - signaling how diverse or, inversely, tightly controlled the electric field is within the protein's active site. Additionally, this clustering algorithm yields a single best representative frame for each cluster, aiding visualization and further analysis of the 3D electric field that the active site samples. The predominant clusters (those representing >5% of the MDs) from each system were considered and subsequently compared using a distance matrix (**Figure 4D**). A distance closer to 0 indicates high similarity in the 3D heterogeneous electric field topologies, while a score of 1 indicates high dissimilarity, for example, between WT and evolved variants.

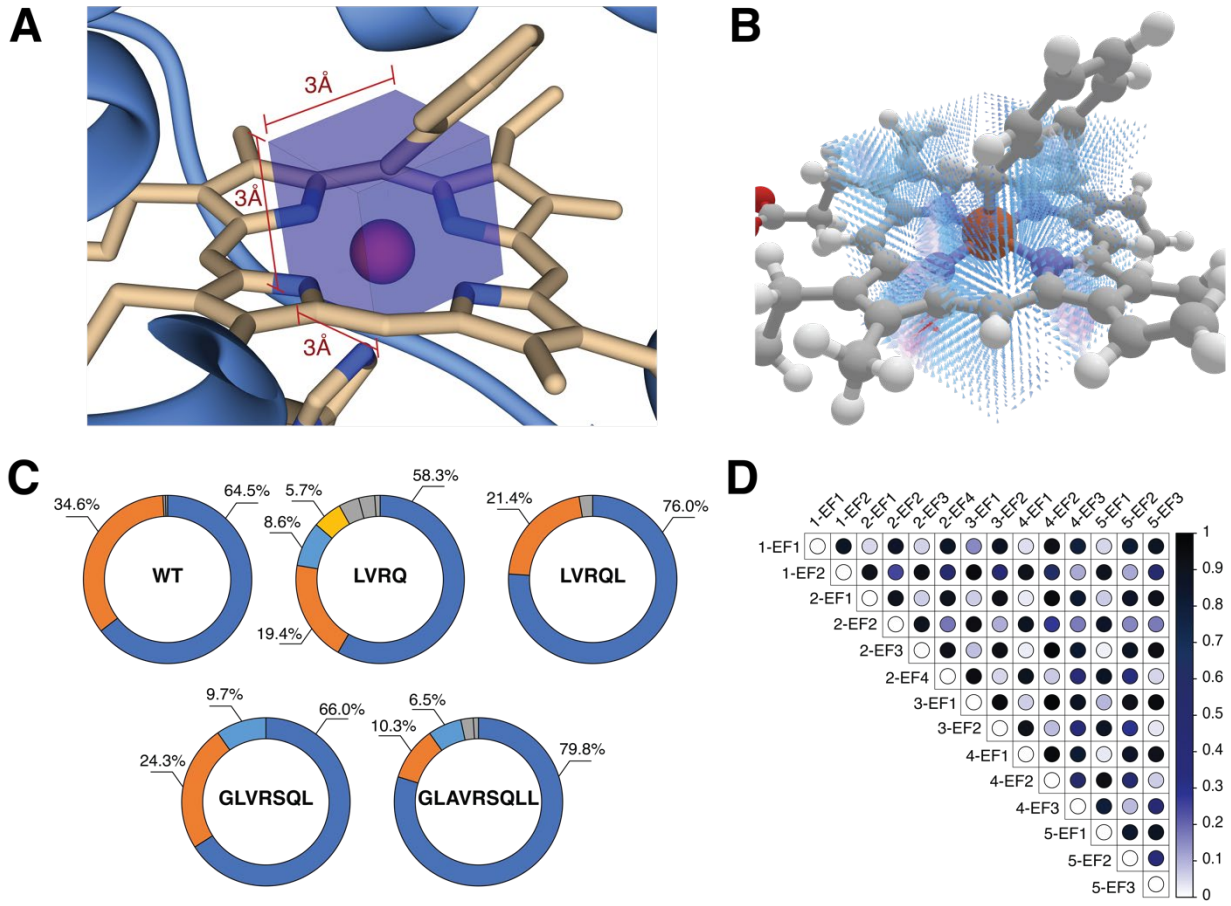

Figure 4. (A) Illustration of a 3Å box centered on the Fe-carbene bond for calculating the 3D heterogeneous electric field topology. (B) Example of a 3D heterogeneous electric field topology calculation. (C) Affinity Propagation clustering of electric field topologies for each variant, with blue indicating the most prevalent, orange the second, and green the third; clusters under 5% are in grey. (D) A pairwise distance matrix comparing the similarity (0) or difference (1) of electric field topology clusters across all systems. The first number in the labels indicate the stage of directed evolution (1=WT, 5=GLAVRSQLL), and the second number indicates how often the field topology is visited along the trajectory (1=the most frequently visited).

WT Protoglobin features two highly distinct 3D electric fields, with clusters WT-EF1 (visited by the system 64.53% of the time) and WT-EF2 (34.58%). The LVRQ variant introduces four electric field clusters, LVRQ-EF1 to LVRQ-EF4, with visitations of 58.32%, 19.36%, 8.60%, and 5.68%, respectively. LVRQ-EF1 and LVRQ-EF3 closely resemble each other and WT-EF1, while LVRQ-EF2 and LVRQ-EF4 diverge significantly, marking the introduction of two novel electric fields. LVRQL further evolves this pattern, showing two main fields: LVRQL-EF1 (76.04%) and LVRQL-EF2 (21.36%), which are derivatives of LVRQ's clusters, illustrating an ongoing modification from WT through DE. GLVRSQL presents three clusters: GLVRSQL-EF1 (65.98%), GLVRSQL-EF2 (24.34%), and GLVRSQL-EF3 (9.68%), with EF1 and EF3 showing regressive similarity to WT-EF1 and EF2, respectively, while GLVRSQL-EF2 (0.92 and 0.62 from WT-EFs) remains distinct, reflecting influences from LVRQ-EF4 and LVRQL-EF2. The final GLAVRSQLL variant has three clusters: GLAVRSQLL-EF1 (79.82%), GLAVRSQLL-EF2 (10.34%), and GLAVRSQLL-EF3 (6.46%). GLAVRSQLL-EF1 demonstrate nuanced similarities to WT-EF1 (0.05) and GLAVRSQLL-EF2 is closest related to GLVRSQL-EF3 and WT-EF2 (0.08 and 0.11, respectively), indicating evolutionary modifications. In contrast, GLAVRSQLL-EF3 introduces a distinct electric field, diverging from WT, which evolved throughout the directed evolution process. It now remains to be seen how these field variations impact the reactivity.

Link between evolving electric fields and reactivity changes

The carbene transfer reaction in the engineered Protoglobin proceeds through the iron porphyrin carbene (IPC) intermediate with the substrate bound nearby.[19] The IPC intermediate contains a highly reactive carbene carbon, which reacts with the double bond in the benzyl acrylate substrate, leading to the formation of two new carbon-carbon bonds and culminating in the formation of a cyclopropane ring embedded within the substrate. The IPC intermediate is capable of adopting three spin states, each potentially influencing the cyclopropanation pathway differently. However, most experimental evidence points to the existence of a closed-shell singlet spin state.[29,30] Unpaired electron states (triplet or open shell singlet) may lead to a stepwise process, while a closed shell singlet state favors a direct, concerted mechanism either synchronous or asynchronous, without intermediates.[31] To explore the cyclopropanation reactivity, we performed hybrid quantum mechanics/molecular mechanics (QM/MM) calculations on the cluster centers for WT Protoglobin and the evolved variants. The calculations indicate the preferred spin state for the IPC complex is the closed shell singlet, favored over the triplet by 16 kcal/mol, with several attempts to converge the open shell singlet leading to the closed shell singlet. The spin preference for closed shell singlet is also supported by similarities in Fe-CC bond lengths between the

QM/MM optimized closed-shell singlet state (1.79 Å) and the crystallized Protoglobin IPC intermediate (1.74 Å), contrasting with the longer bond length (1.93 Å) in the triplet state.[19]

The reactivity calculations using a hybrid QM/MM method were conducted on all electric field clusters for variants containing the substrate within a reactive proximity (<5 Å) to the CC. The results showed that for the WT-RCs and EF1, EF2, and EF4 clusters for the LVRQ variant, the substrate was positioned at distances greater than the reactive range from the CC atom, classifying these states as unreactive (see SI **Table S1**). Consistent with other carbene transfer studies, all reactive clusters for the LVRQ, LVRQL, GLVRSQL, and GLAVRSQLL variants, with closed shell singlet spin state, demonstrated a concerted reaction mechanism, lacking stable intermediates and characterized by the asynchronous formation and breaking of bonds.[31–33] Initially, the CC and substrate C1 atom bond formation and elongation of the Fe-CC bond is favored, followed by complete breaking of the Fe-CC bond, culminating in the bond formation between CC and C2. For the LVRQ variant with EF3, the Gibbs free energy barrier was identified as 21.1 kcal/mol, coupled with a product stabilization energy of -13.6 kcal/mol. The LVRQL variant exhibited a free energy barrier ranging between 19.6 and 22.6 kcal/mol and product stabilization energies between -20.8 and -33.5 kcal/mol. The GLVRSQL variant showed a barrier range of 22.2 to 31.8 kcal/mol and product stabilization energies between -15.7 and -16.2 kcal/mol. Lastly, the GLAVRSQLL variant displayed a barrier range from 18.1 to 35.9 kcal/mol with product stabilization energies between -20.6 and -37.40 kcal/mol.

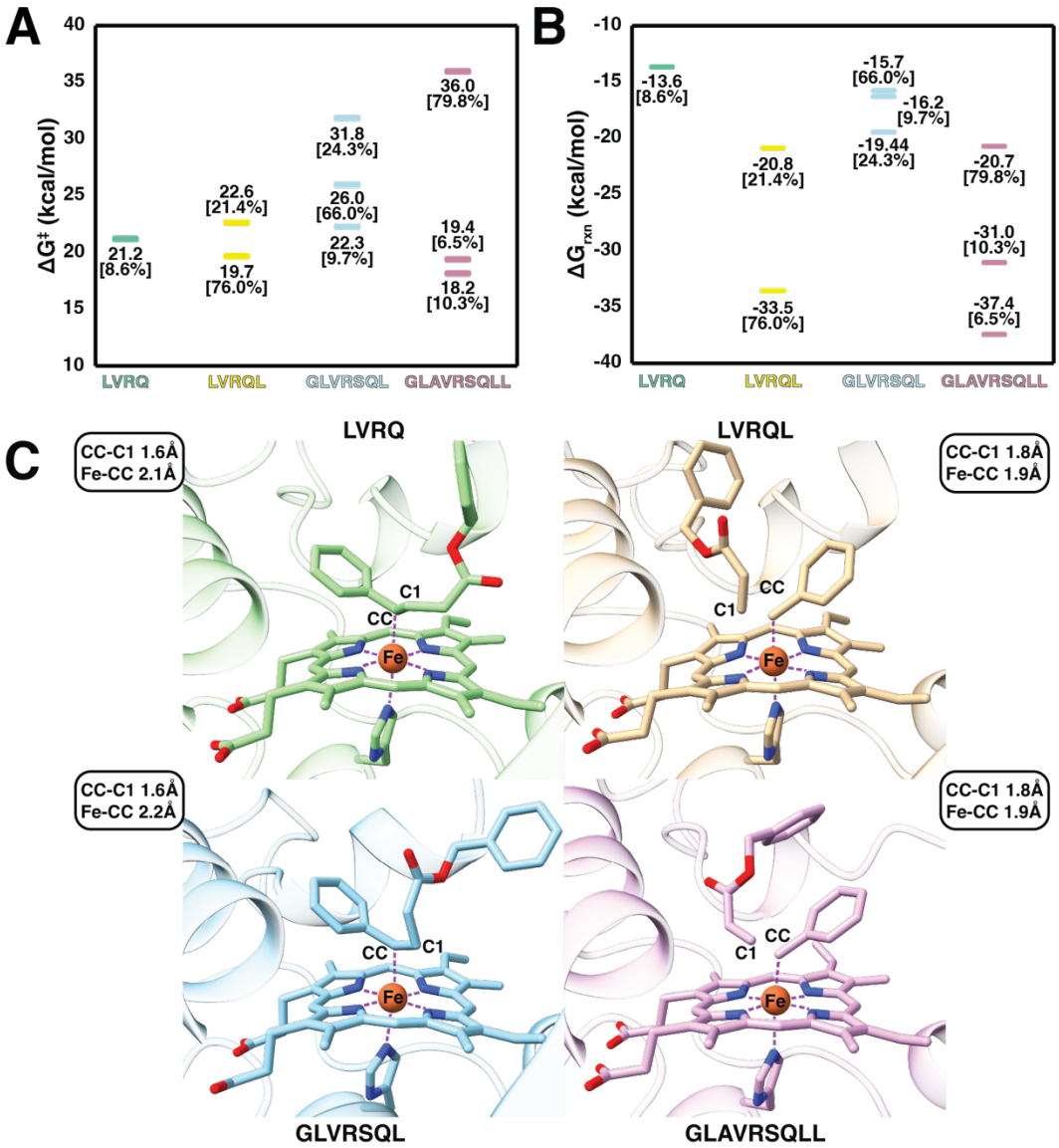

Figure 5. (A) Transition state free energy barriers for reactive clusters from each variant; (B) Product stabilization energies for reactive clusters from each variant. (C) Observed transition states from the best performing cluster centers of each variant. Transition state and product stabilization energies/structures were obtained from reaction path scans.

The results indicate that LVRQ and LVRQL produce lower free energy barriers and, in LVRQL, also a significant product stabilization, aligning with some experimental activities observed.[18] This suggests the observed low experimental reactivity in the LVRQ and LVRQL variants is likely not due to a lack of intrinsic reactivity but rather from rare visiting of reactive configurations, as suggested by the mean distances from molecular dynamics (MD) simulations (7.78 ± 3.53 Å for LVRQ and 8.87 ± 8.72 Å for LVRQL). Conversely, the primary reason for the low experimental yields in GLVRSQL appears to be the absence of effective electric fields necessary for lowering the barrier to the cyclopropanation reaction and stabilizing the product, despite the close proximity of the benzyl acrylate substrate (MD mean distance of 4.44 ± 0.90 Å).

This underscores the principle that mere access of the substrate to the active site is insufficient for high yield; the presence of a conducive electric field is critical for enhancing reactivity. The GLAVRSQLL variant, alongside the close binding of the benzyl acrylate substrate (MD mean distances of 4.53 ± 1.09 Å), has an effective electric field leading to both low energy barriers and favorable product stabilization, indicating its proficiency in catalyzing the cyclopropanation reaction. This agrees with and rationalizes the yield increase from approximately 8% in GLVRSQLL to about 28% in GLAVRSQLL.

Intriguingly, QM/MM calculations also reveal that the nature of the reaction TS within different enzyme variants is significantly influenced by the electric fields present. We identified two distinct types of TSs. The first type, observed in the most efficient EF clusters of the variants LVRQL and GLAVRSQLL, is characterized by the formation of the CC-C1 bond accompanied by a slight elongation of the Fe-CC bond. In contrast, the second type of TS, found in LVRQ and GLVRSQL variants for the same cyclopropanation reaction, showcases a fully formed CC-C1 bond and a complete dissociation of the Fe-CC bond. Thus, the distinct 3D electric fields can facilitate a mechanism change of the cyclopropanation reaction. Moreover, within the GLVRSQL variant, EF3 and EF1 both exhibit TS of the second type, whereas EF2 presents a TS of the first type. Hence, enzyme's dynamically visiting diverse electric fields has the potential for diverse mechanistic pathways to be active within the same enzyme.

Principal Component Analysis of the Fields

To link the 3D heterogeneous electric fields to reactivity in a chemically meaningful manner, we employed Principal Component Analysis (PCA).[34] We mapped cluster centers to PCA components constructed from the compiled set of electric fields across all trajectories for each variant. This yields a single basis for electric field variability within the protein active site. The population density of each mutant, as illustrated in **Figure S3**, is mapped across PC0-9 components. This mapping reveals that every mutant, including GLVRSQL and GLAVRSQLL, exhibits significant variance from the WT Protoglobin along several PC components, confirming that DE influences the electric field and its dynamics within the active site considerably. The most dramatic shift between variants GLVRSQL and GLAVRSQLL, the mutations that incur the greatest change in activity, is observed along component PC9 (**Figure 6A**). The population density of GLAVRSQLL shifts positively along PC9, suggesting a robust alignment of its electric field with this PC. The findings that the most pronounced changes occur in higher-order components point to a multifaceted impact of mutations on the electric field's characteristics, re-emphasizing that the full spectrum of electric field components, rather solely the dominant one, must be analyzed to elucidate the role of fields in the catalytic process.

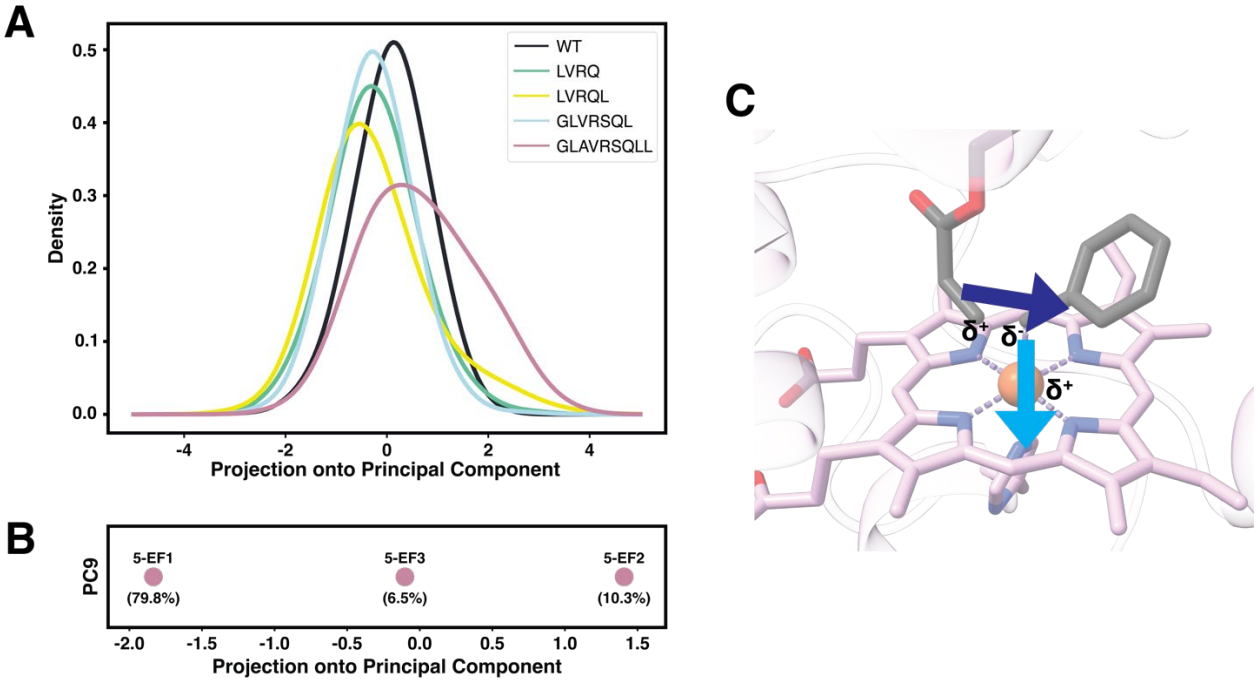

Figure 6. (A) Distribution of structures from replica molecular dynamics of all systems across the Principal Component 9. (B) Projections of GLAVRSQLL electric field cluster centers on PC9. (C) Schematic of the PC9 direction plotted on **TS-GLAVRSQLL-EF2** with the relative partial charges polarization marked on the atoms involved in bond rearrangements.

The isolated PC9 can be analyzed visually. It is curvy, and defined by two main directions: one tracing the path from the CC atom to the Fe and the other – from the C1 atom of the benzyl acrylate to the CC (**Figure 6C, Figure S4**). This field is straightforwardly linked to chemistry: in the TSs, the C1-CC bond is formed, and the electron density shifts from CC to C1 – a shift aided by the field of opposing direction. Similarly, the field pointing from CC to Fe aids the Fe-CC bond breaking in the TS. This implies that the intrinsic electric field alignment with PC9 in GLAVRSQLL facilitates the barrier crossing. This relationship becomes even clearer when we plot the electric field clusters along PC9. We observe that the degree of alignment with PC9 in GLAVRSQLL directly corresponds with the free energy barrier (**Figure 5A, Figure 6B**). Hence, the efficient catalysis observed in GLAVRSQLL is largely driven by a shift in its intrinsic electric field toward the positive direction of PC9, which plays a key role in stabilizing its TS. The development of a PC9-type electric field appears to be a key achievement of DE of Protoglobin.

Conclusions

While computational design often struggles to enable enzymes to catalyze new chemical reactions, DE has emerged as a potent method for imparting novel catalytic abilities to enzymes. This contrast poses a crucial dichotomy: despite being effective, DE is a black box method where mechanisms for improved activity are obfuscated by the enzyme complexity. We shed new light on one possible mechanism by studying Protoglobin, a protein that, through DE, has developed the ability to catalyze carbene transfer reactions, leading to the cyclopropanation of benzyl acrylate. Initial analysis of multiple MD simulations of wild-type Protoglobin and its four evolved variants indicated that merely enhancing substrate access and binding to the active site does not fully

explain the improved cyclopropanation yield. Therefore, we turned our attention to the enzyme's electrostatic preorganization. We have developed a detailed and broadly applicable protocol to measure the 3D electric field topology and dynamics, and analyzed and compared these dynamic fields along the DE path using an affinity propagation clustering algorithm. We discovered significant alterations in the active site electric field as Protoglobin evolved. Through PCA, we identified a chemically meaningful field component that emerges and takes the lead during DE and facilitates crossing the barrier to carbene transfer. The catalytic role of the evolved electric field was confirmed by QM/MM mechanistic calculations. These calculations revealed that the nature of the reaction TS (concerted Fe-CC bond breaking and C1-CC bond formation, or asynchronous and led by the Fe-CC bond breaking) can be altered by the field geometry. In summary, fine-tuning the global electric field in the active site appears to be the key achievement of DE and is, therefore, an aspirational goal for *de novo* enzyme design.

Methods

Our developed methodology has six primary components, visualized in **Figure 3**. We use molecular dynamics (MD) simulations to sample configurations of the protein. We use methods in field analysis to calculate the point electric field and electric field topology at the active site throughout the molecular dynamics trajectories. These topologies are compared using statistical distance metrics to obtain a distance matrix for each trajectory. To analyze how these field topologies change, we then use clustering on the distance matrices to obtain representative "snapshots" of the electric field at the active site. These snapshots are subjected to quantum mechanical/classical mechanical (QM/MM) reaction path calculations and principal component analysis (PCA).

System Preparation and Molecular Dynamics: We performed MD simulations on the carbene-substrate intermediate of the WT Protoglobin and the four directed evolved variants—LVRQ, LVRQL, GLVRSQL, and GLAVRSQLL. When compared to the WT, the LVRQ evolved variant presents mutations W59L, G60V, F145Q, and V63R. The LVRQL variant includes an additional I149L mutation. The GLVRSQL variant incorporates further C45G and C102S mutations. Lastly, the GLAVRSQLL evolved variant introduces additional mutations V60A, G61V, and F175L. We used the crystal structure of the Protoglobin GLVRSQL variant as the template to model the carbene-substrate intermediate of the other evolved variants. Since the crystal structure for WT Protoglobin was not available, we relied on a high-confidence Alphafold[35] model to simulate the carbene-substrate intermediate in WT Protoglobin. The carbene was modeled taking the Micro-ED crystal structure as a reference. The substrate benzyl acrylate was docked into the active site using AutoDock Vina.[36,37] The setup for the MD simulation was done via Amber 22 and AmberTools 22 modules.[38,39] The active site parameters for the carbene-substrate intermediates encompassing the heme, Fe, carbene, and an axial histidine were derived using AmberTool's Metal Center Parameter Builder (MCPB) v3.0.[40] The GAFF tool in Antechamber generated the topology for the substrate benzyl acrylate. The protonation states of the protein in the carbene-substrate intermediate were determined using Chimera routines, and the parameters for the rest of the protein were generated using the AmberFF19SB[41] force field. The LEaP module of Amber 22 neutralized the system by adding counterions. The system was then immersed into an OPC water box of at least 10 Å from the surface of the protein. Periodic boundary conditions were applied to the system, and long-range electrostatic interactions were calculated using the particle mesh Ewald method

with a cut-off distance of 8 Å. The SHAKE algorithm[42] was used to constrain bonds involving a hydrogen atom. The systems were minimized in two steps: using the steepest descent (10,000 steps), and (2) the conjugate gradient (10,000 steps) methods. During this phase, the protein's heavy atoms were restrained using a harmonic potential of 100 kcal mol^{-1} $Å^2$, and the protein's hydrogen atoms, along with solvent molecules, were minimized. Subsequently, the entire system underwent a comprehensive minimization process without any restraints via steepest descent (10,000 steps) and conjugate gradient (10,000 steps) methods. The system was then heated from 0 to 300 K in 50 ps using an NVT ensemble and then remained at 300K for another 50 ps. Next, A weakly constrained MD with constant pressure was performed to achieve uniform density in the systems, followed by equilibration MD in an NPT ensemble for 10 ns with restraints on the benzyl acrylate substrate to equilibrate it in the active site and then without any restraints for 2 ns. Finally, all production runs were performed using the GPU version of the AMBER 22 package. To enhance the credibility and precision of MD analysis, five replica MD simulations, each with a 100 ns duration, were performed.[43] RMSD and distance analysis was done with CPPTRAJ.[44] The binding free energies were calculated with the MMPBSA/MMGBSA module implemented in AMBER 22.[45]

Topological Electric Field Measurements and Comparison by Distance Metric

Most previous studies that incorporate electric fields as an analytical tool for protein activity use the protein structure in the crystalized variant or at a single frame within a larger MD trajectory.[26,28] This misses the effects of dynamics. We used a distance metric to construct a matrix of pairwise distances of electric fields along an entire MD trajectory. Every 5th frame of a 100 ns trajectory was used for the description of electric fields in the active site. The atomic charges were computed for the protein in each frame using ChargeFW2.[46] The field was calculated in a 3 Å box defined that is centered by the heme Fe – carbene carbon bond. The pairwise distances between each electric field's topology (see eq (1) and (2)), were computed, and subsequently fed into a graph clustering algorithm.

Our group has previously developed a distance metric to measure differences between 3-D electric fields.[28] This formulation enjoys important mathematical properties such as rotational, scalar, and translational invariance - essential properties when describing dynamical structures. The method samples points within a rectangular prism where linearizations of the electric field are computed and followed to calculate curvature. These lines are known as *streamlines* and provide a highly parallelizable compute unit, as each streamline can be calculated independently. We compute the curvature at the beginning and end of these streamlines with:

$$\kappa = \frac{\left||r'(t) \times r''(t)|\right|}{\left||r'(t)|\right|^3} \qquad eq\ (1)$$

Mean curvature values of the start and end points are compiled across each individual streamline along with the Euclidian distance between the start and end points to yield a histogram distribution of curvatures and mean distances for each electric field, a form of topology. This method computes the pairwise distance between two of these normalized distributions via the χ^2 distance:

$$\chi^2 : D(f,g) = \frac{1}{2}\sum_{i=1}^{N}\frac{(f[i]-g[i])^2}{f[i]+g[i]} \qquad eq\ (2)$$

With a defined distance comparing electric fields we can then create a graph where the edge lengths are the distances between two electric fields. This method requires the user to specify several parameters, including box size (Å), number of streamlines, and the step size (Å) for each linearization step along a streamline. The number of streamlines used for all calculations is 10000 with a step size of 0.001 - information on testing of box size can be found in the SI **Table S3**. Raw fields were preprocessed prior to input in visualization and clustering schemes. For clustering, we determined an upper boundary CPET distance above which edges were removed (**Figure S6**). This cutoff was the 10%ile distance from the collective distance matrix of all 5 WT runs. For affinity propagation, we also standardized the remaining distances. We used max_it = 10000 and 0.5 dampening. For PCA analysis, raw fields were used.

Affinity Propagation

Affinity Propagation intakes the "affinity" or similitude between different data points in a distance matrix, this can include non-connected graph nodes. We refer the audience to the original implementation of Affinity propagation[47] but will provide a brief outline of the method as follows. Affinity propagation is built on the iterative message passing of responsibility and availability between nodes in a graph. If **X** = {x_1…x_z} represents a set of data points and s(i, j) represents a similarity metric between points i and j. Responsibility r(i, k) describes how representative point k is for point i and availability a(i, k) measures how reasonable it is for k to pick i as a representative for itself. These values are updated using the following equations:

$$r(i,j) \leftarrow s(i,k) - \max_{k' \neq k}\{a(i,k') + s(i,k')\} \text{ (Update Responsibility matrix } r\text{)}$$

$$a(i,k) \leftarrow \min\left(0,\ r(k,k) + \sum_{i' \notin \{i,k\}} \max\left(0,\ r(i',\ k)\right)\right) \text{ (Update availability matrix } a\text{)}$$

Message-passing updates are repeated until convergence of representative structures/boundaries or a maximum number of iterations are reached.

PCA

Principal Component Analysis (PCA) is a widely-used dimensionality reduction algorithm that intakes descriptors on a dataset and performs a basis change to orthogonal components by order of descending variance – these new components are referred to as principal components.[34] PCA yields a few important statistical objects, namely the eigenvalues of the new principal components (PCs) and principal components themselves. Eigenvalues elucidate the variability in the dataset along the new basis and can be used to diagnose dataset dimensionality. The principal components themselves can be analyzed, along with the eigenvalues, to determine the directions of greatest variability in the dataset. We constructed PCA components from the compiled dataset of electric fields at every frame considered in the graph compression (all 5 mutants). This amounted to 25,000 electric fields where each field was centered at the heme-Fe. From here, we constructed a sampling mesh of 10 equidistant points in the six axial directions up to the boundary of 1.5 Å. This results in a 21 x 21 x 21 mesh of points spanning a 3 Å box – and thus, an input dimensionality of > 27,0000 points to the PCA algorithm. Remarkably, 5 components accounted for > 77% of the explained variation, 10 for > 95%, and 25 for > 98%. This shows that a small number of

components likely can be used to understand the variability in the electric fields – though we note that variability does not signify importance and therefore we extend our analysis to include several lower-variance PCs. We selected the 10 most important components and imposed the cluster centers' electric fields on to those components. This allows us to decompose the complex electric fields into simpler motifs for analysis and interpretation. Electric fields for the entire population of a mutant were analyzed along principal components to understand how the dynamic electric field evolves with mutations.

QM/MM Reaction Mechanism

For cluster centers obtained from affinity propagation, the reaction mechanism of carbene transfer and its energetics was elucidated with hybrid QM/MM reaction path optimizations and thermodynamics calculations. ChemShell[48,49] was used for QM/MM calculations in combination with DL_POLY[50] for the energy of the molecular mechanics region and TURBOMOLE[51] for the energy of the quantum mechanical region. The QM region included the Fe, carbene, reduced heme, and substrate, while the rest of the protein was in the MM region (SI **Figure S5**). The AmberFF19SB force field generated the protein MM region parametrization. To have a well-refined reaction path, only cluster centers with the benzyl acrylate within 5 Å of the heme Fe were included in the reaction profile calculation. To determine the reaction profile, we used a collective variable that optimally combined three factors: decreasing the distance between CC and C1, increasing the distance between Fe and CC, and reducing the distance between CC and C2. For the QM reaction path optimization, the TPSS DFT functional[52,53] was employed, with def2-TZVP and def2-SVP basis sets for the Fe atom and the remaining atoms in the QM region, respectively. The transition states and products were freely optimized. Vibrational frequency calculations were used to verify the validity of product and transition states and to compute free energies within the harmonic approximation. Single point calculations were done at the reactant, product, and transition states using the TPSSh functional, with the def2-TZVP basis set for all atoms in the QM region to provide more precise electronic energies. All reported free energies are from single point energies with thermodynamic corrections.

References

1. Wolfenden, R. & Snider, M. J. The Depth of Chemical Time and the Power of Enzymes as Catalysts. *Acc. Chem. Res.* **34**, 938–945 (2001).
2. Chan, H. C. S., Pan, L., Li, Y. & Yuan, S. Rationalization of stereoselectivity in enzyme reactions. *WIREs Comput. Mol. Sci.* **9**, e1403 (2019).
3. Mu, R. *et al.* Application of Enzymes in Regioselective and Stereoselective Organic Reactions. *Catalysts* **10**, 832 (2020).
4. *Cytochrome P450: Structure, Mechanism, and Biochemistry*. (Springer International Publishing, Cham, 2015). doi:10.1007/978-3-319-12108-6.
5. Arnold, F. H. Directed Evolution: Bringing New Chemistry to Life. *Angew. Chem. Int. Ed.* **57**, 4143–4148 (2018).
6. Bunzel, H. A., Anderson, J. L. R. & Mulholland, A. J. Designing better enzymes: Insights from directed evolution. *Curr. Opin. Struct. Biol.* **67**, 212–218 (2021).
7. Lovelock, S. L. *et al.* The road to fully programmable protein catalysis. *Nature* **606**, 49–58 (2022).

8. Turner, N. J. Directed evolution drives the next generation of biocatalysts. *Nat. Chem. Biol.* **5**, 567–573 (2009).
9. Huang, P.-S., Boyken, S. E. & Baker, D. The coming of age of de novo protein design. *Nature* **537**, 320–327 (2016).
10. Baker, D. An exciting but challenging road ahead for computational enzyme design. *Protein Sci.* **19**, 1817–1819 (2010).
11. Ward, T. R. Artificial Enzymes Made to Order: Combination of Computational Design and Directed Evolution. *Angew. Chem. Int. Ed.* **47**, 7802–7803 (2008).
12. Reetz, M. T. Directed Evolution of Artificial Metalloenzymes: A Universal Means to Tune the Selectivity of Transition Metal Catalysts? *Acc. Chem. Res.* **52**, 336–344 (2019).
13. Giger, L. *et al.* Evolution of a designed retro-aldolase leads to complete active site remodeling. *Nat. Chem. Biol.* **9**, 494–498 (2013).
14. Althoff, E. A. *et al.* Robust design and optimization of retroaldol enzymes. *Protein Sci.* **21**, 717–726 (2012).
15. Khersonsky, O. *et al.* Bridging the gaps in design methodologies by evolutionary optimization of the stability and proficiency of designed Kemp eliminase KE59. *Proc. Natl. Acad. Sci.* **109**, 10358–10363 (2012).
16. Röthlisberger, D. *et al.* Kemp elimination catalysts by computational enzyme design. *Nature* **453**, 190–195 (2008).
17. Obexer, R. *et al.* Emergence of a catalytic tetrad during evolution of a highly active artificial aldolase. *Nat. Chem.* **9**, 50–56 (2017).
18. Porter, N. J., Danelius, E., Gonen, T. & Arnold, F. H. Biocatalytic Carbene Transfer Using Diazirines. *J. Am. Chem. Soc.* **144**, 8892–8896 (2022).
19. Danelius, E., Porter, N. J., Unge, J., Arnold, F. H. & Gonen, T. MicroED Structure of a Protoglobin Reactive Carbene Intermediate. *J. Am. Chem. Soc.* **145**, 7159–7165 (2023).
20. Fried, S. D. & Boxer, S. G. Electric Fields and Enzyme Catalysis. *Annu. Rev. Biochem.* **86**, 387–415 (2017).
21. Welborn, V. V. & Head-Gordon, T. Fluctuations of Electric Fields in the Active Site of the Enzyme Ketosteroid Isomerase. *J. Am. Chem. Soc.* **141**, 12487–12492 (2019).
22. *Effects of Electric Fields on Structure and Reactivity: New Horizons in Chemistry*. (Royal Society of Chemistry, Cambridge, 2021).
23. Chaturvedi, S. S., Bím, D., Christov, C. Z. & Alexandrova, A. N. From random to rational: improving enzyme design through electric fields, second coordination sphere interactions, and conformational dynamics. *Chem. Sci.* **14**, 10997–11011 (2023).
24. Hanoian, P., Liu, C. T., Hammes-Schiffer, S. & Benkovic, S. Perspectives on Electrostatics and Conformational Motions in Enzyme Catalysis. *Acc. Chem. Res.* **48**, 482–489 (2015).
25. Warshel, A. *et al.* Electrostatic Basis for Enzyme Catalysis. *Chem. Rev.* **106**, 3210–3235 (2006).
26. Bím, D. & Alexandrova, A. N. Local Electric Fields As a Natural Switch of Heme-Iron Protein Reactivity. *ACS Catal.* **11**, 6534–6546 (2021).
27. Labas, A., Szabo, E., Mones, L. & Fuxreiter, M. Optimization of reorganization energy drives evolution of the designed Kemp eliminase KE07. *Biochim. Biophys. Acta BBA - Proteins Proteomics* **1834**, 908–917 (2013).
28. Hennefarth, M. R. & Alexandrova, A. N. Direct Look at the Electric Field in Ketosteroid Isomerase and Its Variants. *ACS Catal.* **10**, 9915–9924 (2020).

29. Khade, R. L. *et al.* Iron Porphyrin Carbenes as Catalytic Intermediates: Structures, Mössbauer and NMR Spectroscopic Properties, and Bonding. *Angew. Chem. Int. Ed.* **53**, 7574–7578 (2014).
30. Khade, R. L. & Zhang, Y. C−H Insertions by Iron Porphyrin Carbene: Basic Mechanism and Origin of Substrate Selectivity. *Chem. – Eur. J.* **23**, 17654–17658 (2017).
31. De Brito Sá, E., Rimola, A., Rodríguez-Santiago, L., Sodupe, M. & Solans-Monfort, X. Reactivity of Metal Carbenes with Olefins: Theoretical Insights on the Carbene Electronic Structure and Cyclopropanation Reaction Mechanism. *J. Phys. Chem. A* **122**, 1702–1712 (2018).
32. Wei, Y., Tinoco, A., Steck, V., Fasan, R. & Zhang, Y. Cyclopropanations via Heme Carbenes: Basic Mechanism and Effects of Carbene Substituent, Protein Axial Ligand, and Porphyrin Substitution. *J. Am. Chem. Soc.* **140**, 1649–1662 (2018).
33. Rogge, T., Zhou, Q., Porter, N. J., Arnold, F. H. & Houk, K. N. Iron Heme Enzyme-Catalyzed Cyclopropanations with Diazirines as Carbene Precursors: Computational Explorations of Diazirine Activation and Cyclopropanation Mechanism. *J. Am. Chem. Soc.* **146**, 2959–2966 (2024).
34. Bishop, C. M. *Pattern Recognition and Machine Learning*. (Springer, New York, 2006).
35. Jumper, J. *et al.* Highly accurate protein structure prediction with AlphaFold. *Nature* **596**, 583–589 (2021).
36. Eberhardt, J., Santos-Martins, D., Tillack, A. F. & Forli, S. AutoDock Vina 1.2.0: New Docking Methods, Expanded Force Field, and Python Bindings. *J. Chem. Inf. Model.* **61**, 3891–3898 (2021).
37. Trott, O. & Olson, A. J. AutoDock Vina: Improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. *J. Comput. Chem.* **31**, 455–461 (2010).
38. Case, D. A. *et al.* AmberTools. *J. Chem. Inf. Model.* **63**, 6183–6191 (2023).
39. Case, D. A. *et al. Amber 2022*. (University of California, San Francisco, 2022).
40. Li, P. & Merz, K. M. MCPB.py: A Python Based Metal Center Parameter Builder. *J. Chem. Inf. Model.* **56**, 599–604 (2016).
41. Tian, C. *et al.* ff19SB: Amino-Acid-Specific Protein Backbone Parameters Trained against Quantum Mechanics Energy Surfaces in Solution. *J. Chem. Theory Comput.* **16**, 528–552 (2020).
42. Ryckaert, J.-P., Ciccotti, G. & Berendsen, H. J. C. Numerical integration of the cartesian equations of motion of a system with constraints: molecular dynamics of n-alkanes. *J. Comput. Phys.* **23**, 327–341 (1977).
43. Knapp, B., Ospina, L. & Deane, C. M. Avoiding False Positive Conclusions in Molecular Simulation: The Importance of Replicas. *J. Chem. Theory Comput.* **14**, 6127–6138 (2018).
44. Roe, D. R. & Cheatham, T. E. PTRAJ and CPPTRAJ: Software for Processing and Analysis of Molecular Dynamics Trajectory Data. *J. Chem. Theory Comput.* **9**, 3084–3095 (2013).
45. Miller, B. R. *et al. MMPBSA.py* : An Efficient Program for End-State Free Energy Calculations. *J. Chem. Theory Comput.* **8**, 3314–3321 (2012).
46. Raček, T. *et al.* Atomic Charge Calculator II: web-based tool for the calculation of partial atomic charges. *Nucleic Acids Res.* **48**, W591–W596 (2020).
47. Frey, B. J. & Dueck, D. Clustering by Passing Messages Between Data Points. *Science* **315**, 972–976 (2007).

48. Sherwood, P. *et al.* QUASI: A general purpose implementation of the QM/MM approach and its application to problems in catalysis. *J. Mol. Struct. THEOCHEM* **632**, 1–28 (2003).
49. Kästner, J. *et al.* DL-FIND: An Open-Source Geometry Optimizer for Atomistic Simulations. *J. Phys. Chem. A* **113**, 11856–11865 (2009).
50. Todorov, I. T., Smith, W., Trachenko, K. & Dove, M. T. DL_POLY_3: new dimensions in molecular dynamics simulations via massive parallelism. *J. Mater. Chem.* **16**, 1911 (2006).
51. Ahlrichs, R., Bär, M., Häser, M., Horn, H. & Kölmel, C. Electronic structure calculations on workstation computers: The program system turbomole. *Chem. Phys. Lett.* **162**, 165–169 (1989).
52. Perdew, J. P., Tao, J., Staroverov, V. N. & Scuseria, G. E. Meta-generalized gradient approximation: Explanation of a realistic nonempirical density functional. *J. Chem. Phys.* **120**, 6898–6911 (2004).
53. Tao, J., Perdew, J. P., Staroverov, V. N. & Scuseria, G. E. Climbing the Density Functional Ladder: Nonempirical Meta–Generalized Gradient Approximation Designed for Molecules and Solids. *Phys. Rev. Lett.* **91**, 146401 (2003).