Convergence in Determining Enzyme Functional Descriptors across Kemp Eliminase Variants

Yaoyukun Jiang,[1] Sebastian L. Stull,[1] Qianzhen Shao,[1] and Zhongyue J. Yang[1-5,*]

[1]*Department of Chemistry, Vanderbilt University, Nashville, Tennessee 37235, United States*

[2]*Center for Structural Biology, Vanderbilt University, Nashville, Tennessee 37235, United States*

[3]*Vanderbilt Institute of Chemical Biology, Vanderbilt University, Nashville, Tennessee 37235, United States* [4]*Data Science Institute, Vanderbilt University, Nashville, Tennessee 37235, United States* [5]*Department of Chemical and Biomolecular Engineering, Vanderbilt University, Nashville, Tennessee 37235, United States*

ABSTRACT: Molecular simulations have been extensively employed to accelerate biocatalytic discoveries. Enzyme functional descriptors derived from molecular simulations have been leveraged to guide the search for beneficial enzyme mutants. However, the ideal active-site region size for computing the descriptors over multiple enzyme variants remains untested. Here, we conducted convergence tests for dynamics-derived and electrostatic descriptors on eighteen Kemp eliminase variants across six active-site regions with various boundary distances to the substrate. The tested descriptors include the root-mean-square deviation of the active-site region, the solvent accessible surface area ratio between the substrate and active site, and the projection of the electric field on the breaking C–H bond. All descriptors were evaluated using molecular mechanics methods. To understand the effects of electronic structure, the electric field was also evaluated using quantum mechanics/molecular mechanics methods. The descriptor values were computed for eighteen KE variants combined with six active-site regions. Spearman correlation matrices were used to determine the region size condition under which further expansion of the region boundary does not substantially change the ranking of descriptor values. We observed that protein dynamics-derived descriptors, including $RMSD_{active\_site}$ and $SASA_{ratio}$, converge at a distance cutoff of 5 Å from the substrate. The electrostatic descriptor, $EF_{C–H}$, converges at 6 Å using molecular mechanics methods and 7 Å using quantum mechanics/molecular mechanics methods. This study serves as a future reference to determine descriptors for predictive modeling of enzyme engineering.

Keywords: Convergence Test, Enzyme Kinetics, Mutation Effect, Kemp Eliminase

1. Introduction

Enzymes have been widely used as biocatalysts for chemical synthesis,[1-3] biomass conversion,[4-7] polymer upcycling,[8-11] drug functionalization,[12-15] and food allergy treatment.[16-18] Wild-type enzymes usually exhibit low specificity for converting non-native substrate and feeble activity for catalyzing new-to-nature reactions. Experimental strategies of enzyme engineering, such as random mutagenesis,[19-21] gene shuffling/recombination,[22,23] CASTing,[24,25] and directed evolution,[26-29] have been leveraged to optimize enzymes' capability for accommodating non-native substrates or catalyzing new-to-nature reactions. These strategies require extensive efforts for screening and selecting mutants to achieve desired functions. To accelerate biocatalytic discovery, molecular simulations[30-34] have been augmented with the campaign of biocatalytic discovery. The catalytic actions of enzyme catalysis can be elucidated and quantified using descriptors, including folding stability,[35] binding affinity,[36-38] activation barriers,[39,40] protein dynamics and correlated motions,[41-49] electric field (EF),[50-55] charge transfer,[54,56] and more. These descriptors, derived from quantum mechanical (QM) or molecular mechanical (MM) simulations, have guided the search for beneficial mutants.[57,58] They also serve as critical features for data-driven enzyme engineering.

For example, protein dynamics-derived descriptors and electric fields have been extensively studied, because they were found to correlate with enzyme catalytic efficiency.[50,58-60] Additionally, their computation is more efficient than that of activation barriers, whose convergence requires intensive conformational sampling and electronic structure calculations. A common descriptor for protein dynamics is the root-mean-square deviation of the active-site region ($RMSD_{active\_site}$). $RMSD_{active\_site}$ quantifies the structural fluctuation of protein backbones or sidechains relative to a reference structure. The fluctuation is associated with the B-factor of protein structure determined from crystallography. In an analysis of catalytic residues in 178

enzyme active sites,[59] Bartlett et. al. showed that the active-site residues of efficient enzymes generally have a lower B-factor. As such, a lower $RMSD_{active\_site}$ should be expected for efficient mutant enzymes and designer enzymes in catalysis, albeit that the catalytic efficiency may drop under very low RMSD range.[61] Besides $RMSD_{active\_site}$, our group identified a new descriptor to evaluate the overall impact of protein dynamics on substrates.[58] The descriptor, defined as solvent accessible surface area ratio of substrate to active-site residues ($SASA_{ratio}$), can be obtained from molecular dynamics simulations. Using lactonase as a model system, our previous work shows that $SASA_{ratio}$ can guide the search of optimal enzyme mutants with enhanced specificity for non-native substrates. Besides protein dynamics, the role of electrostatic environments was reported as a critical factor in mediating enzyme catalysis.[62] Linear correlation was observed between the magnitude of EF in the reaction center and the free energy barrier in ketosteroid isomerase and serine protease.[50]

Despite the broad applications of simulation-derived descriptors in guiding enzyme engineering, converging the computation of descriptors in QM and MM simulations is a non-trivial task. Failure of achieving convergence hampers reproducibility of computational outcomes and may misguide experimental designs. This issue is particularly significant for QM-based calculations due to their high computational cost. Benchmarks have been performed to investigate the selection of QM regions that converge the computation of electronic structure descriptors (e.g., partial charge,[63-67] charge transfer,[67] charge density,[66] bond valence,[67] and electrostatic potential), energetic properties (e.g., energy barrier, reaction energy, and free energy),[64,65,67-75] geometries,[64,73,74] and NMR shielding[76,77] in various model enzymes (peroxidase, methyltransferases, cytochrome P450, and deacetylase).[67] Rational QM region selection

approaches have also been developed, including charge shift analysis,[78] Fukui shift analysis,[78] and point charge variation analysis[79].

The benchmark studies on descriptors have been mostly performed on wild-type enzymes.[72] However, to understand or predict mutation effects, it is essential to perform convergence tests over multiple enzyme variants. Ideally, the selected active-site region for computing QM or MM properties should be large enough so that further expanding the region size does not substantially change the order of descriptor values across different enzyme variants. In this work, using 18 variants of Kemp eliminase,[80] we investigated whether the ranking of descriptor values across enzyme variants approaches convergence as the increase of active-site region sizes used in descriptor computation. We first sampled conformational ensembles for 18 variants using classical molecular dynamics. Based on the sampled conformers, protein dynamics-derived descriptors (i.e., $RMSD_{active\_site}$ and $SASA_{ratio}$) and electronic structure-derived descriptors (i.e., electric field along the breaking C–H bond) were evaluated using different sizes of active-site region based on MM or QM methods. For each descriptor, the Spearman correlation matrix was computed to examine the trend of convergence. The study informs the conditions under which different descriptors can be calculated with high fidelity for predicting the impact of mutations on catalytic functions.

2. Computational Methods

Protein Structure and Preparation The crystal structure of KE07-R7-2 was obtained from the Protein Data Bank (PDB ID: 5D38).[56] All the crystallizing water molecules were removed. To make the amino acid sequence consistent with the original KE07 design,[30] the N-terminal alanine was changed to methionine and the residues following Leu253 on the C-terminal were removed.

The crystal structure[30] of KE07 in complex with the substrate 5-nitrobenzisoxazole was aligned relative to the KE07-R7-2 crystal structure using PyMol.[81] The coordinates of the substrate were used to construct the KE07-R7-2-substrate complex. The complex was then prepared with the AMBER 18 *tleap*[82] utility for MD simulations. AMBER ff14SB force field was used for the protein.[83] Parameters for the substrate were obtained using the generalized AMBER force field.[84,85] The atomic charges were determined by the AM1-BCC model.[86] The missing atoms were also complemented with *tleap*.

Molecular Dynamics Simulations MD simulations for each of the 18 variant-substrate complexes were conducted with a high throughput enzyme modeling platform, EnzyHTP.[87] The 18 variants include one KE07-R7-2 as the "wild-type" and 17 of its mutants, including S48N, H201A, H201K, K222A, R16Q, N25S, I52A, M62A, H84Y, K132N, I199S, I199F, I199A, K132M, K162A, L170A, E185A (Supporting Information, Table S1 and .zip file). Specifically, EnzyHTP automatically generates the structures of enzyme mutants based on the original structure and performs MD simulations using AMBER 18.[82] The SHAKE algorithm was applied to constrain all the hydrogen-containing bonds.[88] To sample the near transition state conformations throughout the simulations, geometric restraints between the substrate and key amino acid residues were applied from minimization to production runs (Supporting Information, Figure S1). The enzyme complexes were then solvated in a periodic octahedron box with a 10 Å buffer of TIP3P water and were neutralized with Na^+ counterions. For each variant complex, the whole solvent box was first relaxed using steepest descent method for 10000 steps followed by conjugate gradient method for another 10000 steps. After minimization, each box was heated from 0 to 293.15 K within 36 ps with constant volume, equilibrated for 4 ps under constant volume at 293.15 K, and further equilibrated at 293.15 K and 1 atm for 1 ns. In addition to the geometric restraints

mentioned above, the backbone C_α, C and N of the amide group were also restrained with a 2 kcal·mol^{-1}·Å^{-2} weight from the minimization to equilibration. After equilibration, we carried out production runs for 110 ns and output the trajectories every 100 ps. The snapshots derived from the last 100 ns of the production run were used for analyses. This yields a total of 1000 snapshots for each production run. All simulations were performed with a time step of 2 fs. The Langevin thermostat[89] and Berendsen barostat[90] were used throughout the simulations. For each of the 18 variant-substrate complexes, five parallel MD runs were conducted with different random seeds, yielding a total sampling time of 500 ns and 5000 snapshots.

QM/MM Calculations We conducted QM/MM single-point electronic structure calculations for 500 snapshots sampled from MD production runs with a 1 ns interval. QM/MM single-point energies were calculated using TeraChem.[91,92] The electrostatic interactions between the QM and MM region were treated with the electrostatic embedding method.[62] The QM/MM boundaries cut the backbone C-N bond of the amide group. To cap the unbonded atoms in the QM region, explicit H atoms were placed along the bond vector connecting the QM and MM atoms, and the resulting N–H and C–H bond lengths were set to be 1.09 Å. At the same time, the point charges originally belonging to the QM-region-bonded amide C and N atoms in the MM region were removed, and their charges were redistributed evenly on the remaining MM atoms except for those covalently bonded to the deleted MM amide C and N atoms. The electronic structures were described using the range-separated exchange-correlation functional ωPBEh[93] (ω = 0.2 bohr^{-1}) with 6-31G(d).[94] This combination of method and basis set has been validated in the study of large-scale electronic structure effects in catechol *O*-methyltransferase, cytochrome P450cam, lysozyme, and DNA methyltransferase.[65,67] The restrained electrostatic potential (RESP) point charges[95] of each snapshot were calculated for QM residue electric field analyses.

Descriptor Calculations and Analyses We selected six active-site regions whose boundary's distance to the substrate surface ranges from 3 to 8 Å with a 1 Å interval. For all 18 variants, the active-site regions were classified based on the averaged MD structure of KE07-R7-2. A residue is selected in the region if any one of its heavy atoms is within the distance cutoff from its nearest substrate heavy atom. Based on each of the active-site region, we calculated the enzyme functional descriptors, including mass weighted root-mean-square deviation of an active-site region ($RMSD_{active\_site}$, in Å), solvent accessible surface area ratio between substrate and active-site residues ($SASA_{ratio}$, in $Å^2$), and electric field along the breaking C–H bond ($EF_{C–H}$, in $MV \cdot cm^{-1}$). The values of each descriptor were first evaluated on individual conformational snapshots, and then averaged over sampled classical MD or QM/MM snapshots (Supporting Information, Table S3-S6).

For $RMSD_{active\_site}$, we included all the heavy atoms of the amino acid residues. The reference structure was averaged from sampled MD snapshots. The $SASA_{ratio}$ was calculated based on the ratio of $SASA_{sub}$ (substrate's SASA) to $SASA_{protein}$ (protein residues' SASA). SASA was quantified using the Shrake and Rupley algorithm[96] embedded in the python library MDTraj.[97] The probe radius was 1.4 Å and the surface of each atom was represented by 5000 grid points. $EF_{C–H}$ was calculated to be the projected EF strength at the middle point of the breaking C–H bond of the substrate 5-nitrobenzisoxazole. The bond vector direction points from C to H. We separately computed $EF_{C–H}$ based on RESP charges either derived from molecular mechanics force field or single-point electronic structure calculation. For MM-derived $EF_{C–H}$, the $EF_{C–H}$ was summed over from all atoms in the selected active-site region based on the RESP charges used in the classical force field. For QM/MM-derived $EF_{C–H}$, the $EF_{C–H}$ was summed over from all atoms in the QM and MM region. The EF contributions from the capping H atoms were not included.

For each active-site region, the averaged descriptor values were computed then ranked across the 18 enzyme variants. Spearman correlation matrix for each descriptor was computed that contains the correlation coefficients for each pair of the active-site regions.

3. Results and Discussion

3.1 Kemp Eliminase Variants as the Model System

As the first known *de novo*-designed enzyme, Kemp eliminase catalyzes the conversion of benzisoxazole to cyanophenol via C–H deprotonation followed by ring opening (Figure 1 top right).[30] Three generations of Kemp eliminase have been reported,[30,42,51,52,56,80,98-100] including the KE family designed using the "inside-out" protocol by Baker, Houk, Tawfik, and co-workers;[30] the HG family using iterative protocol by Hilvert, Houk, Mayo and co-workers;[31] and the AlleyCat family using the minimalist approach by Korendovych, Degrado, and coworkers.[32,100] From their initial reports, the most efficient enzyme variants were identified to be KE07-R7-2 (k_{cat}/K_M = 2590 $M^{-1}s^{-1}$), HG-3 (k_{cat}/K_M = 430 M^{-1} s^{-1}), and AlleyCat (i.e., k_{cat}/K_M = 128.4 $M^{-1}s^{-1}$). All three families of Kemp eliminase involve a general acid-base mechanism, in which the substrate 5-nitrobenzisoxazole is deprotonated by a nearby carboxylate (side chain of Glu or Asp) to form 2-hydroxy-5-nitrobenzonitrile via one single transition state (Figure 1, *top-right*). Nonetheless, they involve a different set of active site residues for substrate deprotonation and binding.

We chose the model system to be a member of the KE family, KE07-R7-2,[30,56] and seventeen of its variants with single amino acid substitution reported by Head-Gordon and coworkers.[80] KE07-R7-2 was derived from seven rounds of directed evolution based on a computationally designed enzyme scaffold KE07. In KE07-R7-2 and its variants, the carboxylic sidechain of Glu101 serves as the catalytic base (Figure 1, *bottom-right*). These variants were

selected in the benchmark for three reasons. First, the mutational spots of the variants span over a wide range of spatial proximity to the substrate (i.e., 3 - 23 Å, Figure 1, *left* and Supporting Information, Table S1). Both close and distal mutations are thus considered in the study. Second, although modeling has been performed for KE07-R7-2 to infer mutational hotspots based on correlated residue motion,[80] protein dynamics and electronic structures for the 17 variants of KE07-R7-2 have not been investigated. Third, the kinetic parameters (i.e., k_{cat} or K_M) for these variants are known experimentally.[80] This implies that the mutation does not abolish the structural and catalytic integrity of Kemp eliminase. The crystal structure for KE07-R7-2 can be used as a scaffold for mimicking the mutant structures.

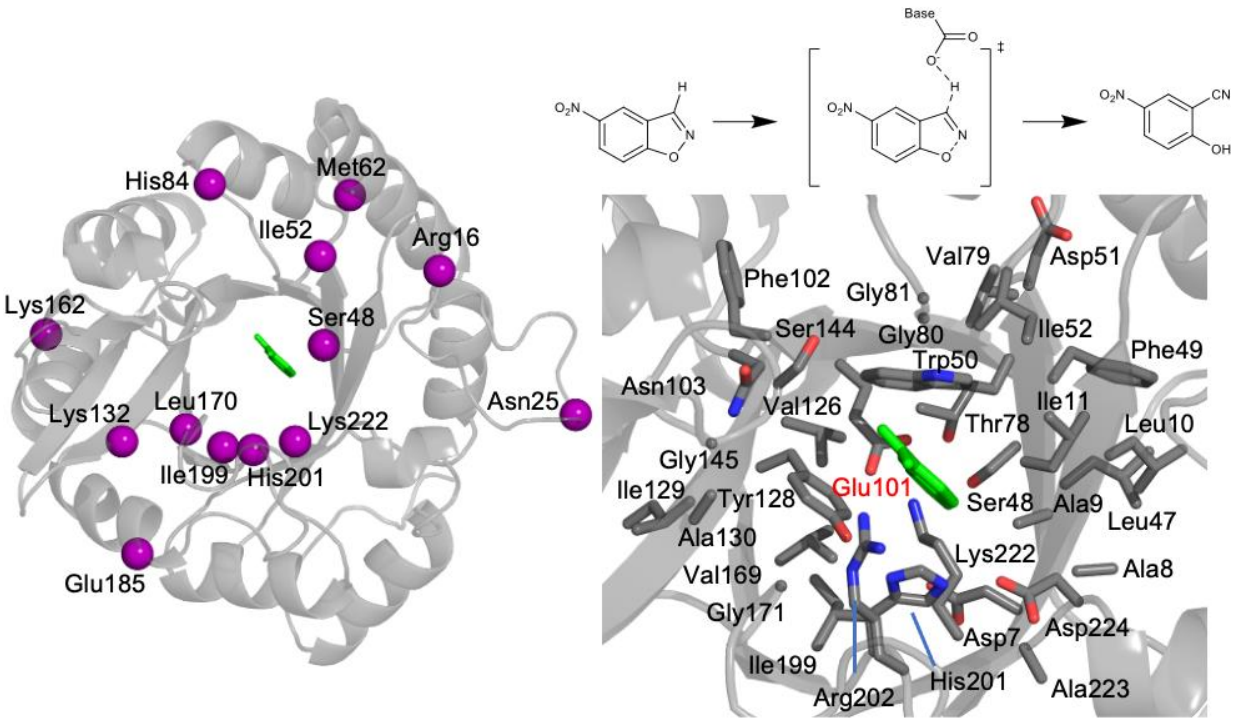

Figure 1. Mutation spots, catalyzed reaction, and active site residues of Kemp eliminase KE07-R7-2. (*Left*) Spatial distribution of mutation spots. The C_α atom of each site is shown in purple sphere, and the substrate is shown in green sticks. (*Top-right*) Catalyzed Kemp elimination

reaction. The single transition state involves the deprotonation of a carbon atom. The transition state is stabilized by a general base from an amino acid side chain. The partial negative charge on the oxygen atom is stabilized by a hydrogen bond donor, which can be an amino acid side chain or a solvent water molecule. (*Bottom-right*) Active site residues are shown in stick. The catalytic base is labeled in red, the substrate is shown in green, and the rest of the residues are shown in gray.

For KE07-R7-2 and its variants, the greater enzyme active site region entails 32 residues, including 6 polar, 20 non-polar, and 6 charged residues (Figure 1, *bottom-right* and Supporting Information, Table S2). By design, Glu101, Lys222 and Trp50 directly participate in the reaction or stabilize the transition state.[30] Glu101 is the general base that deprotonates the substrate. Lys222 is the H-bond donor to stabilize the phenoxide intermediate. Trp50 is the π-stacking residue to stabilize the substrate binding and charge-separated transition state. Four polar residues are observed within 5 Å of the substrate, including Tyr128, Ser48, His201, and Arg202. They likely stabilize the substrate binding or transition state with electrostatic or polar interactions. In addition, a total of eight polar (i.e., Glu101, Lys222, Tyr128, Ser48, His201, Arg202, Asp224, Asn103) and four charged (i.e., Glu101, Lys222, Arg202, Asp224) residues are found within 5.5 Å from the substrate. These residues mediate the electric field environment exerted on the breaking C–H bond. Besides dispersion interactions, the nonpolar residues likely contribute to the active site dynamics as described by $RMSD_{active\_site}$ and $SASA_{ratio}$.

3.2 Region Selection for Descriptor Calculation

To calculate simulation-derived descriptors, an active-site region should be defined first. In this study, the calculations of $RMSD_{active\_site}$, $SASA_{ratio}$, and MM-derived $EF_{C–H}$ involve only

the residues classified within a defined active-site region. The calculation of QM/MM-derived $EF_{C–H}$ involves treatment of the active-site region residues using quantum mechanics and the rest of the enzyme residues using molecular mechanics. To benchmark the region size effect, the active-site regions were defined based on the residues' spatial proximity to the substrate (see Computational Method, *Descriptor Calculations and Analyses*). We selected six active-site regions whose boundaries to the substrate range from 3 to 8 Å (with 1 Å interval) – they are named C3 to C8, respectively (Figure 2). The residues were consistently selected by referencing KE07-R7-2. For C3, only Glu101 is included. Glu101 serves as the catalytic base to deprotonate the residue. Notably, throughout the MD simulations, a distance constraint was applied between Glu101 and the substrate to maintain their favorable catalytic pose. Compared to C3, C4 involves an expansion of 7 additional residues. Among them, Lys222 and Trp50 appear in the original design of theozyme.[30] These two residues, cooperating with His201, Tyr128, and Ser48, likely facilitate proton transfer needed for the general acid-base mechanism. Unlike C3 which bears a -1 charge, C4 is charge neutral due to the addition of Lys222. In C5, only one additional residue Arg202 is included. This indicates that the catalytic core of KE involves a relatively compact inner cluster of residues surrounding the substrate. The positive charge introduced by Arg202 in C5 is neutralized by Asp224 in the C6 region. Notably, among the newly added residues in C6, three (i.e., Leu10, Phe49, and Val169) out of five are non-polar. This trend is also observed in C7 and C8. For the new additions, only two residues (i.e., Ser144 and Thr78) out of eight are polar in C7; two (i.e., Asp7 and Asp51) out of ten are polar in C8. The excessive number of non-polar residues in the greater active-site region contribute to the stability of Kemp eliminase.

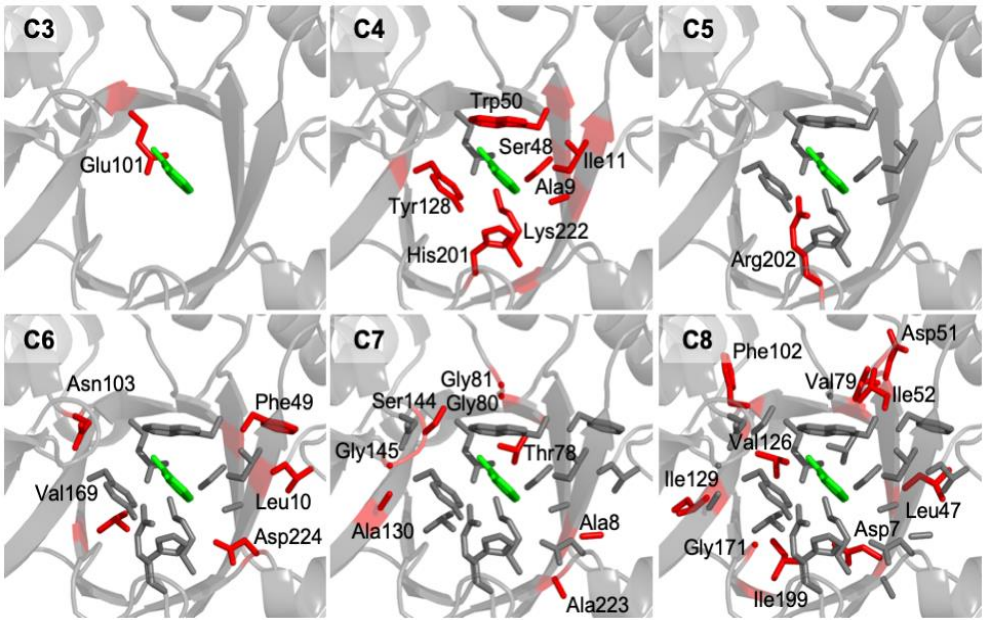

Figure 2. Six active-site regions with various boundary distances to the substrate. The distances used here range from 3 to 8 Å; the regions are named C3 to C8, respectively. For each region, the selected residues are shown in stick. Compared to an adjacent region with a smaller size, the newly-added residues shown in red stick and labeled with a residue name; the existing residues are shown in gray stick.

Table 1. Residue number, atom number, and net charge for the six active-site regions of KE07-R7-2 with different region sizes. Substrate atoms are counted in the atom number.

Cutoff (Å)	Residue number	Atom number	Charge
3	1	31	-1
4	8	155	0
5	9	179	1
6	14	260	0
7	22	336	0
8	32	495	-2

The total number of atoms ranges from 31 in C3 to 495 in C8 (Table 1). The region size tested here is comparable to or greater than the optimal region sizes determined from previous benchmark studies, including DNA methyltransferase by Mehmood et. al.[67] (300 atoms), histone deacetylase by Morgenstern et al.[66] (200 atoms), and catechol O-methyltransferase by Kulik et al.[64] (500 to 600 atoms) and Jindal et al.[72] (60 atoms).

For each of the six active-site regions, we computed the average descriptor values for the eighteen KE07-R7-2 variants based on their conformational ensembles (Supporting Information, Tables S3 to 6). We investigated how the ranking of descriptor values across the eighteen variants varies with the increase of region size. Instead of benchmarking a certain molecular property against its reference value, this study intends to identify a condition of region size under which further expanding the region boundary minimally changes the ranking of descriptor values across enzyme variants. Notably, the region size condition for a converged trend does not guarantee the convergence of individual property values. Nonetheless, the mutation effect can be reasonably inferred under this condition to guide enzyme engineering.

3.3 Descriptor of Protein Dynamics: $RMSD_{active\_site}$ and $SASA_{ratio}$

We first investigated the dynamics-derived descriptors, $RMSD_{active\_site}$ and $SASA_{ratio}$. They represent different aspects of protein dynamics. $RMSD_{active\_site}$ informs the conformational fluctuation of active site residues, while $SASA_{ratio}$ informs the dynamic positioning and fitness of substrate in the active site.

Figure 3 shows the Spearman correlation matrix for $RMSD_{active\_site}$ (*left*) and $SASA_{ratio}$ (*right*). Each element of the matrix represents a Spearman correlation coefficient (i.e., ρ) between descriptor values derived from two regions with a distinct size. For $RMSD_{active\_site}$, a high ρ value

(i.e., ≥ 0.70) is observed for almost all pairs of regions except those that involve C3. The moderate ρ values between C3 and C5 – C8 (i.e., 0.5-0.7) are caused by the small size of C3 that involves only one residue in the region (i.e., Glu101). The correlation coefficients tend to be higher for regions that are close in size (e.g., ρ > 0.9 for C5-C6, C6-C7, and C7-C8) and lower for regions with a larger size gap (e.g., R = 0.53 for C3-C8 and 0.73 for C4-C8). Notably, it is unexpected that the correlation coefficient is still as high as 0.53 between C3 and C8 because their numbers of residues differ by 31 and of atoms by 464. This indicates that the $RMSD_{active\_site}$ ranking calculated from C3 can still partially inform the ranking of dynamic fluctuation exhibited by larger-sized regions. This is likely caused by the collective motions of residues in the enzyme active site, where all residues are somewhat interconnected in a complex, dynamic network.

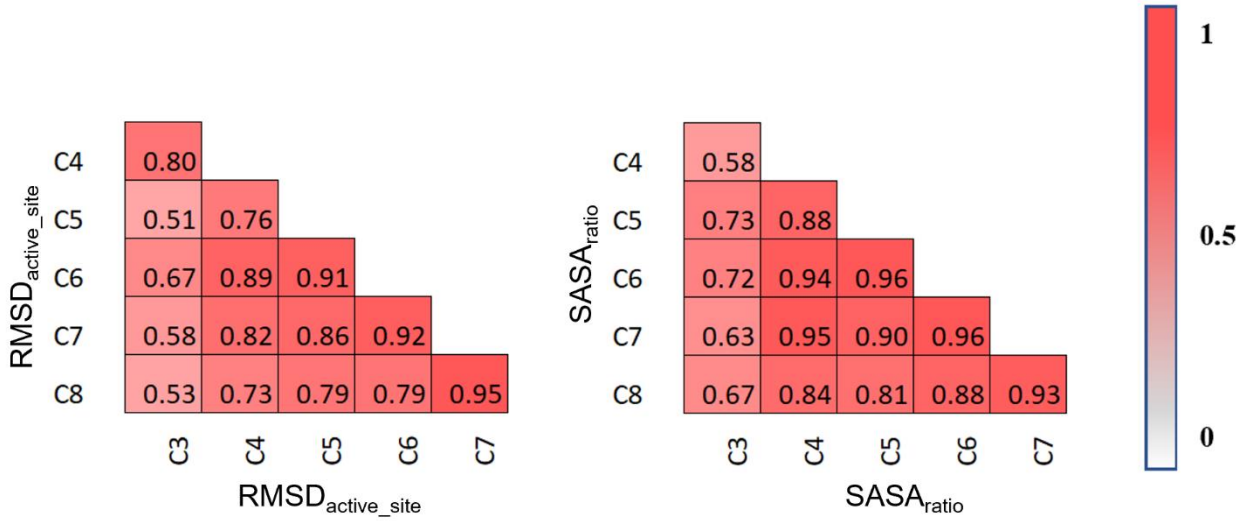

Figure 3. Spearman correlation matrices for protein dynamics-derived descriptors, $RMSD_{active\_site}$ (*left*) and $SASA_{ratio}$ (*right*). Each matrix element represents a Spearman correlation coefficient for a pair of active-site regions with a distinct size. The magnitude of the correlation is coded by a gradient color bar that ranges from 0 (white) to 1 (red).

Unlike $RMSD_{active\_site}$, which emphasizes protein dynamics, $SASA_{ratio}$ represents the interplay of dynamic motion between substrate and its surrounding active-site residues. The Spearman correlation matrix of $SASA_{ratio}$ shows a similar trend to that of $RMSD_{active\_site}$ (Figure 3, *right*). For each pair of regions, the Spearman correlation coefficient of $SASA_{ratio}$ is generally greater than that of $RMSD_{active\_site}$. The ρ values are greater for correlations between larger regions that are closer in size (e.g., ρ = 0.96, 0.96, and 0.93 for C5-C6, C6-C7, and C7-C8, respectively). Notably, the $SASA_{ratio}$ is computed by the SASA ratio of substrate to active-site residues. For different active-site regions, the SASA value of the substrate always remains constant. This helps dampen the perturbation of expanding region size on the ranking of descriptor values across variants.

To determine a convergence cutoff for computing dynamics-derived descriptors, we investigated the change of Spearman correlation coefficients between adjacent active-site regions versus the increase of region size (Figure 4). For both $RMSD_{active\_site}$ and $SASA_{ratio}$, the correlation coefficient appears greater than 0.90 after C5 (i.e., 5.0 Å from the substrate). With a Spearman ρ value greater than 0.90, the ranking of descriptor values computed from one active-site region is largely preserved in another. As such, the convergence cutoff for dynamics-derived descriptors is determined to be 5.0 Å. Notably, from C5 to C8, the atomic charge varies from 1 (C5), to 0 (C6 and C7), then to -2 (i.e., C8). The correlation coefficients remain high even between regions of different charges. This observation confirms that the dynamics-derived descriptors used here are approximately independent from electrostatic effects – they are insensitive to electrostatic perturbation in the protein environment.

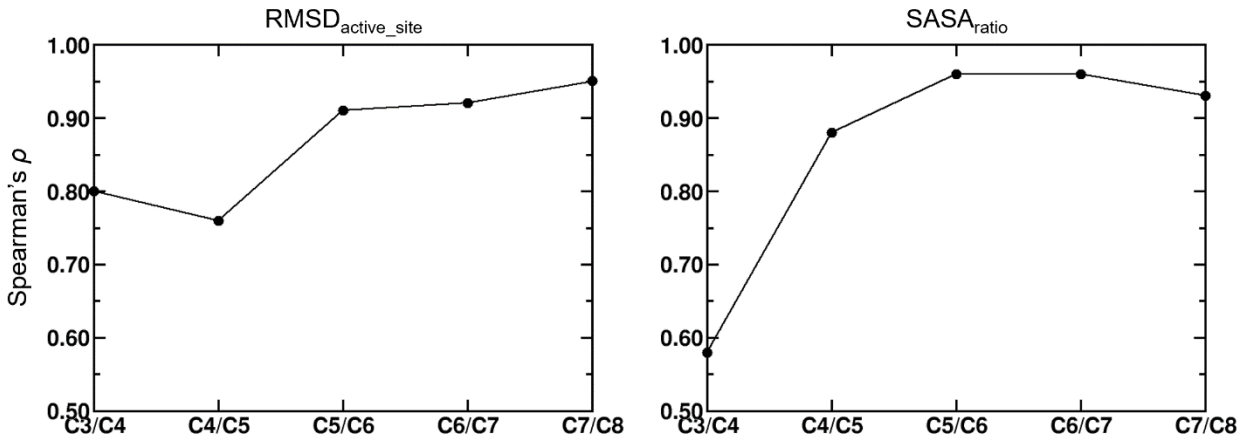

Figure 4. Spearman correlation coefficients for the dynamics-derived descriptors, $RMSD_{active\_site}$ (*left*) and $SASA_{ratio}$ (*right*) between regions that are close in size.

3.4 Descriptor of Electrostatic Environment: Electric Field along Breaking C–H Bond

Next, we investigated the descriptor for enzyme electrostatics, $EF_{C–H}$, the electric field along the breaking C–H bond. The interior electric field in Kemp eliminase has been proposed as a factor to stabilize the developing dipole moment along the C–H bond.[52] Optimizing the electric field through mutagenesis has also been demonstrated as an effective strategy to improve enzyme catalytic efficiency.[52,101]

Figure 5 shows the Spearman correlation matrix for $EF_{C–H}$ that were separately computed using MM (*left*) and QM/MM (*right*) method. MM-derived $EF_{C–H}$ involves only the local residues that are classified in the active-site region. This approach is similar to the distance cutoff method used in Rosetta score functions for computing electrostatic interactions.[102] QM/MM-derived $EF_{C–H}$ employs QM to treat residues in the active-site region and MM for residues in the rest of the enzyme. This approach incorporates the effects of long-range electrostatics.

Unlike dynamics-derived descriptors, low correlation coefficients are more frequently observed between active-site regions of different sizes, especially between regions with a larger size gap. For example, the Spearman ρ values for C3-C6 (i.e., differ by 13 residues), C3-C7 (i.e., differ by 21 residues), and C3-C8 (i.e., differ by 31 residues) are 0.07, 0.07, and 0.05, respectively, for MM-derived $EF_{C–H}$ (Figure 5, *left*); they are 0.19, 0.28, and 0.25, respectively, for QM/MM-derived $EF_{C–H}$ (Figure 5, *right*). The low correlation strength indicates that the ranking of $EF_{C–H}$ values derived from a smaller active-site region cannot be used to infer the ranking from a larger active-site region. Different from dynamics-derived descriptors, the electric field depends more sensitively on the active-site regions used in the calculation. From C3 to larger active-site regions, individual residues added to the active site region, especially polar and charged residues, can significantly affect the representation of mutation effects on interior enzyme electrostatics.

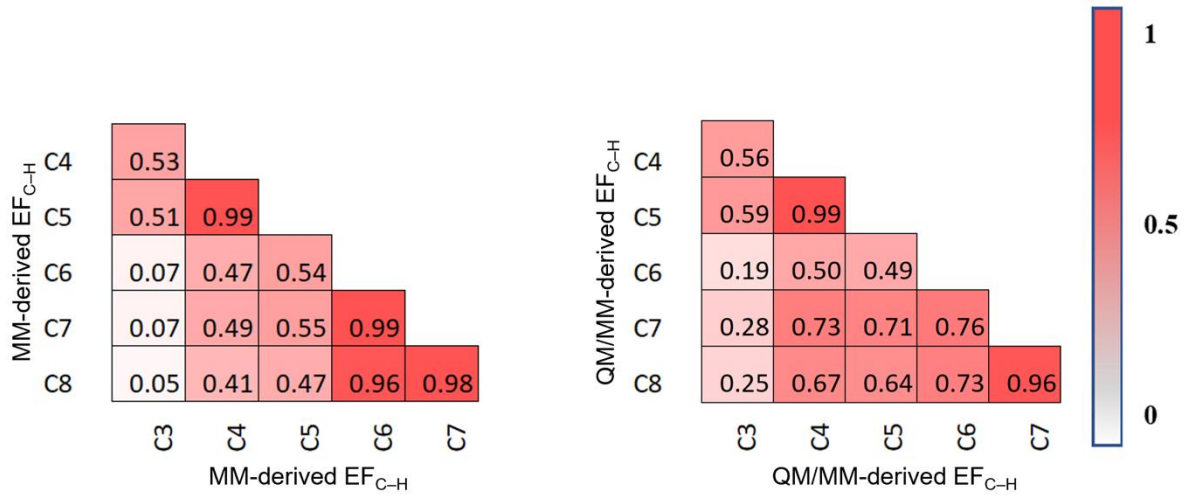

Figure 5. Spearman correlation matrix for MM-derived $EF_{C–H}$ (*left*) and QM/MM-derived $EF_{C–H}$ (*right*). Each matrix element represents a Spearman correlation coefficient for a pair of active-site regions with different region sizes. The magnitude of the correlation is coded by a gradient color bar that ranges from 0 (white) to 1 (red).

Similar to dynamics-derived descriptors, the Spearman ρ values are greater for correlations of $EF_{C–H}$ rankings between larger regions that are closer in size. The ρ values for C4-C5, C6-C7, and C7-C8 are 0.99, 0.99, and 0.98, respectively, for MM-derived $EF_{C–H}$ (Figure 5, *left*); and are 0.99, 0.76, and 0.96, respectively, for QM/MM-derived $EF_{C–H}$ (Figure 5, *right*). Interestingly, the rankings derived from C4 and C5 are highly consistent, albeit their difference in the total charge of active-site residues by -1. The charge difference is caused by the addition of Arg202 in C5. Despite having a +1 charge, Arg202 has a trivial influence on $EF_{C–H}$ due to it being perpendicular to the breaking C–H bond vector. For C6-C7 and C7-C8, the newly added residues are mostly nonpolar and are distant from the breaking C–H bond in the substrate (i.e., >6.3 Å). As electric field strength is inversely proportional to the square of the distance, the impact of remote residues dies off quickly. As such, a consistent ranking of $EF_{C–H}$ values is observed between regions beyond C6. Unexpectedly, the Spearman ρ value for C5-C6 (i.e., 0.54 and 0.49 for MM- and QM/MM-derived $EF_{C–H}$, respectively) is significantly lower than that for C4-C5 or C6-C7 (Figure 5). This is because the newly added charged residue in C6, Asp224, is positioned along the direction of the breaking C–H bond vector. As such, the impact of Asp224 on the ranking of $EF_{C–H}$ values is substantial.

By comparing the Spearman ρ values for C4-C5, C5-C6, and C7-C8, the results show that for both MM- and QM/MM-derived $EF_{C–H}$ values, the ranking is dependent more on the spatial distribution of charged residues relative to the breaking C–H bond than on the total atomic charge in the active-site regions. This finding can potentially help rational identification of residues for tuning interior enzyme electric fields for selective bond activation. To determine a convergence cutoff for computing electrostatic descriptors (i.e., for MM- and QM/MM-derived $EF_{C–H}$), we investigated the change of Spearman correlation coefficients between adjacent active-site regions

versus the increase of region size (Figure 6). Consistent with the dynamics-derived descriptors, we adopted a Spearman ρ value of 0.90 as the criterion for determining the convergence cutoff. As such, the convergence cutoff values for MM- and QM/MM-derived $EF_{C–H}$ are determined to be 6.0 and 7.0 Å, respectively. QM/MM-derived $EF_{C–H}$ demands a larger active-site region for convergence due to the involvement of charge transfer and polarization between residues in the QM region. To predict mutation effects on interior enzyme electrostatics, the use of minimal QM region in QM/MM calculation is not sufficient, albeit the incorporation of the whole enzyme in the model.

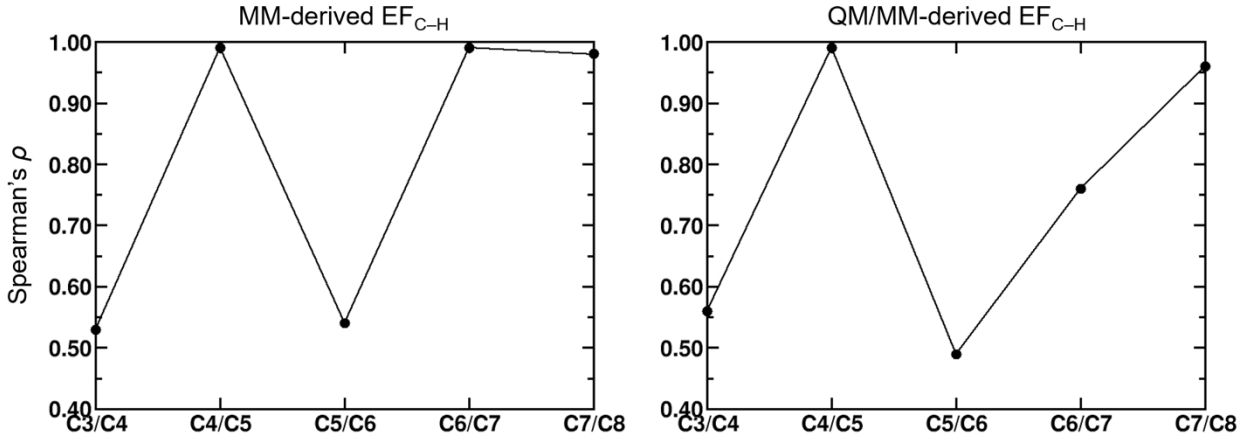

Figure 6. Spearman correlation coefficients for the dynamics-derived descriptors, MM-derived $EF_{C–H}$ (*left*) and QM/MM-derived $EF_{C–H}$ (*right*) between regions that are close in size.

Due to the high computational cost of QM/MM calculations, we investigated how predictive the ranking of MM-derived $EF_{C–H}$ (i.e., including all atoms in the enzyme) is for the ranking of QM/MM-derived $EF_{C–H}$ values under different QM region sizes (Supporting Information, Figure S2). The Spearman correlation coefficient between the MM- versus QM/MM-derived $EF_{C–H}$ values is 0.45 in C7 and 0.44 in C8. Under large QM region, the association still exists between the MM- and QM/MM-derived $EF_{C–H}$ values, albeit with a moderate correlation

strength. Considering the low computational cost of MM-derived $EF_{C–H}$ values, we would recommend a hybrid approach for future practice of computational enzyme engineering. This hybrid approach involves using MM-derived $EF_{C–H}$ values for pre-screening of a large number of mutants, followed by an assessment of QM/MM-derived $EF_{C–H}$ values to identify mutants for experimental tests. Based on the convergence test, a large-QM region should be used, but the region size could potentially be reduced by using rational QM determination approaches such as charge shift analysis,[78] Fukui shift analysis,[78] and point charge variation analysis[79].

4. Conclusions

In this work, we investigated how large an active-site region should be to converge the description of mutation effects on enzyme dynamics and electrostatics. For eighteen KE07-R7-2 variants, dynamics-derived descriptors ($RMSD_{active\_site}$ and $SASA_{ratio}$, both derived from classical MD) and electrostatic descriptors (MM- and QM/MM-derived $EF_{C–H}$) were computed across six active-site regions with various boundary distances (i.e., 3-8 Å) to the substrate. For each descriptor, we employed a Spearman correlation matrix to determine the region size condition under which further expansion of the region boundary does not substantially change the ranking of descriptor values.

Using a Spearman ρ value of 0.9 as a criterion for convergence, we observed that the ranking for $RMSD_{active\_site}$ and $SASA_{ratio}$ converges at 5 Å; MM- and QM/MM-derived $EF_{C–H}$ converge at 6.0 and 7.0 Å, respectively. Under large QM regions (i.e., 7 or 8 Å from the substrate), the ranking of MM-derived $EF_{C–H}$ (i.e., including all atoms in the enzyme) is weakly predictive to the ranking of $EF_{C–H}$ values from QM/MM computation. As such, we recommend a hybrid approach for future practice of computational enzyme engineering, which involves a pre-screening

of a large number of mutants based on MM-derived $EF_{C–H}$ values, followed by an assessment of QM/MM-derived $EF_{C–H}$ values on a smaller number of pre-screened mutants. Notably, the convergence of rankings does not ensure the convergence of measured descriptor values. Nonetheless, the ranking is most useful to guide experimental selection of function-enhancing enzyme mutants. Additionally, the current study emphasizes a designer enzyme, Kemp eliminase. Future studies should entail more types of enzymes with various catalytic actions.

ASSOCIATED CONTENT

Supporting Information. Restraints applied in the MD simulation; list of mutation sites of KE07-R7-2; list of amino acid residues involved in the region selection of KE07-R7-2; average descriptor values for 18 KE07-R7-2 variants based on C3-8 regions; Spearman correlation coefficients between whole MM- and QM/MM-derived $EF_{C–H}$. (PDF)

Initial topology and coordinate files for 18 KE07-R7-2. (ZIP)

AUTHOR INFORMATION

Corresponding Author

*Email: zhongyue.yang@vanderbilt.edu Phone: 615-343-9849

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

This research was supported by the startup grant from Vanderbilt University. Z. J. Yang, Y. Jiang, and Q. Shao are supported by the National Institute of General Medical Sciences of the National

Institutes of Health under award number R35GM146982. This work was carried out in part using computational resources from the Extreme Science and Engineering Discovery Environment (XSEDE), which is supported by National Science Foundation grant number TG-BIO200057.[103]

References

1 Koeller, K. M. & Wong, C.-H. Enzymes for chemical synthesis. *Nature* **409**, 232-240, doi:10.1038/35051706 (2001).

2 Strohmeier, G. A., Pichler, H., May, O. & Gruber-Khadjawi, M. Application of designed enzymes in organic synthesis. *Chem. Rev.* **111**, 4141-4164, doi:10.1021/cr100386u (2011).

3 Petchey, M. R. & Grogan, G. Enzyme-Catalysed Synthesis of Secondary and Tertiary Amides. *Advanced Synthesis & Catalysis* **361**, 3895-3914, doi:https://doi.org/10.1002/adsc.201900694 (2019).

4 Chundawat, S. P., Beckham, G. T., Himmel, M. E. & Dale, B. E. Deconstruction of lignocellulosic biomass to fuels and chemicals. *Annu. Rev. Chem. Biomol. Eng.* **2**, 121-145, doi:10.1146/annurev-chembioeng-061010-114205 (2011).

5 Yang, B., Dai, Z., Ding, S.-Y. & Wyman, C. E. Enzymatic hydrolysis of cellulosic biomass. *Biofuels* **2**, 421-449, doi:10.4155/bfs.11.116 (2011).

6 Sweeney, M. D. & Xu, F. Biomass Converting Enzymes as Industrial Biocatalysts for Fuels and Chemicals: Recent Developments. *Catalysts* **2**, 244-263 (2012).

7 Horn, S. J., Vaaje-Kolstad, G., Westereng, B. & Eijsink, V. G. Novel enzymes for the degradation of cellulose. *Biotechnol Biofuels* **5**, 45, doi:10.1186/1754-6834-5-45 (2012).

8 Austin, H. P. *et al.* Characterization and engineering of a plastic-degrading aromatic polyesterase. *Proc. Natl. Acad. Sci. U. S. A.* **115**, E4350-E4357, doi:10.1073/pnas.1718804115 (2018).

9 Knott, B. C. *et al.* Characterization and engineering of a two-enzyme system for plastics depolymerization. *Proc. Natl. Acad. Sci. U. S. A.* **117**, 25476-25485, doi:10.1073/pnas.2006753117 (2020).

10 Tiso, T. *et al.* Towards bio-upcycling of polyethylene terephthalate. *Metab Eng* **66**, 167-178, doi:10.1016/j.ymben.2021.03.011 (2021).

11 Ellis, L. D. *et al.* Chemical and biological catalysis for plastics recycling and upcycling. *Nat. Catal.* **4**, 539-556, doi:10.1038/s41929-021-00648-4 (2021).

12 Fessner, N. D. P450 Monooxygenases Enable Rapid Late-Stage Diversification of Natural Products via C-H Bond Activation. *ChemCatChem* **11**, 2226-2242, doi:10.1002/cctc.201801829 (2019).

13 Hong, B., Luo, T. & Lei, X. Late-Stage Diversification of Natural Products. *ACS Cent. Sci.* **6**, 622-635, doi:10.1021/acscentsci.9b00916 (2020).

14 Romero, E. *et al.* Enzymatic Late-Stage Modifications: Better Late Than Never. *Angew. Chem. Int. Ed.* **60**, 16824-16855, doi:https://doi.org/10.1002/anie.202014931 (2021).

15 Craven, E. J. *et al.* Programmable late-stage C−H bond functionalization enabled by integration of enzymes with chemocatalysis. *Nat. Catal.* **4**, 385-394, doi:10.1038/s41929-021-00603-3 (2021).

16 Gordon, S. R. *et al.* Computational design of an alpha-gliadin peptidase. *J. Am. Chem. Soc.* **134**, 20513-20520, doi:10.1021/ja3094795 (2012).
17 Sun, S. *et al.* Selection of a versatile Lactobacillus plantarum for wine production and identification and preliminary characterisation of a novel histamine-degrading enzyme. *International Journal of Food Science & Technology* **55**, 2608-2618, doi:https://doi.org/10.1111/ijfs.14514 (2020).
18 Samadi, N. *et al.* Gastric Enzyme Supplementation Inhibits Food Allergy in a BALB/c Mouse Model. *Nutrients* **13**, doi:10.3390/nu13030738 (2021).
19 Bloom, J. D. *et al.* Evolving strategies for enzyme engineering. *Curr. Opin. Struct. Biol.* **15**, 447-452, doi:10.1016/j.sbi.2005.06.004 (2005).
20 Alejaldre, L., Pelletier, J. N. & Quaglia, D. Methods for enzyme library creation: Which one will you choose? *BioEssays* **43**, 2100052, doi:https://doi.org/10.1002/bies.202100052 (2021).
21 Ravikumar, Y., Nadarajan, S. P., Yoo, T. H., Lee, C. S. & Yun, H. Unnatural amino acid mutagenesis-based enzyme engineering. *Trends Biotechnol* **33**, 462-470, doi:10.1016/j.tibtech.2015.05.002 (2015).
22 Kolkman, J. A. & Stemmer, W. P. Directed evolution of proteins by exon shuffling. *Nat Biotechnol* **19**, 423-428, doi:10.1038/88084 (2001).
23 Akbulut, N., Tuzlakoglu Ozturk, M., Pijning, T., Issever Ozturk, S. & Gumusel, F. Improved activity and thermostability of Bacillus pumilus lipase by directed evolution. *J Biotechnol* **164**, 123-129, doi:10.1016/j.jbiotec.2012.12.016 (2013).
24 Reetz, M. T., Bocola, M., Carballeira, J. D., Zha, D. & Vogel, A. Expanding the range of substrate acceptance of enzymes: combinatorial active-site saturation test. *Angew. Chem., Int. Ed. Engl.* **44**, 4192-4196, doi:10.1002/anie.200500767 (2005).
25 Reetz, M. T. *et al.* Expanding the substrate scope of enzymes: combining mutations obtained by CASTing. *Chemistry* **12**, 6031-6038, doi:10.1002/chem.200600459 (2006).
26 Arnold, F. H. & Volkov, A. A. Directed evolution of biocatalysts. *Curr Opin Chem Biol* **3**, 54-59, doi:10.1016/s1367-5931(99)80010-6 (1999).
27 Packer, M. S. & Liu, D. R. Methods for the directed evolution of proteins. *Nat. Rev. Genet.* **16**, 379-394, doi:10.1038/nrg3927 (2015).
28 Arnold, F. H. Directed Evolution: Bringing New Chemistry to Life. *Angew. Chem., Int. Ed. Engl.* **57**, 4143-4148, doi:10.1002/anie.201708408 (2018).
29 Wang, Y. *et al.* Directed Evolution: Methodologies and Applications. *Chem. Rev.* **121**, 12384-12444, doi:10.1021/acs.chemrev.1c00260 (2021).
30 Rothlisberger, D. *et al.* Kemp elimination catalysts by computational enzyme design. *Nature* **453**, 190-195, doi:10.1038/nature06879 (2008).
31 Privett, H. K. *et al.* Iterative approach to computational enzyme design. *Proc. Natl. Acad. Sci. U. S. A.* **109**, 3790-3795, doi:10.1073/pnas.1118082108 (2012).
32 Korendovych, I. V. *et al.* Design of a switchable eliminase. *Proc. Natl. Acad. Sci. U. S. A.* **108**, 6823-6827, doi:10.1073/pnas.1018191108 (2011).
33 Korendovych, I. V. in *Protein Engineering: Methods and Protocols* (eds Uwe T. Bornscheuer & Matthias Höhne) 15-23 (Springer New York, 2018).
34 Alonso-Cotchico, L., Rodri Guez-Guerra, J., Lledos, A. & Marechal, J. D. Molecular Modeling for Artificial Metalloenzyme Design and Optimization. *Acc. Chem. Res.* **53**, 896-905, doi:10.1021/acs.accounts.0c00031 (2020).

35 Khersonsky, O. *et al.* Automated Design of Efficient and Functionally Diverse Enzyme Repertoires. *Mol Cell* **72**, 178-186 e175, doi:10.1016/j.molcel.2018.08.033 (2018).
36 Le, P., Zhao, J. & Franzen, S. Correlation of heme binding affinity and enzyme kinetics of dehaloperoxidase. *Biochemistry* **53**, 6863-6877, doi:10.1021/bi5005975 (2014).
37 Luo, Q., Chen, D., Boom, R. M. & Janssen, A. E. M. Revisiting the enzymatic kinetics of pepsin using isothermal titration calorimetry. *Food Chemistry* **268**, 94-100, doi:https://doi.org/10.1016/j.foodchem.2018.06.042 (2018).
38 Richard, J. P. Protein Flexibility and Stiffness Enable Efficient Enzymatic Catalysis. *J. Am. Chem. Soc.* **141**, 3320-3331, doi:10.1021/jacs.8b10836 (2019).
39 Amrein, B. A. *et al.* CADEE: Computer-Aided Directed Evolution of Enzymes. *IUCrJ* **4**, 50-64, doi:10.1107/S2052252516018017 (2017).
40 Yao, J., Chen, X., Zheng, F. & Zhan, C. G. Catalytic Reaction Mechanism for Drug Metabolism in Human Carboxylesterase-1: Cocaine Hydrolysis Pathway. *Mol Pharm* **15**, 3871-3880, doi:10.1021/acs.molpharmaceut.8b00354 (2018).
41 Hur, S. & Bruice, T. C. The near attack conformation approach to the study of the chorismate to prephenate reaction. *Proc. Natl. Acad. Sci. U. S. A.* **100**, 12015-12020, doi:10.1073/pnas.1534873100 (2003).
42 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. U. S. A.* **109**, 10358-10363, doi:10.1073/pnas.1121063109 (2012).
43 Vaissier Welborn, V. & Head-Gordon, T. Computational Design of Synthetic Enzymes. *Chem. Rev.* **119**, 6613-6630, doi:10.1021/acs.chemrev.8b00399 (2019).
44 Mehmood, R., Vennelakanti, V. & Kulik, H. J. Spectroscopically Guided Simulations Reveal Distinct Strategies for Positioning Substrates to Achieve Selectivity in Nonheme Fe(II)/α-Ketoglutarate-Dependent Halogenases. *ACS Catal.* **11**, 12394-12408, doi:10.1021/acscatal.1c03169 (2021).
45 Liu, C. T. *et al. Escherichia coli* dihydrofolate reductase catalyzed proton and hydride transfers: Temporal order and the roles of Asp27 and Tyr100. *Proc. Natl. Acad. Sci. U. S. A.* **111**, 18231-18236, doi:doi:10.1073/pnas.1415940111 (2014).
46 Masterson, J. E. & Schwartz, S. D. Evolution alters the enzymatic reaction coordinate of dihydrofolate reductase. *J. Phys. Chem. B* **119**, 989-996, doi:10.1021/jp506373q (2015).
47 Liao, Q. *et al.* Loop Motion in Triosephosphate Isomerase Is Not a Simple Open and Shut Case. *J. Am. Chem. Soc.* **140**, 15889-15903, doi:10.1021/jacs.8b09378 (2018).
48 Gao, S. *et al.* Hydrogen–Deuterium Exchange within Adenosine Deaminase, a TIM Barrel Hydrolase, Identifies Networks for Thermal Activation of Catalysis. *J. Am. Chem. Soc.* **142**, 19936-19949, doi:10.1021/jacs.0c07866 (2020).
49 Bunzel, H. A. *et al.* Emergence of a Negative Activation Heat Capacity during Evolution of a Designed Enzyme. *J. Am. Chem. Soc.* **141**, 11745-11748, doi:10.1021/jacs.9b02731 (2019).
50 Fried, S. D. & Boxer, S. G. Electric Fields and Enzyme Catalysis. *Annu. Rev. Biochem.* **86**, 387-415, doi:10.1146/annurev-biochem-061516-044432 (2017).
51 Bhowmick, A., Sharma, S. C. & Head-Gordon, T. The Importance of the Scaffold for de Novo Enzymes: A Case Study with Kemp Eliminase. *J. Am. Chem. Soc.* **139**, 5793-5800, doi:10.1021/jacs.6b12265 (2017).

52 Vaissier, V., Sharma, S. C., Schaettle, K., Zhang, T. & Head-Gordon, T. Computational Optimization of Electric Fields for Improving Catalysis of a Designed Kemp Eliminase. *ACS Catal.* **8**, 219-227, doi:10.1021/acscatal.7b03151 (2018).
53 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, doi:10.1021/jacs.9b05323 (2019).
54 Yang, Z., Liu, F., Steeves, A. H. & Kulik, H. J. Quantum Mechanical Description of Electrostatics Provides a Unified Picture of Catalytic Action Across Methyltransferases. *J. Phys. Chem. Lett.* **10**, 3779-3787, doi:10.1021/acs.jpclett.9b01555 (2019).
55 Bím, D. & Alexandrova, A. N. Local Electric Fields As a Natural Switch of Heme-Iron Protein Reactivity. *ACS Catal.* **11**, 6534-6546, doi:10.1021/acscatal.1c00687 (2021).
56 Hong, N. S. *et al.* The evolution of multiple active site configurations in a designed enzyme. *Nat. Commun.* **9**, 3900, doi:10.1038/s41467-018-06305-y (2018).
57 Carlin, D. A. *et al.* Kinetic Characterization of 100 Glycoside Hydrolase Mutants Enables the Discovery of Structural Features Correlated with Kinetic Constants. *PloS One* **11**, e0147596, doi:10.1371/journal.pone.0147596 (2016).
58 Jiang, Y., Yan, B., Chen, Y., Juarez, R. J. & Yang, Z. J. Molecular Dynamics-Derived Descriptor Informs the Impact of Mutation on the Catalytic Turnover Number in Lactonase Across Substrates. *J. Phys. Chem. B* **126**, 2486-2495, doi:10.1021/acs.jpcb.2c00142 (2022).
59 Bartlett, G. J., Porter, C. T., Borkakoti, N. & Thornton, J. M. Analysis of catalytic residues in enzyme active sites. *J. Mol. Biol.* **324**, 105-121, doi:10.1016/s0022-2836(02)01036-7 (2002).
60 Lodola, A. *et al.* Structural Fluctuations in Enzyme-Catalyzed Reactions: Determinants of Reactivity in Fatty Acid Amide Hydrolase from Multivariate Statistical Analysis of Quantum Mechanics/Molecular Mechanics Paths. *J. Chem. Theory Comput.* **6**, 2948-2960, doi:10.1021/ct100264j (2010).
61 Yabukarski, F. *et al.* Assessment of enzyme active site positioning and tests of catalytic mechanisms through X-ray-derived conformational ensembles. *Proc. Natl. Acad. Sci. U. S. A.* **117**, 33204-33215, doi:10.1073/pnas.2011350117 (2020).
62 Warshel, A. & Levitt, M. Theoretical studies of enzymic reactions: dielectric, electrostatic and steric stabilization of the carbonium ion in the reaction of lysozyme. *J. Mol. Biol.* **103**, 227-249, doi:10.1016/0022-2836(76)90311-9 (1976).
63 Vanpoucke, D. E., Olah, J., De Proft, F., Van Speybroeck, V. & Roos, G. Convergence of atomic charges with the size of the enzymatic environment. *J. Chem. Inf. Model* **55**, 564-571, doi:10.1021/ci5006417 (2015).
64 Kulik, H. J., Zhang, J., Klinman, J. P. & Martínez, T. J. How Large Should the QM Region Be in QM/MM Calculations? The Case of Catechol O-Methyltransferase. *J. Phys. Chem. B* **120**, 11381-11394, doi:10.1021/acs.jpcb.6b07814 (2016).
65 Karelina, M. & Kulik, H. J. Systematic Quantum Mechanical Region Determination in QM/MM Simulation. *J. Chem. Theory Comput.* **13**, 563-576, doi:10.1021/acs.jctc.6b01049 (2017).
66 Morgenstern, A., Jaszai, M., Eberhart, M. E. & Alexandrova, A. N. Quantified electrostatic preorganization in enzymes using the geometry of the electron charge density. *Chem. Sci.* **8**, 5010-5018, doi:10.1039/C7SC01301A (2017).

67 Mehmood, R. & Kulik, H. J. Both Configuration and QM Region Size Matter: Zinc Stability in QM/MM Models of DNA Methyltransferase. *J. Chem. Theory Comput.* **16**, 3121-3134, doi:10.1021/acs.jctc.0c00153 (2020).
68 Solt, I. *et al.* Evaluating boundary dependent errors in QM/MM simulations. *J. Phys. Chem. B* **113**, 5728-5735, doi:10.1021/jp807277r (2009).
69 Sumowski, C. V. & Ochsenfeld, C. A convergence study of QM/MM isomerization energies with the selected size of the QM region for peptidic systems. *J Phys Chem A* **113**, 11734-11741, doi:10.1021/jp902876n (2009).
70 Liao, R. Z. & Thiel, W. Convergence in the QM-only and QM/MM modeling of enzymatic reactions: A case study for acetylene hydratase. *J. Comput. Chem.* **34**, 2389-2397, doi:10.1002/jcc.23403 (2013).
71 Sadeghian, K. *et al.* Ribose-protonated DNA base excision repair: a combined theoretical and experimental study. *Angew. Chem., Int. Ed. Engl.* **53**, 10044-10048, doi:10.1002/anie.201403334 (2014).
72 Jindal, G. & Warshel, A. Exploring the Dependence of QM/MM Calculations of Enzyme Catalysis on the Size of the QM Region. *J. Phys. Chem. B* **120**, 9913-9921, doi:10.1021/acs.jpcb.6b07203 (2016).
73 Benediktsson, B. & Bjornsson, R. QM/MM Study of the Nitrogenase MoFe Protein Resting State: Broken-Symmetry States, Protonation States, and QM Region Convergence in the FeMoco Active Site. *Inorg Chem* **56**, 13417-13429, doi:10.1021/acs.inorgchem.7b02158 (2017).
74 Kang, H. & Zheng, M. Influence of the quantum mechanical region size in QM/MM modelling: A case study of fluoroacetate dehalogenase catalyzed CF bond cleavage. *Computational and Theoretical Chemistry* **1204**, 113399, doi:https://doi.org/10.1016/j.comptc.2021.113399 (2021).
75 Demapan, D., Kussmann, J., Ochsenfeld, C. & Cui, Q. Factors That Determine the Variation of Equilibrium and Kinetic Properties of QM/MM Enzyme Simulations: QM Region, Conformation, and Boundary Condition. *J. Chem. Theory Comput.* **18**, 2530-2542, doi:10.1021/acs.jctc.1c00714 (2022).
76 Flaig, D., Beer, M. & Ochsenfeld, C. Convergence of Electronic Structure with the Size of the QM Region: Example of QM/MM NMR Shieldings. *J. Chem. Theory Comput.* **8**, 2260-2271, doi:10.1021/ct300036s (2012).
77 Hartman, J. D., Neubauer, T. J., Caulkins, B. G., Mueller, L. J. & Beran, G. J. Converging nuclear magnetic shielding calculations with respect to basis and system size in protein systems. *J Biomol NMR* **62**, 327-340, doi:10.1007/s10858-015-9947-2 (2015).
78 Qi, H. W., Karelina, M. & Kulik, H. J. Quantifying Electronic Effects in QM and QM/MM Biomolecular Modeling with the Fukui Function. *Acta Physico-Chimica Sinica* **34**, 81-91, doi:10.3866/Pku.Whxb201706303 (2018).
79 Brandt, F. & Jacob, C. R. Systematic QM Region Construction in QM/MM Calculations Based on Uncertainty Quantification. *J. Chem. Theory Comput.* **18**, 2584-2596, doi:10.1021/acs.jctc.1c01093 (2022).
80 Bhowmick, A., Sharma, S. C., Honma, H. & Head-Gordon, T. The role of side chain entropy and mutual information for improving the de novo design of Kemp eliminases KE07 and KE70. *Phys. Chem. Chem. Phys.* **18**, 19386-19396, doi:10.1039/c6cp03622h (2016).
81 The PyMOL Molecular Graphics System, Version 2.4 (2015).

82 AMBER 2018 (University of California, San Francisco, 2018).
83 Maier, J. A. *et al.* ff14SB: Improving the Accuracy of Protein Side Chain and Backbone Parameters from ff99SB. *J. Chem. Theory Comput.* **11**, 3696-3713, doi:10.1021/acs.jctc.5b00255 (2015).
84 Wang, N. X. & Wilson, A. K. The behavior of density functionals with respect to basis set. I. The correlation consistent basis sets. *J. Chem. Phys.* **121**, 7632-7646, doi:10.1063/1.1792071 (2004).
85 Wang, J., Wang, W., Kollman, P. A. & Case, D. A. Automatic atom type and bond type perception in molecular mechanical calculations. *J. Mol. Graphics Modell.* **25**, 247-260, doi:10.1016/j.jmgm.2005.12.005 (2006).
86 Jakalian, A., Jack, D. B. & Bayly, C. I. Fast, efficient generation of high-quality atomic charges. AM1-BCC model: II. Parameterization and validation. *J. Comput. Chem.* **23**, 1623-1641, doi:10.1002/jcc.10128 (2002).
87 Shao, Q., Jiang, Y. & Yang, Z. J. EnzyHTP: A High-Throughput Computational Platform for Enzyme Modeling. *J. Chem. Inf. Model* **62**, 647-655, doi:10.1021/acs.jcim.1c01424 (2022).
88 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, doi:https://doi.org/10.1016/0021-9991(77)90098-5 (1977).
89 Loncharich, R. J., Brooks, B. R. & Pastor, R. W. Langevin dynamics of peptides: the frictional dependence of isomerization rates of N-acetylalanyl-N'-methylamide. *Biopolymers* **32**, 523-535, doi:10.1002/bip.360320508 (1992).
90 Berendsen, H. J. C., Postma, J. P. M., Gunsteren, W. F. v., DiNola, A. & Haak, J. R. Molecular dynamics with coupling to an external bath. *J. Chem. Phys.* **81**, 3684-3690, doi:10.1063/1.448118 (1984).
91 Ufimtsev, I. S. & Martinez, T. J. Quantum Chemistry on Graphical Processing Units. 3. Analytical Energy Gradients, Geometry Optimization, and First Principles Molecular Dynamics. *J. Chem. Theory Comput.* **5**, 2619-2628, doi:10.1021/ct9003004 (2009).
92 Titov, A. V., Ufimtsev, I. S., Luehr, N. & Martinez, T. J. Generating Efficient Quantum Chemistry Codes for Novel Architectures. *J. Chem. Theory Comput.* **9**, 213-221, doi:10.1021/ct300321a (2013).
93 Rohrdanz, M. A., Martins, K. M. & Herbert, J. M. A long-range-corrected density functional that performs well for both ground-state properties and time-dependent density functional theory excitation energies, including charge-transfer excited states. *J. Chem. Phys.* **130**, 054112, doi:10.1063/1.3073302 (2009).
94 Hariharan, P. C. & Pople, J. A. The influence of polarization functions on molecular orbital hydrogenation energies. *Theor. Chim. Acta.* **28**, 213-222, doi:10.1007/BF00533485 (1973).
95 Cornell, W. D., Cieplak, P., Bayly, C. I. & Kollman, P. A. Application of RESP charges to calculate conformational energies, hydrogen bond energies, and free energies of solvation. *J. Am. Chem. Soc.* **115**, 9620-9631, doi:10.1021/ja00074a030 (1993).
96 Shrake, A. & Rupley, J. A. Environment and exposure to solvent of protein atoms. Lysozyme and insulin. *J. Mol. Biol.* **79**, 351-371, doi:10.1016/0022-2836(73)90011-9 (1973).

97 McGibbon, R. T. *et al.* MDTraj: A Modern Open Library for the Analysis of Molecular Dynamics Trajectories. *Biophys. J.* **109**, 1528-1532, doi:10.1016/j.bpj.2015.08.015 (2015).
98 Alexandrova, A. N., Rothlisberger, D., Baker, D. & Jorgensen, W. L. Catalytic Mechanism and Performance of Computationally Designed Enzymes for Kemp Elimination. *J. Am. Chem. Soc.* **130**, 15907-15915, doi:10.1021/ja804040s (2008).
99 Khersonsky, O. *et al.* Evolutionary Optimization of Computationally Designed Enzymes: Kemp Eliminases of the KE07 Series. *J. Mol. Biol.* **396**, 1025-1042, doi:10.1016/j.jmb.2009.12.031 (2010).
100 Caselle, E. A. *et al.* Kemp Eliminases of the AlleyCat Family Possess High Substrate Promiscuity. *ChemCatChem* **11**, 1425-1430, doi:10.1002/cctc.201801994 (2019).
101 Welborn, V. V., Pestana, L. R. & Head-Gordon, T. Computational optimization of electric fields for better catalysis design. *Nat. Catal.* **1**, 649-655, doi:10.1038/s41929-018-0109-2 (2018).
102 Fleishman, S. J. *et al.* RosettaScripts: a scripting language interface to the Rosetta macromolecular modeling suite. *PloS One* **6**, e20161, doi:10.1371/journal.pone.0020161 (2011).
103 Towns, J. *et al.* XSEDE: Accelerating Scientific Discovery. *Computing in Science & Engineering* **16**, 62-74, doi:10.1109/MCSE.2014.80 (2014).

Graphic Abstract