Machine Learning Identifies Chemical Characteristics That Promote Enzyme Catalysis (Supporting Information)

Brian M. Bonk[1,2], James W. Weis[2,3,4], Bruce Tidor*[1,2,3,4]

[1] Department of Biological Engineering, Massachusetts Institute of Technology, 77 Massachusetts Avenue, Cambridge, Massachusetts 02139, United States

[2] Computer Science and Artificial Intelligence Laboratory, Massachusetts Institute of Technology, 77 Massachusetts Avenue, Cambridge, Massachusetts 02139, United States

[3] Computational and Systems Biology, Massachusetts Institute of Technology, 77 Massachusetts Avenue, Cambridge, Massachusetts 02139, United States

[4] Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology, 77 Massachusetts Avenue, Cambridge, Massachusetts 02139, United States

SUPPORTING METHODS

Structure Preparation. The crystal structure of *Spinacia oleracea* KARI bound to the transition-state analog N-hydroxy-N-isopropyloxamate was obtained from the Protein Data Bank[27–29] with the accession code 1YVE[26] and prepared as described previously by Silver.[12] Although the enzyme crystallizes as a homodimer with two identical active sites,[26] only the chain A monomer was used for all simulations in order to improve computational efficiency. This choice was justified by the significant separation between the active sites of the two monomers[26] (Figure S3). Histidine side-chain orientation and protonation for the following chain A residues was selected to maximize hydrogen-bonding potential, resulting in no changes to histidine orientation, no doubly protonated histidine side chains, and neutral histidine protonation as indicated: 103-δ, 215-δ, 226-δ, 232-δ, 280-ε, 328-ε, 484-δ, 506-ε, and 564-ε. Crystallographic water molecules that were neither in the active site nor made at least three hydrogen bonds with the protein (using a maximum heavy-atom hydrogen-bond distance of 3.33 Å) were removed. The 61 water molecules remaining had residue identifiers of 72, 75, 87, 93, 106, 109, 179, 194, 379, 405, 429, 440, 474, 481, 838–841, 852, 862, 878, 883, 887, 894, 895, 941–949, 965, 967–969, 975, 998, 999, 1023–1025, 1032, 1072, 1089, 1093–1095, 1097, 1105, 1108, 1206, 1250, 1252, 1253, 1257, 1304, 1305, and 1779.

A model of the substrate-bound enzyme was then constructed by running an *in vacuo* QM ground-state minimization of the substrate, two magnesium centers, five magnesium-coordinating water molecules, and the side chains of three surrounding active-site residues, Asp 315, Glu 319, and Glu 496. Glu 496 was protonated, consistent with previous studies indicating its importance in stabilizing the transition and product state by forming a hydrogen bond with the substrate O8.[30] The GAUSSIAN03 computer program[31] was used to perform *in vacuo* QM calculations at the rhf/3-21g* level of theory, using ground-state energy minimization (keyword OPT) to obtain reactant and product structures and a saddle-point search (keyword QST3) to obtain the transition-state structure. Both types of optimization were performed using the Berny algorithm.[32,33] To ensure low-energy pathways to the reactant and product state of isomerization, the resulting transition state was validated by following the vibrational eigenmode corresponding to the single negative eigenvalue.

Each of the optimized and validated QM-derived structures was combined with the prepared crystallographic structure for the rest of the enzyme by alignment of the carbon atoms of the QM-optimized substrate to the crystallographic transition-state analog, followed by ten rounds of sliding, restrained minimization. During this minimization, which consisted of 100 steps of steepest descent minimization followed by 100 steps of adopted basis Newton-Raphson minimization, all substrate, magnesium ion, and coordinating aspartate and glutamate oxygen atoms were held fixed, and the remaining active-site residues were harmonically restrained using a force constant of 50 kcal/(mol·Å^2). Harmonic restraints were reset after each round of minimization.

Simulation Methodology. CHARMM version 41[34,35] compiled with the SQUANTUM option was used to perform all molecular dynamics simulations. The QM portion of the energy function was calculated with the AM1 semiempirical quantum mechanical force field;[36] the MM portion of the energy function was computed using the CHARMM36 all-atom force field.[37] Additional AM1 parameters were used for the magnesium ions.[38] The following atoms made up the QM region: substrate (acetolactate), both magnesium centers, five magnesium-coordinating active-site water molecules, the side chains of Asp 315, Glu 319, and Glu 496, and the nicotinamide group of NADPH (Figure 1C). The Generalized Hybrid Orbital method[39] was used to treat the QM/MM boundary atoms, included the Cα atoms of residues Asp 315, Glu 319, and Glu 496, as well as the C5' atom of the ribose ring in NADPH linking to the nicotinamide group. The substrate O6 was deprotonated and the coordinating Glu 496 was protonated, paralleling previous QM/MM studies of KARI.[30] All molecular dynamics simulations were performed *in vacuo* with a distance dependent dielectric (4*r*) using the leapfrog integrator at 300 K with a 1 fs integration time step.

Seed Trajectory Generation. The initial reactive trajectories used to bootstrap the TIS simulations were found by computing a potential of mean force (PMF) along the order parameter λ, defined as the difference of the distance between the substrate breaking bond (C4–C5) and the forming bond (C5–C7), which has units of angstroms. This PMF was computed using umbrella sampling and the weighted histogram analysis method.[40] The umbrella sampling was performed in CHARMM41 using the RXNCOR module with windows 0.05 Å in width and harmonic restraints of 300 kcal/(mol·Å^2). The resulting PMF provided an estimate of the location of the transition state along the order parameter λ, roughly within the $-0.05 < \lambda < +0.05$ region. Candidate seed trajectories were then generated by integrating forward and backward for 2000 fs without restraints starting from a randomly chosen frame from the umbrella sampling window ensembles centered at λ values of −0.05, 0.00, and +0.05. Trajectories were selected as successful seed trajectories if they connected the reactant basin ($\lambda < -1$) and product basin ($\lambda > +1$).

Training Data Set Generation and Time Point Selection. Three randomly-selected connecting seed trajectories from the collection described above were used as starting trajectories for the generation of a larger ensemble of reactive and almost-reactive trajectories. Each seed was used to generate 9 reactive ensembles and 9 almost-reactive ensembles of 20,000 trajectories each. The combined data set contained 461,422 almost-reactive and 618,578 reactive trajectories. The greater number of reactive trajectories resulted because the sampling process for almost-reactive trajectories could also generate some reactive ones, but the sampling process for reactive trajectories could not generate almost-reactive ones. When the almost-reactive process produced a reactive trajectory, it was removed from that set and added to the reactive data set. To ensure a balanced number of reactive and almost-reactive trajectories in each training and testing data set, the reactive trajectories were randomly sampled without replacement to produce a set of 461,422 reactive trajectories.

For the reactive ensembles, the product interface was defined as $\lambda_R = +1.00$, and for the almost-reactive ensembles, the product interface was defined as $\lambda_{AR} = -0.20$ (Figure 1A). In both ensembles, the reactant interface was defined as $\lambda = -1.00$. To collect time points early in the reactant basin for analysis, integration was not stopped once a trajectory reached the reactant and product interface (and had been accepted into the Markov chain), but continued forward and backward for a total of 200 fs in each direction. A MATLAB wrapper that launched individual CHARMM41 trajectory runs was used to perform all TIS computations.

To ensure that candidate features (see below) were computed at analogous time points between reactive and almost-reactive trajectory ensembles, in a postprocessing step, all almost-reactive and reactive trajectories from all 27 pairs of ensembles were time-shifted such that the 0 fs time point corresponded to the bottom of the last "trough" in λ (when plotted vs. time) before the prospective alkyl migration event, a geometric feature that all the collected trajectories shared (Figure 1D). Chemically, the last trough represents the point at which the C4–C5 bond is most compressed, before, like a spring, launching into the prospective bond-breaking event (whether or not that event occurred). This trough was found by first finding the point in the trajectory closest to the transition region at $\lambda = 0$, then scanning along the trajectory backward from this point until the first change in sign of the derivative of λ with respect to time was found with a value of λ less than 0 (i.e., was located in the reactant basin). All other time points were defined relative to this first trough at time 0. Cartesian coordinate frames of atomic positions were collected in 5 fs increments from the 0 fs time point, going backward to –150 fs and forward to +35 fs from the t=0 fs point, for a total of 38 total time points. This collection of subsampled time points was used for all subsequent analysis.

Feature Computation. At each of the 38 time points between –150 and +35 fs, the set of 68 structural features in Table S1 were computed for each of the trajectories in each of the 27 reactive and 27 almost-reactive ensembles. The 68 features are illustrated structurally in Figure S2A (distances), Figure S2B (angles), and Figure S2C (dihedrals). For the generation of Table S5 and Figure S8, the velocity magnitudes of the 341 atoms (including all hydrogens) within 5 Å (based on the starting aligned, minimized model) of the migrating methyl (AC6/C5) were also computed at the same time points. These data were pooled across ensembles to produce one combined reactive and one combined almost-reactive data set at each of the 38 time points, which were used in machine learning and subsequent analysis described below and stored as a row in a data matrix. A separate data matrix was constructed for each time point by augmenting the 68 computed features with the trajectory outcome (1 for reactive or 0 for almost-reactive), as well as the ensemble and trajectory indices. For model training, the data matrix at each time point was randomly sampled without replacement to produce five equal partitions containing 73,827 trajectories each, and for model testing, the remaining trajectories were randomly sampled to produce five equal partitions containing 18,456 trajectories each.

Machine Learning. For feature regularization and discovery, the LASSO method[41] was used with the *lassoglm* implementation in MATLAB. For an intercept β_0 and predictor coefficients β_j, LASSO solves the general problem,

$$\min_{\beta_0,\beta}\left(\frac{1}{N}\sum_{i=1}^{N}\rho_{\beta_0,\beta}(X_i,Y_i)+\lambda\sum_{j=1}^{p}|\beta_j|\right)$$

where p is the number of input predictor features, N is the number of observables (the number of reactive and almost-reactive trajectories used in a given LASSO training set), the X_i are each a p-dimensional vector of predictor features (generally interatomic distances, angles, and dihedrals), the Y_i are scalar outcomes (1 for a trajectory that was reactive and 0 for one that was almost-reactive), λ is a nonnegative regularization (penalty strength) parameter, and an underlying logistic learning model was composed of an intercept β_0, a set of p feature coefficients β_j, and the loss function $\rho_{\beta_0,\beta}(X_i,Y_i)$.

Due to the binary nature of the response variables, a logistic loss function was used,

$$\rho_{\beta_0,\beta}(x,y)=-y(\beta_0+\sum_{j=1}^{p}\beta_j x_j)+\log\left(1+e^{(\beta_0+\sum_{j=1}^{p}\beta_j x_j)}\right)$$

where x and y denote individual observations of X_i and Y_i. Note that only the values of the predictor coefficients β_j were penalized using LASSO, and not the value of the intercept β_0. In order to select a given number of features with LASSO, the regularization parameter λ was adjusted until a specific number m (1, 5, 10, 15, 20, 25, or 30) of non-zero coefficients β_j remained (using a tolerance of 1.0×10^{-4}). These m LASSO-selected predictor features with non-zero coefficients were then fit using the *fitglm* function in MATLAB to a logistic classifier of the form:

$$\mu=\frac{e^{\left(\beta_0+\sum_{j=1}^{m}\beta_j x_j\right)}}{1+e^{\left(\beta_0+\sum_{j=1}^{m}\beta_j x_j\right)}}$$

where μ is the probability of evaluating to 1 (reactive) given a specific linear combination of predictor features x_j. Trajectories were considered reactive if this probability evaluated to greater than 0.5 ($\beta_0+\sum_{j=1}^{p}\beta_j x_j > 0.0$) and non-reactive if this probability evaluated to less than or equal to 0.5 ($\beta_0+\sum_{j=1}^{p}\beta_j x_j \leq 0.0$). The logistic classifier essentially defines a hyperplane with the equation $\beta_0+\sum_{j=1}^{p}\beta_j x_j = 0$ that partitions the reactant well in two, with reactive predictions on one side and non-reactive on the other.

After fitting predictor coefficients, the area under the curve of the receiver operating characteristic (AUC) was computed for each logistic classifier using the *perfcurve* function in MATLAB to vary the classifier threshold β_0 in order to generate a receiver operating characteristic, and subsequently compute the area under the resulting curve. Other classifier performance metrics were computed using the *classperf* function in MATLAB, where accuracy was defined as the number of correctly classified trajectories divided by the total number of trajectories, sensitivity was defined as the number of correctly classified reactive trajectories divided by the total number of reactive trajectories, and specificity was defined as the number of correctly classified almost-reactive trajectories divided by the total number of almost-reactive trajectories.

Cluster Assignment. Reactive clusters were assigned by k-means clustering, with the *kmeans* function in MATLAB using $k = 5$ applied to the matrix of consensus feature Z-scores weighted by their corresponding logistic coefficient β_j for all correctly classified reactive trajectories. The number of clusters (5) was chosen based on a hierarchical clustering analysis also performed in MATLAB (data not shown). The Euclidian distance of the consensus feature set from each almost-reactive trajectory to each of the five k-means centers was computed, and each almost-reactive trajectory was then assigned to the cluster with the shortest Euclidian distance to its respective centroid.

Rate Constant Computations. The TIS rate constant was computed as the product of two terms––a flux term and a probability term denoted $P(\lambda_B|\lambda_1)$.[21] The flux term represents the number of crossings through interface λ_1 coming directly from state A (also referred to as the reactant basin, defined as all points for which $\lambda \leq \lambda_A = -0.8$), normalized by the total time spent in state A. The probability term represents the probability for a trajectory to reach interface λ_B given that it crossed interface λ_1, and for computational efficiency can be decomposed into a series of conditional probabilities:

$$\mathcal{P}(\lambda_B|\lambda_1) = \prod_{i=1}^{n-1} \mathcal{P}(\lambda_{i+1}|\lambda_i)\mathcal{P}(\lambda_B|\lambda_n)$$

For the flux factor calculations, a total of 10 independent 1 ns molecular dynamics simulations were performed starting from reactant structures derived from each of 6 randomly selected seed trajectories generated as described above. The λ_A interface was set equal to the λ_1 interface at $\lambda = -0.8$. For the control flux factor computations (as illustrated in Figure S1A), the effective positive flux was computed as the number of times the trajectory crossed the $\lambda_A = -0.8$ interface, having come from the region below the λ_A interface, divided by the total amount of time spent below the λ_A interface. For the constrained test flux factor computations (as illustrated in Figure S1C), the top 10 LASSO-selected features at the t=0 time point were written out during the dynamics run, and the effective positive flux was computed as the number of times the trajectory crossed the $\lambda_1 = -0.8$ interface, having come from the region A', where region A' refers to all points in phase space which lie at the last trough (i.e., the first point at which $\frac{d\lambda}{dt} = 0$ and $\frac{d^2\lambda}{dt^2} > 0$) before crossing $\lambda_A = -0.8$, having first crossed $\lambda_0 = -1$, and for which the logistic classifier with coefficients and features listed in Table S2 evaluated to true. Derivatives of λ with respect to time were computed using finite differences.

For the probability factor calculations, a total of 29 $P(\lambda_{i+1}|\ \lambda_i)$ interface ensembles from each of the six seed trajectories were computed, with the λ_i interfaces spaced between $\lambda = -0.8$ and $\lambda = 0$. The placement of these interfaces relative to the potential of mean force surface used to generate initial seed is shown in Figure S4. To ensure sufficient sampling, interfaces between $\lambda = -0.8$ and $\lambda = -0.15$ were spaced in 0.025-Å increments and the remaining interfaces between −0.15 and 0 spaced in 0.05-Å increments. For each interface ensemble, a total of 5000 shooting moves was attempted. In each λ_i ensemble, candidate trajectories were generated using full shooting moves and accepted if they both crossed the $\lambda_A = -0.8$ interface and crossed the $\lambda = \lambda_i$ interface having first come from crossing interface λ_A. For the unconstrained control ensembles (Figure S1B), no further acceptance rules were applied.

For constrained ensembles (Figure S1D), once the trajectory connected the $\lambda_A = -0.8$ and λ_{i+1} interfaces, the trajectory was only included in the ensemble if the logistic classifier evaluated with features and coefficients in Table S2 evaluated to true at the first point at which $\frac{d\lambda}{dt} = 0$ and $\frac{d^2\lambda}{dt^2} > 0$ before crossing $\lambda_A = -0.8$, having first crossed interface $\lambda_0 = -1$.

Integration was stopped when the candidate trajectories crossed their respective $\lambda = \lambda_{i+1}$ interface or the λ_0 interface, which was accomplished by modifying the RXNCOR module of CHARMM41.[34,35] All shooting moves and acceptance criteria were implemented using a MATLAB wrapper around CHARMM41 (i.e., CHARMM was only used for the actual molecular dynamics integration). The number of accepted trajectories varied between the interface ensembles, seed trajectories, and whether or not the additional sampling constraint was applied, ranging between 10 and 95%.

Figure S1: (A) Illustration of computation of the TIS flux factor. The red and gray line represents a long molecular dynamics trajectory originating in region A. Portions of the trajectory in red indicate the time points in region A used to normalize the flux factor. Black dots represent effective crossings of the λ_A interface. (B) Illustration of computation of a $P(\lambda_{i+1}|\ \lambda_i)$ ensemble. Each red and white line indicates an attempted shooting move. Black dots indicate shooting points. Red lines indicate accepted shooting moves, while white lines indicate rejected shooting moves. (C) Illustration of procedure used to compute the constrained flux factor. The dark red region indicates the reactive subregion A' identified using machine learning. Portions of the trajectory in red indicate the time point in either region A' used to compute the constrained flux factor. Black dots represent effective crossings of the λ_A interface. (D) Illustration of a constrained $P(\lambda_{i+1}|\ \lambda_i)$ ensemble. The dark red region indicates the reactive subregion A' identified using machine learning.

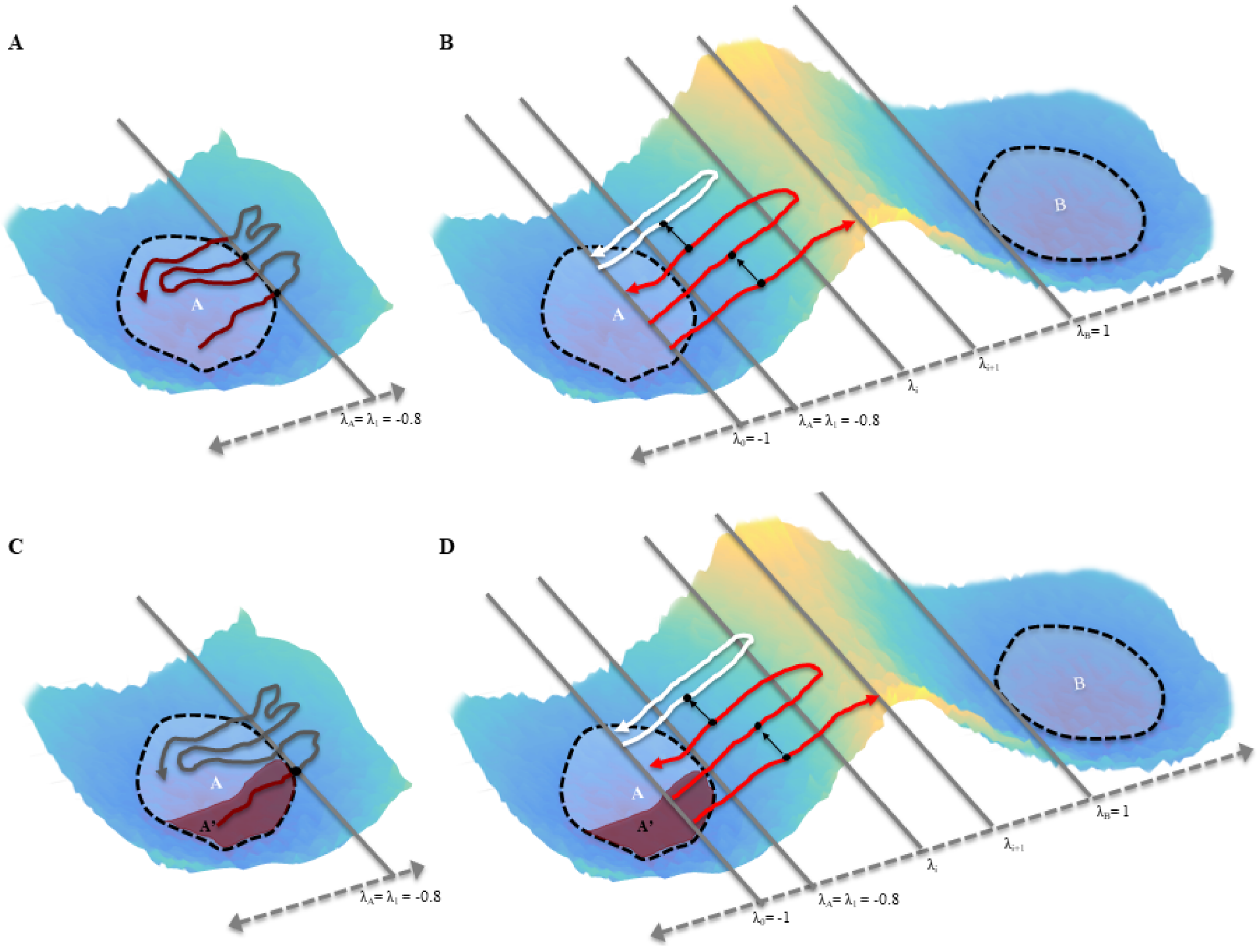


Table S1. Feature names, feature indices, and feature types computed at each time point. Residue name AC6 refers to the substrate, residue name NDP refers to the NADPH cofactor, and the residue name MG6 refers to the 5 active site waters and two magnesium ions. Structural representations of features are shown in Figure S2.

Feature Index	Feature Name	Feature Type
1	Dist AC6/O2,NDP/N7N	Substrate-environment
2	Dist AC6/O2,NDP/O7N	Substrate-environment
3	Dist AC6/O3,MG6/H24	Substrate-environment
4	Dist AC6/O6,MG6/M16	Substrate-environment
5	Dist AC6/O8,GLU496/Hε2	Substrate-environment
6	Dist AC6/O8,MG6/M17	Substrate-environment
7	Dist GLU319/Oε1,AC6/C5	Substrate-environment
8	Dist MG6/H25,AC6/O6	Substrate-environment
9	Dist MG6/H26,AC6/O6	Substrate-environment
10	Dist MG6/H27,AC6/O6	Substrate-environment
11	Dist MG6/H28,AC6/O6	Substrate-environment
12	Dist MG6/H31,AC6/O6	Substrate-environment
13	Dist MG6/H32,AC6/O6	Substrate-environment
14	Dist MG6/M16,AC6/O3	Substrate-environment
15	Dist MG6/M17,AC6/O6	Substrate-environment
16	Dist NDP/H4N2,AC6/C4	Substrate-environment
17	Ang AC6/O6,MG6/M16,AC6/O3	Substrate-environment
18	Ang AC6/O8,MG6/M17,AC6/O6	Substrate-environment
19	Ang MG6/M17,AC6/O6,MG6/M16	Substrate-environment
20	Dist AC6/C1,AC6/C4	Intra-substrate
21	Dist AC6/C1,AC6/O2	Intra-substrate
22	Dist AC6/C1,AC6/O3	Intra-substrate
23	Dist AC6/C4,AC6/C7	Intra-substrate
24	Dist AC6/C4,AC6/O6	Intra-substrate
25	Dist AC6/C5,AC6/C4	Intra-substrate
26	Dist AC6/C5,AC6/C7	Intra-substrate
27	Dist AC6/C7,AC6/C9	Intra-substrate
28	Dist AC6/C7,AC6/O8	Intra-substrate
29	Ang AC6/C1,AC6/C4,AC6/C7	Intra-substrate
30	Ang AC6/C4,AC6/C7,AC6/C5	Intra-substrate
31	Ang AC6/C4,AC6/C7,AC6/C9	Intra-substrate
32	Ang AC6/C5,AC6/C4,AC6/C1	Intra-substrate
33	Ang AC6/C5,AC6/C7,AC6/C9	Intra-substrate
34	Dihe AC6/C1,AC6/C5,AC6/C7,AC6/C4	Intra-substrate
35	Dihe AC6/C5,AC6/C4,AC6/C7,AC6/C9	Intra-substrate
36	Dist NDP/H4N2,NDP/C4N	Intra-cofactor
37	Dist NDP/N7N,NDP/O2N	Intra-cofactor
38	Ang NDP/C4N,NDP/N1N,NDP/C1NQ	Intra-cofactor
39	Ang NDP/C6N,NDP/C3N,NDP/C7N	Intra-cofactor
40	Ang NDP/N7N,NDP/H72N,NDP/O2N	Intra-cofactor
41	Dihe NDP/C2N,NDP/C3N,NDP/C7N,NDP/N7N	Intra-cofactor
42	Dihe NDP/C2NQ,NDP/C1NQ,NDP/N1N,NDP/C6N	Intra-cofactor
43	Dihe NDP/C4N,NDP/C3N,NDP/C7N,NDP/O7N	Intra-cofactor
44	Dihe NDP/H1NQ,NDP/C1NQ,NDP/N1N,NDP/C2N	Intra-cofactor
45	Dist MG6/O18,MG6/M17	Water-metal
46	Dist MG6/O19,MG6/M17	Water-metal
47	Dist MG6/O20,MG6/M17	Water-metal
48	Dist MG6/O21,MG6/M16	Water-metal
49	Dist MG6/O22,MG6/M16	Water-metal
50	Ang MG6/H23,MG6/O22,MG6/M16	Water-metal
51	Ang MG6/H24,MG6/O22,MG6/M16	Water-metal
52	Ang MG6/H25,MG6/O21,MG6/M16	Water-metal
53	Ang MG6/H26,MG6/O21,MG6/M16	Water-metal
54	Ang MG6/H27,MG6/O20,MG6/M17	Water-metal
55	Ang MG6/H28,MG6/O20,MG6/M17	Water-metal
56	Ang MG6/H29,MG6/O18,MG6/M17	Water-metal
57	Ang MG6/H30,MG6/O18,MG6/M17	Water-metal
58	Ang MG6/H31,MG6/O19,MG6/M17	Water-metal
59	Ang MG6/H32,MG6/O19,MG6/M17	Water-metal
60	Dist GLU496/Oε2,GLU496/Hε2	Other environment
61	Dist GLN136/Nε2,NDP/O7N	Other environment
62	Dist MG6/H25,MG6/O21	Other environment
63	Dist MG6/H26,MG6/O21	Other environment
64	Dist MG6/H27,MG6/O20	Other environment
65	Dist MG6/H28,MG6/O20	Other environment
66	Dist MG6/H31,MG6/O19	Other environment
67	Dist MG6/H32,MG6/O19	Other environment
68	Ang GL136/Nε2,GLN136/Hε22,NDP/O7N	Other environment

Figure S2: Structural representation of (A) distances computed, (B) angles computed, and (C) dihedrals computed at each time point. Numbering of features corresponds to that of Table S1. Coloring of features corresponds to the feature type with red indicating substrate-environment interactions, orange indicating intrasubstrate conformations, blue indicating intra-cofactor conformations, green indicating water-metal interactions and gold indicating other environment interactions.

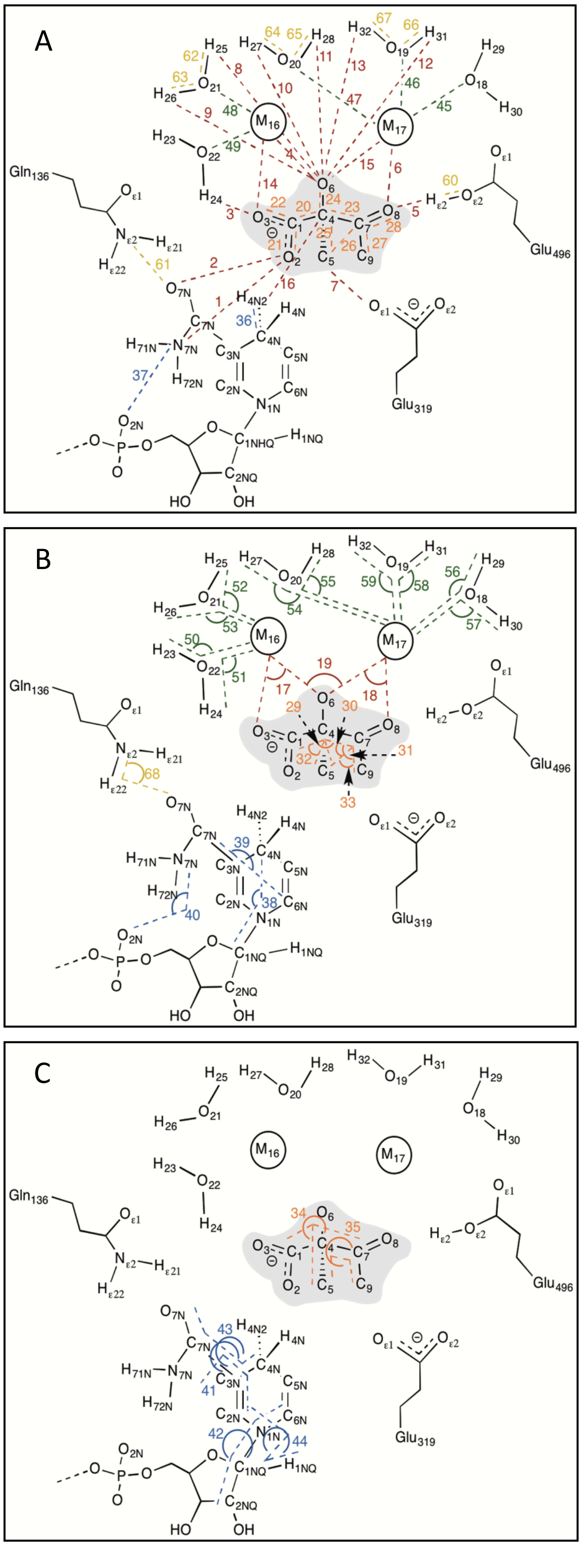


Figure S3: Illustration of both KARI homodimer subunits (PDB ID: 1YVE), with active-site residues Asp 315, Glu 319, Glu 496, bound transition state analog N-hydroxy-N-isopropyloxamate and NADPH cofactor shown as sticks to indicate active-site separation and to support the choice of using a single subunit in simulations.

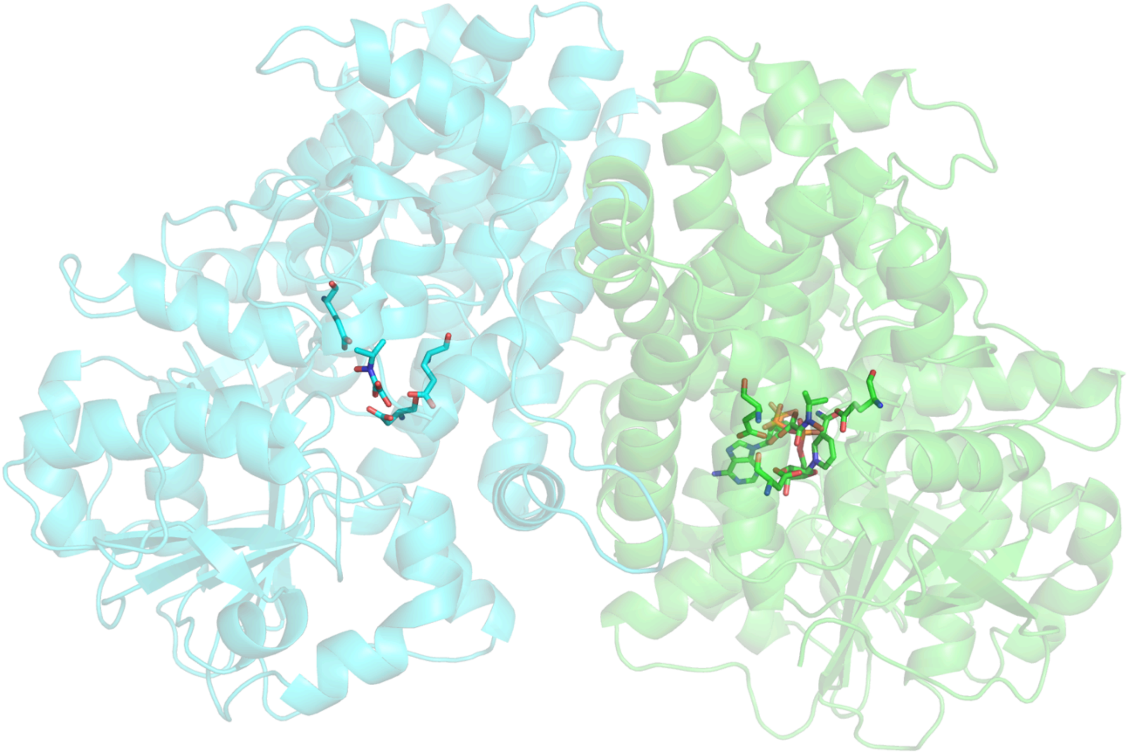


Table S2: Top 10 LASSO selected features at 0 fs time point and coefficients β_j used to define reactive region A’ in constrained TIS simulations. Note that classification was performed on the fly through the TIS Markov chain and thus features were not normalized by Z-scores, so non-standardized coefficients β_j are reported. The bias β_0 used was −18.603.

j	Feature	β_j
1	Distance GLU`319/Oε1,AC6/C5	2.1944
2	Distance MG6/M16,AC6/O3	−12.093
3	Distance AC6/C1,AC6/C4	13.447
4	Distance AC6/C4,AC6/O6	20.561
5	Angle NDP/C4N,NDP/N1N,NDP/C1NQ	−2.8234
6	Distance AC6/O8,GLU`496/Hε2	−3.4298
7	Distance AC6/C5,AC6/C4	−8.8403
8	Distance AC6/O8,MG6/M17	8.8193
9	Dihedral AC6/C5,AC6/C4,AC6/C7,AC6/C9	−3.7307
10	Distance MG6/H28,AC6/O6	−0.5615

Figure S4: Placement of interfaces used in TIS probability factor calculations superimposed onto the potential of mean force surface used to generate initial seed trajectories. Key interfaces λ_0=-1, λ_A = -0.8 and λ_B = 1 are labeled.

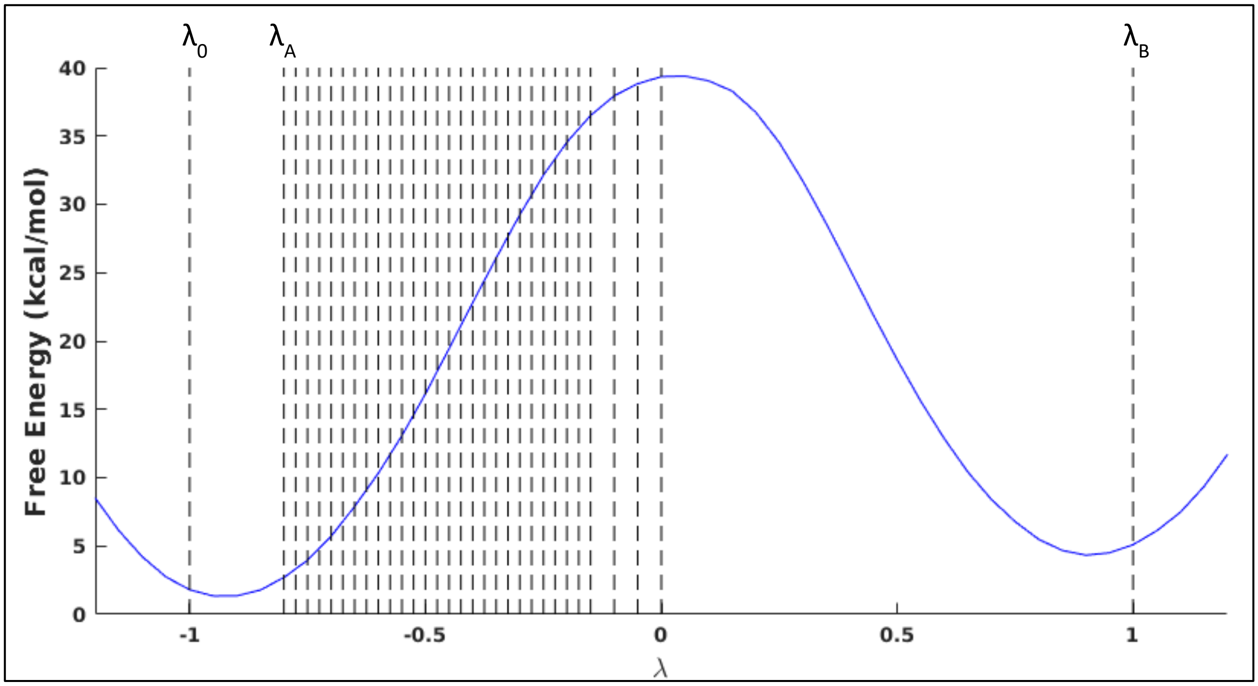


Table S3: Top 30 consensus features for the –150 to 0 fs time window. Feature rank indicates ranking according to the number of occurrences in the 20 LASSO-selected feature sets.

Rank	Feature Name	Feature Type	Occ.
1	Dist GLU319/Oε1,AC6/C5	Substrate-environment	24
2	Dist MG6/H32,AC6/O6	Substrate-environment	23
3	Dist MG6/H26,AC6/O6	Substrate-environment	22
4	Ang MG6/H31,MG6/O19,MG6/M17	Water-metal	20
5	Dist MG6/H28,AC6/O6	Substrate-environment	19
6	Dist AC6/O8,GLU496/Hε2	Substrate-environment	19
7	Ang NDP/C4N,NDP/N1N,NDP/C1NQ	Intra-cofactor	19
8	Dist MG6/H27,AC6/O6	Substrate-environment	18
9	Dist AC6/C5,AC6/C4	Intra-substrate	18
10	Ang MG6/M17,AC6/O6,MG6/M16	Substrate-environment	18
11	Dihe AC6/C5,AC6/C4,AC6/C7,AC6/C9	Intra-substrate	17
12	Ang AC6/O6,MG6/M16,AC6/O3	Substrate-environment	17
13	Dist MG6/O20,MG6/M17	Water-metal	17
14	Ang AC6/C1,AC6/C4,AC6/C7	Intra-substrate	16
15	Dist AC6/C7,AC6/C9	Intra-substrate	16
16	Ang AC6/O8,MG6/M17,AC6/O6	Substrate-environment	16
17	Ang GLN136/Nε2,GLN136/Hε22,NDP/O7N	Other environment	16
18	Ang AC6/C5,AC6/C7,AC6/C9	Intra-substrate	15
19	Dist MG6/M17,AC6/O6	Substrate-environment	15
20	Ang MG6/H29,MG6/O18,MG6/M17	Water-metal	15
21	Dist GLN136/Nε2,NDP/O7N	Other environment	15
22	Dist MG6/H31,AC6/O6	Substrate-environment	13
23	Dist AC6/C4,AC6/C7	Intra-substrate	13
24	Dist AC6/O6,MG6/M16	Substrate-environment	13
25	Dist AC6/C1,AC6/C4	Intra-substrate	12
26	Ang MG6/H32,MG6/O19,MG6/M17	Water-metal	12
27	Dist MG6/M16,AC6/O3	Substrate-environment	10
28	Dist MG6/O19,MG6/M17	Water-metal	10
29	Ang MG6/H23,MG6/O22,MG6/M16	Water-metal	10
30	Ang MG6/H25,MG6/O21,MG6/M16	Water-metal	10

Figure S5: Structural representations of top 30 most consistently predictive (A) distances and (B) angles and dihedrals during the –150 to 0 fs time window. Labeling of features corresponds to ranking in Table S3. Coloring of features corresponds to the feature type with red indicating substrate-environment interactions, orange indicating intra-substrate conformations, blue indicating intra-cofactor conformations, green indicating water-metal interactions and gold indicating other environment interactions.

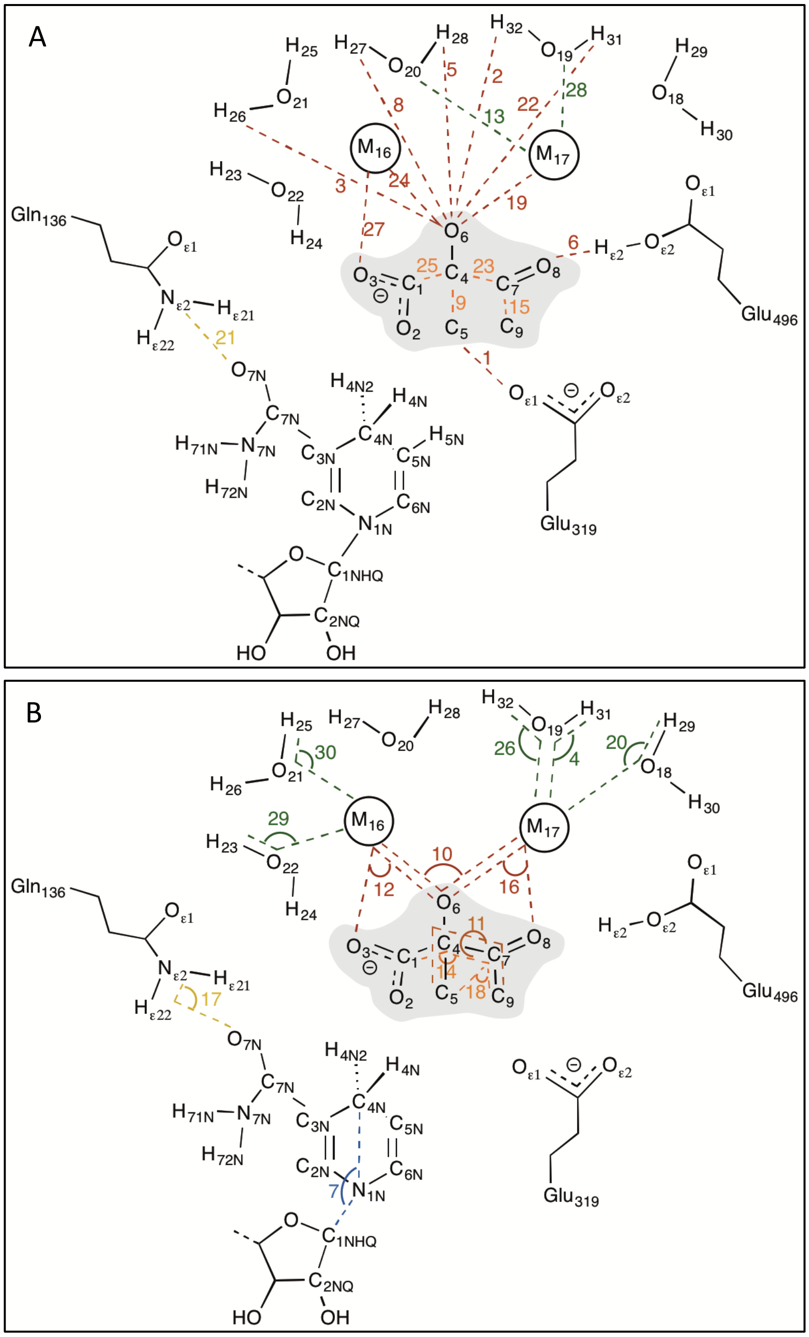
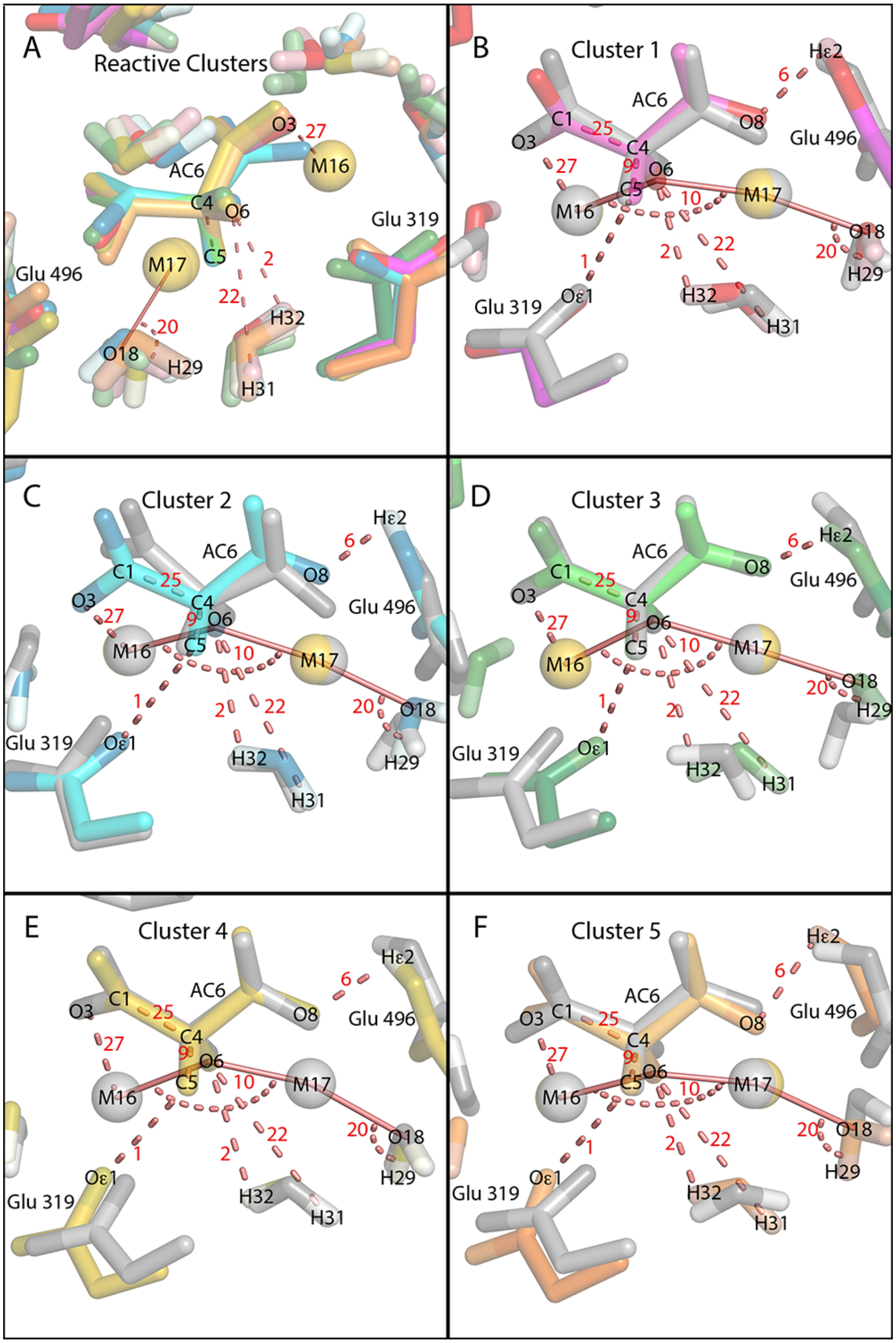

A
H_{25} H_{27} O_{20} H_{28} H_{32} O_{19} H_{31} H_{29}
O_{21} H_{26} O_{18} H_{30}
M_{16} M_{17}
H_{23} O_{22} H_{24}
Gln_{136} $O_{\varepsilon1}$ $N_{\varepsilon2}$ $H_{\varepsilon21}$ $H_{\varepsilon22}$
O_6 O_3 C_1 C_4 C_7 O_8 O_2 C_5 C_9 ⊖
$H_{\varepsilon2}$ $O_{\varepsilon2}$ $O_{\varepsilon1}$ Glu_{496}
$O_{\varepsilon1}$ ⊖ $O_{\varepsilon2}$ Glu_{319}
O_{7N} C_{7N} H_{71N} N_{7N} H_{72N} H_{4N2} H_{4N} C_{4N} H_{5N} C_{3N} C_{5N} C_{2N} C_{6N} N_{1N}
O C_{1NHQ} C_{2NQ} HO OH
1 2 3 5 6 8 9 13 15 19 21 22 23 24 25 27 28
B
H_{25} H_{27} O_{20} H_{28} H_{32} O_{19} H_{31} H_{29}
O_{21} H_{26} O_{18} H_{30}
M_{16} M_{17}
H_{23} O_{22} H_{24}
Gln_{136} $O_{\varepsilon1}$ $N_{\varepsilon2}$ $H_{\varepsilon21}$ $H_{\varepsilon22}$
O_6 O_3 C_1 C_4 C_7 O_8 O_2 C_5 C_9 ⊖
$H_{\varepsilon2}$ $O_{\varepsilon2}$ $O_{\varepsilon1}$ Glu_{496}
$O_{\varepsilon1}$ ⊖ $O_{\varepsilon2}$ Glu_{319}
O_{7N} C_{7N} H_{71N} N_{7N} H_{72N} H_{4N2} H_{4N} C_{4N} C_{3N} C_{5N} C_{2N} C_{6N} N_{1N}
O C_{1NHQ} C_{2NQ} HO OH
4 7 10 11 12 14 16 17 18 20 26 29 30

Table S4: Mean standardized logistic regression coefficients fit to classifier trained using the top 30 most consistently predictive features between –150 and 0 fs (listed in Table S3 and illustrated structurally in Figure S5) at the –150, –100, –50 and 0 fs time points relative to the last trough in the order parameter prior to the prospective catalytic event. Coefficients shown represent the mean values across 5 cross-validation partitions.

Standardized Regression Coefficient	Time Before Last Trough			
	-150 fs	-100 fs	-50 fs	0 fs
β_0	-0.059	-0.195	-0.269	-0.094
β_1	-0.361	-0.527	0.354	0.470
β_2	-0.198	-0.569	-0.374	0.874
β_3	-0.303	-0.957	-0.497	-0.035
β_4	0.615	0.706	0.117	0.453
β_5	0.094	0.069	0.401	-0.477
β_6	-0.365	-0.265	-0.147	-0.613
β_7	0.273	-0.251	-0.423	-0.397
β_8	0.293	-0.428	-1.134	-0.356
β_9	0.068	0.446	0.533	-1.030
β_{10}	0.318	-1.060	-0.666	-0.025
β_{11}	-0.307	0.058	-1.379	-0.289
β_{12}	-0.723	0.414	0.179	-0.510
β_{13}	0.236	-0.129	0.610	0.050
β_{14}	-0.256	0.214	-0.348	-0.107
β_{15}	-0.132	-0.460	-0.227	0.269
β_{16}	-0.330	-0.237	1.049	0.106
β_{17}	0.065	0.302	0.137	0.039
β_{18}	0.193	-0.704	0.665	0.026
β_{19}	-0.426	0.252	-0.425	0.007
β_{20}	0.033	0.141	0.477	0.704
β_{21}	0.319	-0.327	-0.471	-0.013
β_{22}	-0.135	-0.630	-0.162	-1.100
β_{23}	0.790	0.281	-0.089	0.200
β_{24}	-0.048	-0.179	-0.127	-0.014
β_{25}	0.083	-0.047	-0.182	0.504
β_{26}	0.592	0.592	0.434	-0.244
β_{27}	0.142	0.093	0.241	-0.944
β_{28}	-0.208	0.477	0.437	-0.083
β_{29}	-0.148	-0.370	-0.327	-0.061
β_{30}	0.183	0.151	0.338	-0.174

Figure S6: Representative structures for the reactive cluster and corresponding almost-reactive clusters described in Figure 3B–E. Feature numbering corresponds to that of Table S3. (A) Representative structures from all five reactive clusters. Representative structures from (B) cluster 1, (C) cluster 2, (D) cluster 3, (E) cluster 4, (F) cluster (5) and their corresponding almost-reactive clusters, respectively. In all panels, magenta corresponds to cluster 1, cyan corresponds to cluster 2, green corresponds to cluster 3, yellow corresponds to cluster 4, orange corresponds to cluster 5 and gray corresponds to the corresponding almost-reactive cluster for the reactive cluster shown in each histogram. In all panels, structures were aligned to minimize the root mean square difference between the two magnesium centers.



Supplementary Commentary on Figure 3E and Figure S6:

Figure S6 shows a number of interesting structural variations, particularly considering that they represent the final bond compression of reactive trajectories. The five water molecules that coordinate to one or the other magnesium ion each show significant variability across the clusters. Some of these are differences in water molecule positioning are represented in the features significant for cluster identify (e.g, features 2 and 22) and others in those significant for reactivity within a cluster (e.g., feature 20). Additionally, there is substantial variability in the internal substrate conformation across the different reactive clusters, with cluster 2 being an especially unusual outlier.

Illustrated in Figure 3E (column 5; feature 27), clusters 1, 2, and 5 exhibit significantly shorter values for the distance MG6/M16–AC6/O3 for the reactive than the almost-reactive trajectories. Structurally, Figure S6B, C, and F show that this corresponds to a different conformation of the substrate carboxylate group and a different engagement of magnesium ion M16 between reactive and almost-reactive trajectories. This shorter distance corresponds to a somewhat different orientation for the entire substrate relative to the two magnesium ions that also affects substrate hydroxyl O6 and the metal coordination environment. By contrast, clusters 3 and 4 show much less difference in the distribution of MG6/M16–AC6/O3 (feature 27) between reactive and almost-reactive sets (Figure 3E) and this can also be seen structurally in Figure S6D and E.

Also illustrated in Figure 3E (column 2; feature 9), all five clusters show that the length of the breaking bond, AC6/C5–AC6/C4, spans a wider range of values for the nearly-reactive trajectories and is on the shorter side of that distribution for the reactive ones. Keeping in mind that these conformations are for the 0 fs time point, when the bond is fully compressed before launching toward the barrier, this represents the notion that reactive trajectories require substantial potential energy by stored in the bond that is not always seen for almost-reactive trajectories (that is, this extra compression is necessary but not sufficient).

Figure 3E indicates that the adjacent substrate bond, AC6/C1–AC6/C4 (column 4; feature 25), is distributed somewhat longer in reactive than almost-reactive trajectories for clusters 2, 3, and 5; examining the corresponding structures in Figure S6C, D, and F doesn’t show a clear effect of this on conformation. Figure 3E also indicates that a water molecule orientation, angle MG6/H29–MG6/O18–MG6/M17 (column 3; feature 20), is distributed substantially larger for reactive than almost-reactive trajectories in cluster 5, and much more so than in any of the other clusters. Figure S6F seems to indicate that this allows engagement of a lone pair from O18 to interact much more favorably with magnesium, and perhaps affect the polarization of the substrate, in a typical reactive rather than almost-reactive trajectory. Some of the other clusters appear to show a difference in the interaction between that water molecule and magnesium ion, although it may not show up in the angle indicated. Finally, Figure 3E indicates that the distance from Glu 319 Oε1 to the substrate’s migrating methyl group C5 (column 1; feature 1) is distributed longer in reactive than almost-reactive trajectories for clusters 1, 2, and 5 (and partially for clusters 3 and 4). Figure S6B–F indicates the interaction, but it is unclear how much of it is steric (the Glu side chain must be far enough away from the methyl at compression to be adequately poised to push it toward product upon bond expansion) and how much is stabilizing of the methyl during the transition.

Figure S7. (A) Cumulative log(P) for increasing interface placement for each of the 5 seed trajectories tested. Red lines indicate trajectories sampled with the reactant basin constrained to only include the region where the 10 feature classifier evaluated to true. Blue lines indicate unconstrained control simulations. (B) Individual values of P(λ$_{i+1}$| λ$_i$) for each λ$_i$ ensemble computed. Error bars correspond to two standard errors of the mean across three independent Markov chains at each λ$_i$ ensemble. Red bars indicate test simulations, while blue bars indicate unconstrained control simulations.

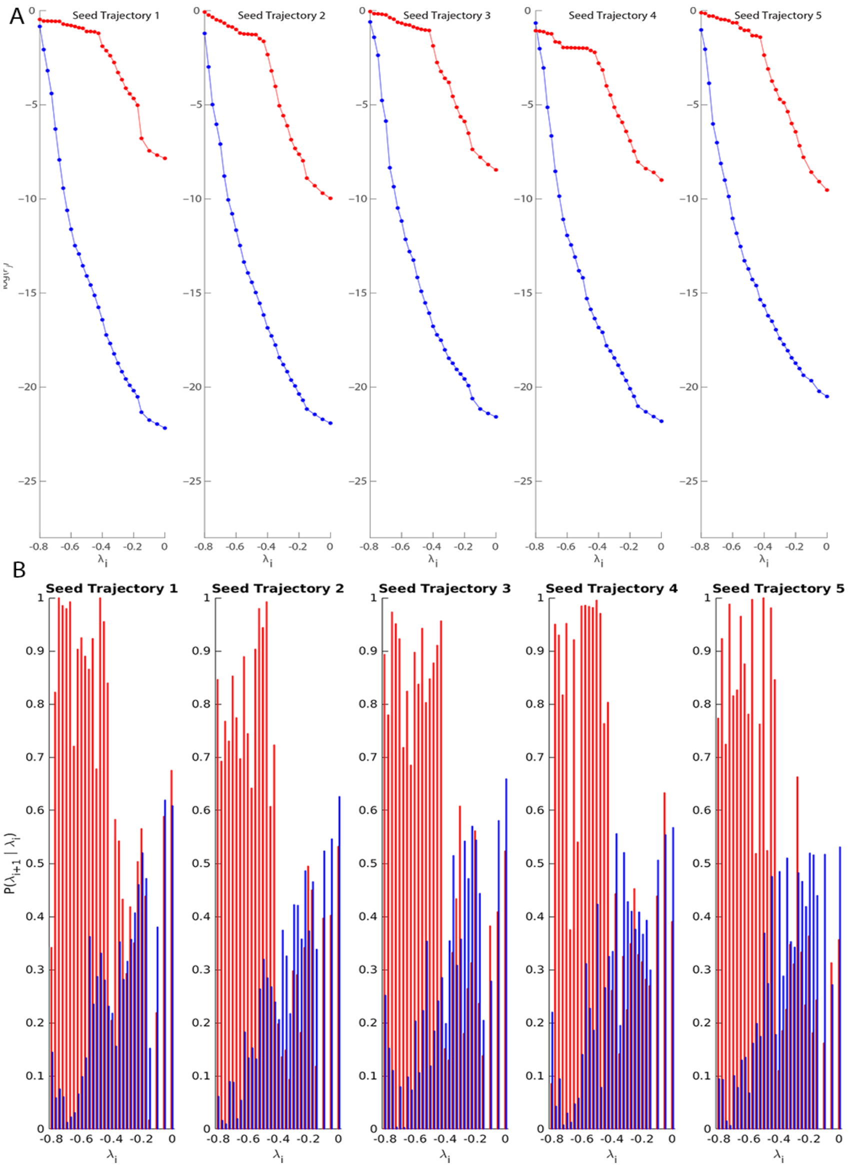
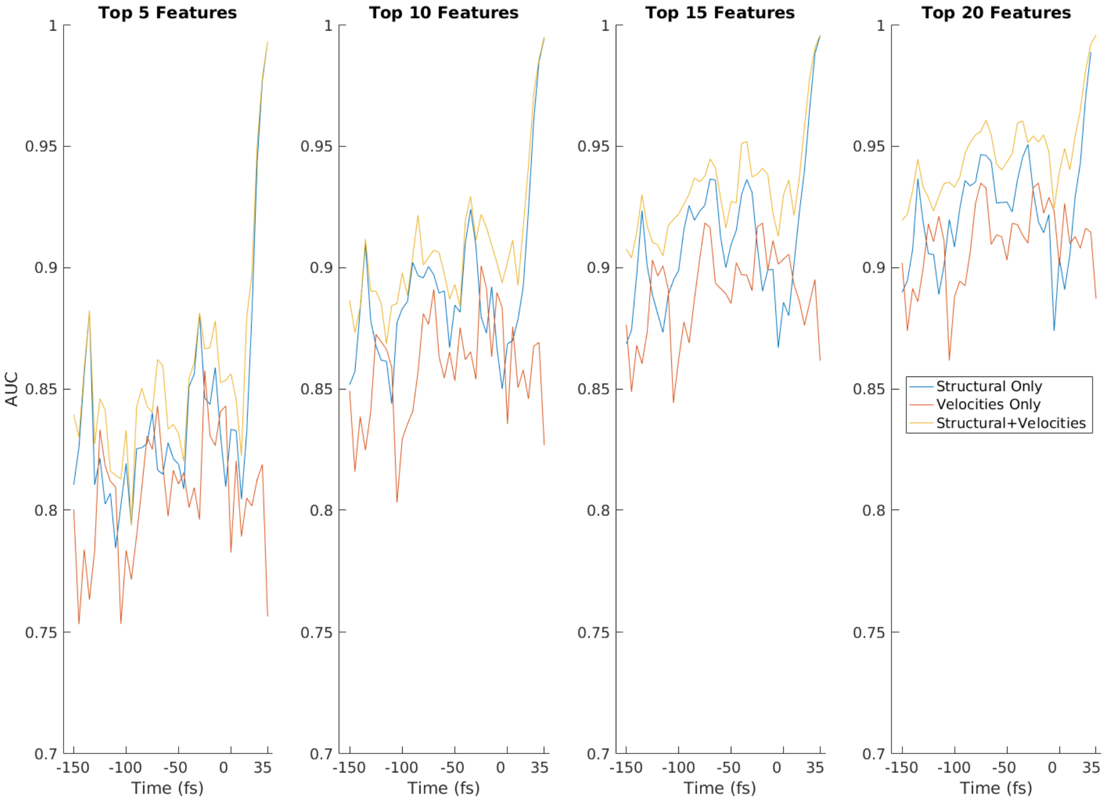


Table S5: Top 20 atomic velocity magnitudes at the 0 fs time point ranked by individual AUC. The feature set was comprised of the velocity magnitudes of the 341 atoms within 5 angstroms of the migrating methyl, AC6/C5 (including all hydrogens). Note that of these 341 velocities, only 17 exhibited individual AUCs greater than 0.60 at the “last trough” of the prelaunch window, and of these 17, 5 involved atoms included in the "consensus set”.

Rank	Atom Name	AUC	Involved in Consensus Feature Set?
1	AC6/O6	0.854	Yes
2	AC6/C4	0.738	Yes
3	Glu 319/Oε1	0.6713	Yes
4	AC6/H12	0.6687	No
5	NDP/P2A	0.6435	No
6	NDP/H2A	0.6406	No
7	Met 254/C	0.637	No
8	Thr 520/HN	0.6363	No
9	AC6/C7	0.6213	Yes
10	Glu 319/C	0.6187	No
11	Gln 136/CA	0.6154	No
12	NDP 600/O3	0.6082	No
13	Glu 319/Oε2	0.6077	No
14	Glu 319/O	0.6034	No
15	Glu 496/Oε2	0.6022	No
16	AC6/C1	0.6022	Yes
17	Lys 252 /HG1	0.602	No
18	Pro 251/O	0.5991	No
19	NDP 600/H1NQ	0.5984	No
20	AC6/O8	0.5953	Yes

Figure S8: Comparison of AUCs versus reaction progress for top LASSO-selected features from set consisting of (a) the 68 structural descriptors listed in Table S1 only, (b) velocity magnitudes of the 341 atoms within 5 Å of the migrating methyl or (c) the combined structural/velocity feature set.



REFERENCES

[1] Basner, J. E.; Schwartz, S. D. How enzyme dynamics helps catalyze a reaction in atomic detail: A transition path sampling study. *J. Am. Chem. Soc.* **2005**, *127* (40), 13822.

[2] Ruscio, J. Z.; Kohn, J. E.; Ball, K. A.; Head-Gordon, T. The influence of protein dynamics on the success of computational enzyme design. *J. Am. Chem. Soc.* **2009**, *131* (39), 14111.

[3] Kamerlin, S. C. L.; Warshel, A. At the dawn of the 21st century: Is dynamics the missing link for understanding enzyme catalysis? *Proteins* **2010**, *78* (6), 1339.

[4] Porter, J. L.; Rusli, R. A.; Ollis, D. L. Directed evolution of enzymes for industrial biocatalysis. *ChemBioChem* **2016**, *17* (3), 197.

[5] Hammer, S. C.; Knight, A. M.; Arnold, F. H. Design and evolution of enzymes for non-natural chemistry. *Curr. Opin. Green Sustain. Chem.* **2017**, *7* (Supplement C), 23.

[6] Molina-Espeja, P.; Viña-Gonzalez, J.; Gomez-Fernandez, B. J.; Martin-Diaz, J.; Garcia-Ruiz, E.; Alcalde, M. Beyond the outer limits of nature by directed evolution. *Biotechnol. Adv.* **2016**, *34* (5), 754.

[7] Lerner, R. A.; Benkovic, S. J.; Schultz, P. G. At the crossroads of chemistry and immunology: Catalytic antibodies. *Science* **1991**, *252* (5006), 659.

[8] Nevinsky, G. A.; Buneva, V. N. Natural catalytic antibodies – Abzymes. In *Catalytic Antibodies*; Keinan, E., Ed.; Wiley-VCH Verlag GmbH & Co. KGaA: **2004**; pp 505–569.

[9] Maeda, Y.; Makhlynets, O. V.; Matsui, H.; Korendovych, I. V. Design of catalytic peptides and proteins through rational and combinatorial approaches. *Annu. Rev. Biomed. Eng.* **2016**, *18* (1), 311.

[10] Kiss, G.; Çelebi-Ölçüm, N.; Moretti, R.; Baker, D.; Houk, K. N. Computational enzyme design**.** *Angew. Chem. Int. Ed.* **2013**, *52* (22), 5700.

[11] Baker, D. An exciting but challenging road ahead for computational enzyme design**.** *Protein Sci.* **2010**, *19* (10), 1817.

[12] Silver, N. W. Ensemble methods in computational protein and ligand design: Applications to the Fc[gamma] immunoglobulin, HIV-1 protease, and ketol-acid reductoisomerase system. Doctoral Dissertation, Massachusetts Institute of Technology, **2011**.

[13] 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.* **2003**, *100* (21), 12015.

[14] Sadiq, S. K.; Coveney, P. V. Computing the role of near attack conformations in an enzyme-catalyzed nucleophilic bimolecular reaction**.** *J. Chem. Theory Comput.* **2015**, *11* (1), 316.

[15] Zhang, J.; Zhang, Z.; Yang, Y. I.; Liu, S.; Yang, L.; Gao, Y. Q. Rich dynamics underlying solution reactions revealed by sampling and data mining of reactive trajectories**.** *ACS Cent. Sci.* **2017**, *3* (5), 407.

[16] van Erp, T. S.; Moqadam, M.; Riccardi, E.; Lervik, A. Analyzing complex reaction mechanisms using path sampling**.** *J. Chem. Theory Comput.* **2016**, *12* (11), 5398.

[17] Lau, E. Y.; Bruice, T. C. Importance of correlated motions in forming highly reactive near attack conformations in catechol *O*-methyltransferase. *J. Am. Chem. Soc.* **1998**, *120* (48), 12387.

[18] Bruice, T. C.; Lightstone, F. C. Ground state and transition state contributions to the rates of intramolecular and enzymatic reactions**.** *Acc. Chem. Res.* **1999**, *32* (2), 127.

[19] Bruice, T. C. A view at the millennium: The efficiency of enzymatic catalysis**.** *Acc. Chem. Res.* **2002**, *35* (3), 139.

[20] Dellago, C.; Bolhuis, P. G.; Csajka, F. S.; Chandler, D. Transition path sampling and the calculation of rate constants *J. Chem. Phys.* **1998**, *108* (5), 1964.

[21] van Erp, T. S.; Moroni, D.; Bolhuis, P. G. A novel path sampling method for the calculation of rate constants. *J. Chem. Phys.* **2003**, *118*, 7762.

[22] Dumas, R.; Biou, V.; Halgand, F.; Douce, R.; Duggleby, R. G. Enzymology, structure, and dynamics of acetohydroxy acid isomeroreductase. *Acc. Chem. Res.* **2001**, *34* (5), 399.

[23] Chen, C.-T.; Liao, J. C. Frontiers in microbial 1-butanol and isobutanol production**.** *FEMS Microbiol. Lett.* **2016**, *363* (5), fnw020.

[24] Bastian, S.; Liu, X.; Meyerowitz, J. T.; Snow, C. D.; Chen, M. M.; Arnold, F. H. Engineered ketol-acid reductoisomerase and alcohol dehydrogenase enable anaerobic 2-methylpropan-1-ol production at theoretical yield in Escherichia coli. *Metab. Eng.* **2011**, *13* (3), 345.

[25] Tadrowski, S.; Pedroso, M. M.; Sieber, V.; Larrabee, J. A.; Guddat, L. W.; Schenk, G. Metal ions play an essential catalytic role in the mechanism of ketol-acid reductoisomerase. *Chem.- Eur. J.* **2016**, *22* (22), 7427.

[26] Biou, V.; Dumas, R.; Cohen-Addad, C.; Douce, R.; Job, D.; Pebay-Peyroula, E. The crystal structure of plant acetohydroxy acid isomeroreductase complexed with NADPH, two magnesium ions and a herbicidal transition state analog determined at 1.65 Å resolution. *EMBO J.* **1997**, *16* (12), 3405.

[27] Bernstein, F. C.; Koetzle, T. F.; Williams, G. J.; Meyer, E. E. J.; Brice, M. D.; Rodgers, J. R.; Kennard, O.; Shimanouchi, T.; Tasumi, M. The Protein Data Bank: A computer-based archival file for macromolecular structures. *J. Mol. Biol.* **1977**, *112*, 535.

[28] Berman, H. M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T. N.; Weissig, H.; Shindyalov, I. N.; Bourne, P. E. The Protein Data Bank**.** *Nucleic Acids Res.* **2000**, *28* (1), 235.

[29] Rose, P. W.; Prlić, A.; Altunkaya, A.; Bi, C.; Bradley, A. R.; Christie, C. H.; Costanzo, L. D.; Duarte, J. M.; Dutta, S.; Feng, Z.; Green, R. K.; Goodsell, D. S.; Hudson, B.; Kalro, T.; Lowe, R.; Peisach, E.; Randle, C.; Rose, A. S.; Shao, C.; Tao, Y. P.; Valasatava, Y.; Voigt, M.; Westbrook, J. D.; Woo, J.; Yang, H.; Young, J. Y.; Zardecki, C.; Berman, H. M.; Burley, S. K. The RCSB protein data bank: Integrative view of protein, gene and 3D structural information**.** *Nucleic Acids Res.* **2017**, *45*, D271.

[30] Proust-De Martin, F.; Dumas, R.; Field, M. J. A hybrid-potential free-energy study of the isomerization step of the acetohydroxy acid isomeroreductase reaction**.** *J. Am. Chem. Soc.* **2000**, *122* 32, 7688.

[31] Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; Montgomery, J. A., Jr.; Vreven, T.; Kudin, K. N.; Burant, J. C.; Millam, J. M.; Iyengar, S. S.; Tomasi, J.; Barone, V.; Mennucci, B.; Cossi, M.; Scalmani, G.; Rega, N.; Petersson, G. A.; Nakatsuji, H.; Hada, M.; Ehara, M.; Toyota, K.; Fukuda, R.; Hasegawa, J.; Ishida, M.; Nakajima, T.; Honda, Y.; Kitao, O.; Nakai, H.; Klene, M.; Li, X., Knox, J. E.; Hratchian, H. P.; Cross, J. B.; Bakken, V.; Adamo, C.; Jaramillo, J.; Gomperts, R.; Stratmann, R. E.; Yazyev, O.; Austin, A. J.; Cammi, R.; Pomelli, C.; Ochterski, J. W.; Ayala, P. Y.; Morokuma, K.; Voth, G. A.; Salvador, P.; Dannenberg, J. J.; Zakrzewski, V. G.; Dapprich, S.; Daniels, A. D.; Strain, M. C.; Farkas, O.; Malick, D. K.; Rabuck, A. D.; Raghavachari, K.; Foresman, J. B.; Ortiz, J. V.; Cui, Q.; Baboul, A. G.; Clifford, S.; Cioslowski, J.; Stefanov, B. B.; Liu, G.; Liashenko, A.; Piskorz, P.; Komaromi, I.; Martin, R. L.; Fox, D. J.; Keith, T.; Al-Laham, M. A.; Peng, C. Y.; Nanayakkara, A.; Challacombe, M.; Gill, P. M. W.; Johnson, B.; Chen, W.; Wong, M. W.; Gonzalez, C.; Pople, J. A. *Gaussian 03: revision B.05*; Gaussian Inc.: Pittsburgh, PA, **2003**.

[32] Peng, C.; Schlegel, H. B. Combining synchronous transit and quasi-Newton methods to find transition states. *Isr. J. Chem.* **1993**, *33* (4), 449.

[33] Peng, C.; Ayala, P. Y.; Schlegel, H. B.; Frisch, M. J. Using redundant internal coordinates to optimize equilibrium geometries and transition states. *J. Comput. Chem.* **1996**, *17* (1), 49.

[34] Brooks, B. R.; Bruccoleri, R. E.; Olafson, B. D.; States, D. J.; Swaminathan, S.; Karplus, M. CHARMM: A program for macromolecular energy, minimization, and dynamics calculations. *J. Comput. Chem.* **1983**, *4* (2), 187.

[35] Brooks, B. R.; Brooks, C. L.; Mackerell, A. D.; Nilsson, L.; Petrella, R. J.; Roux, B.; Won, Y.; Archontis, G.; Bartels, C.; Boresch, S.; Caflisch, A.; Caves, L.; Cui, Q.; Dinner, A. R.; Feig, M.; Fischer, S.; Gao, J.; Hodoscek, M.; Im, W.; Kuczera, K.; Lazaridis, T.; Ma, J.; Ovchinnikov, V.; Paci, E.; Pastor, R. W.; Post, C. B.; Pu, J. Z.; Schaefer, M.; Tidor, B.; Venable, R. M.; Woodcock, H. L.; Wu, X.; Yang, W.; York, D. M.; Karplus, M. CHARMM: The biomolecular simulation program**.** *J. Comput. Chem.* **2009**, *30* (10), 1545.

[36] Dewar, M. J. S.; Zoebisch, E. G.; Healy, E. F.; Stewart, J. J. P. Development and use of quantum mechanical molecular models. 76. AM1: A new general purpose quantum mechanical molecular model**.** *J. Am. Chem. Soc.* **1985**, *107* (13), 3902.

[37] Huang, J.; MacKerell, A. D. CHARMM36 all-atom additive protein force field: Validation based on comparison to NMR data. *J. Comput. Chem.* **2013**, *34* (25), 2135.
[38] Stewart, J.P. Optimization of parameters for semiempirical methods IV: extension of MNDO, AM1, and PM3 to more main group elements
J. Mol. Model. **2004**, *10*, 155.
[39] Gao, J.; Amara, P.; Alhambra, C.; Field, M. A generalized hybrid orbital (GHO) method for the treatment of boundary atoms in combined QM/MM calculations**.** *J. Phys. Chem. A* **1998**, *102*, 4714.
[40] Kumar, S.; Rosenberg, J. M.; Djamal, B.; Swendsen, R. H.; Kollman, P. A. The weighted histogram analysis method for free-energy calculations on biomolecules. I. The method. *J. Comput. Chem.* **1992**, *13* (8).
[41] Tibshirani, R. Regression shrinkage and selection via the lasso**.** *J. R. Stat. Soc. Ser. B Stat. Methodol.* **1996**, *58*, 267.
[42] Warshel, A. Electrostatic origin of the catalytic power of enzymes and the role of preorganized active sites**.** *J. Biol. Chem.* **1998**, *273* (42), 27035.
[43] Kamerlin, S. C. L.; Sharma, P. K.; Chu, Z. T.; Warshel, A. Ketosteroid isomerase provides further support for the idea that enzymes work by electrostatic preorganization**.** *Proc. Natl. Acad. Sci. U. S. A.* **2010**, *107* (9), 4075.
[44] Tyagi, R.; Lee, Y.-T.; Guddat, L. W.; Duggleby R. G. Probing the mechanism of the bifunctional enzyme ketol-acid reductoisomerase by site-directed mutagenesis of the active site. *FEBS J.* **2005** 272 (2), 593.
[45] Zoi, I.; Suarez, J.; Antoniou, D.; Cameron, S. A.; Schramm, V. L.; Schwartz, S. D. Modulating enzyme catalysis through mutations designed to alter rapid protein dynamics. *J. Am. Chem. Soc.* **2016**, *138* (10), 3403.
[46] Harijan, R. K.; Zoi, I.; Antoniou, D.; Schwartz, S. D.; Schramm, V. L. Catalytic-site design for inverse heavy-enzyme isotope effects in human purine nucleoside phosphorylase**.** *Proc. Natl. Acad. Sci. U.S.A.* **2017**, *114* (25), 6456.