The Role of Side Chain Entropy and Mutual Information for Improving the De Novo Design of Kemp Eliminases KE07 and KE70

Asmit Bhowmick[1], Sudhir Sharma[2], Hallie Hhonma[3], and Teresa Head-Gordon[1,2,3,4*]

[1]Department of Chemical and Biomolecular Engineering, [2]Department of Chemistry, and

[3]Department of Bioengineering, University of California Berkeley

[4]Chemical Sciences Division, Lawrence Berkeley National Labs

Berkeley, California 94720, USA

Side chain entropy and mutual entropy information between residue pairs have been calculated for two *de novo* designed Kemp eliminase enzymes, KE07 and KE70, and for their most improved versions at the end of laboratory directed evolution (LDE). We find that entropy, not just enthalpy, helped to destabilize the preference for the reactant state complex of the designed enzyme as well as favoring stabilization of the transition state complex for the best LDE enzymes. Furthermore, residues with the highest side chain couplings as measured by mutual information, when experimentally mutated, were found to diminish or annihilate catalytic activity, some of which were far from the active site. In summary, our findings demonstrate how side chain fluctuations and their coupling can be an important design feature for *de novo* enzymes, and furthermore could be utilized in the computational steps in lieu of or in addition to the LDE steps in future enzyme design projects.

*Corresponding author:

Stanley 274

510-666-2744 (V)

thg@berkeley.edu

INTRODUCTION

The ability to control for protein structure, energetics and dynamical motions is a fundamental problem that limits our ability to rationally design catalysts for new chemical reactions not known to have a natural biocatalyst. Current computational approaches for *de novo* enzyme design seek to engineer a small catalytic construct into an accommodating protein scaffold, as exemplified by the Rosetta strategy applied to the design of many different catalytic motifs[1, 2]. In this study we consider the Rosetta design of the Kemp elimination reaction[3] involving the deprotonation of a small ligand substrate (5-nitro benzisoxazole) by a base (Figure 1a), in which the designed catalytic construct was engineered into a TIM barrel scaffold[2]. Two well-studied *de novo* enzymes for this reaction are KE07 and KE70, in which some minimal activity was observed in the designed enzymes and proved an important validation of the Rosetta approach. Nonetheless the catalytic activity was very low, and a number of follow-on studies have provided some important insight into the active site energetic features that limited the catalytic activity of the original designs of KE07 and KE70[4-7].

What proved more beneficial to improving the catalytic performance of KE07 and KE70 was application of laboratory directed evolution (LDE)[8-10], an experimental strategy based on the principle of natural selection[11]. The goal of LDE is to alter the protein sequence through multiple rounds of mutagenesis and selection to isolate the few new sequences that exhibit enhanced catalytic performance. Given the limitations of our understanding of the structure-function relationship[12], LDE provides an attractive alternative to rational design approaches to biocatalysis, is highly flexible in application to different biocatalysis reactions, and provides an effective way of improving upon *de novo* enzymes generated from computational designs[9, 13]. Although LDE can be an opaque process because it offers no direct rationale as to why mutations are successful, many hypotheses and useful heuristics have been proposed and developed for improving binding selectivity or protein stability using LDE[14-17]. For example, previous efforts to rationalize and ultimately decrease the sequence space for LDE focused on the interplay of sequence site entropy, i.e. the plasticity for evolutionary-driven substitutions, and the likelihood that these sites would thus be more prone to increased structural flexibility[18, 19], and which was borne out by mutations that reduced the entropy of these sites[20, 21]. For KE07 and KE70, LDE improved the Michaelis-Menten specificity constant k_{cat}/K_M by a factor of ~200 and ~400, respectively, in the best evolved enzymes.

The primary question we address in this work is what is missing in the original computational *de novo* design that is captured instead during the LDE process to improve the Michaelis-Menten specificity constant k_{cat}/K_M for KE07 and KE70? Using the framework of

transition state theory[22], biocatalytic improvements as measured by k_{cat}/K_M should arise through reduction in the activation free energy, $\Delta G_T^{\dagger} = \Delta G^{\dagger} + \Delta G_{EL}$, where E and L represent the enzyme and ligand, respectively. The activation free energy is comprised of a positive $\Delta G^{\dagger} = G_{EL^{\dagger}} - G_{EL}$ that quantifies the catalytic barrier between the reactant EL state and transition $EL^{\dagger}$ complex, and therefore relates directly to k_{cat}; in addition $\Delta G_{EL} = G_{EL} - G_{E+L}$ measures the binding affinity of the ligand to the enzyme active site and thus relates to K_M. Therefore knowing $\Delta G_T^{\dagger}$ or the activation enthalpy, $\Delta H_T^{\dagger}$, and activation entropy, $\Delta S_T^{\dagger}$, components, we can connect directly to the k_{cat}/K_M ratio through

$$\frac{k_{cat}}{K_M} = \frac{kT}{h}\exp\left(\frac{-\Delta G_T^{\dagger}}{RT}\right) = \frac{kT}{h}\exp\left(\frac{-\Delta H_T^{\dagger}}{RT}\right)\exp\left(\frac{\Delta S_T^{\dagger}}{R}\right) \quad (1)$$

and therefore the success of the LDE process applied to KE07 and KE70 must have a rational thermodynamic basis via Eq. (1).

While it is broadly accepted that optimizing enthalpic interactions is paramount for good substrate binding and lowering of the transition state barrier to the chemical reaction, the role of dynamics for improving catalytic performance is more controversial. One aspect of the controversy pertains to the definition of dynamics, for example whether it refers to equilibrium statistical fluctuations[23-25], dynamical coupling[26] and/or maximizing the reactive flux through the transition state surface[27]. Probably the most commonly implied definition of important functional motions for biocatalysis is a thermodynamic one, i.e. statistical fluctuations that are embodied in an entropy change that along with enthalpy contributes to the changes in the free energy state function as per Eq. (1).

In order to support the design of good enthalpic interactions between the substrate and the enzyme, it would seem desirable to impose some limits on the conformational flexibility to aid the catalytic function[28, 29]. A survey of 178 enzymes led to the conclusion that active site residues of naturally occurring enzymes are the least flexible within a sequence, supported by their low B-

factors in the crystalline environment[30]. At the same time, evidence also exists that increased conformational flexibility can also be a factor in improved biocatalytic performance. Room temperature X-ray crystallography[31], in good agreement with NMR[32, 33], has shown that protein interiors are very fluid, especially at the level of side chain motions, and that alternate side chain conformers in ligand binding and catalysis can be critical for function[34], and conformational flexibility forms the basis of computational approaches to conformational selection in allostery[35-37]. Hence, even though configurational entropy may well be important for biocatalysis, it still remains poorly understood how statistical fluctuations can be utilized to improve the *de novo* design process.

In this study we consider the question of how LDE improvements in the catalytic activity of KE07 and KE70 changes the active site energetics as well as side chain entropy and side chain coupling captured through mutual information. We find that the best KE07 and KE70 enzymes at the end of LDE process exhibit enthalpies *and* entropies that both destabilize the reactive state and stabilize the transition state with respect to the designed enzymes, showing that the original enzymes were over-designed for the EL reactant state, whereas the LDE process created enzymes that preferred the $EL^{\dagger}$ complex instead, especially for the KE70 enzyme. Furthermore, we find that residues with the highest mutual information proved to be critical for enzyme catalysis, which we tested on the best evolved enzyme for KE07. We show that new amino acid chemistries with high mutual information in the active site, some of which have not been reported in previous studies of the same enzyme, proved critical to function since experimental mutations at these sites destroyed enzyme activity. Of greater interest is that other residues identified as having high mutual information that are far from the active site were found to diminish or annihilate catalytic activity when mutated in the best evolved KE07 enzyme. In summary, our findings demonstrate how differences in not only energetics, but side chain fluctuations and their coupling, can be an important design feature for *de novo* enzymes, and furthermore could be utilized in future computational enzyme design projects.

TRANSITION STATE THEORY

We rely on the analysis of enzyme performance using transition state theory via Eq. (1)[22]. For the calculation of the enthalpy, we assume that the PV term is negligible such that it can be quantified using only potential energy calculations. We therefore calculate all protein-protein interactions for KE07 and KE70 using the generalized Amber force field, while the model for all protein and 5-nitro benzisoxazole interactions with aqueous solvent is based on our GB-HPMF implicit solvent model,

which has been well-validated in previous work[38,39]. We use electrostatic models of the 5-nitro benzisoxazole ligand in the reactant state and transition state based on partial charges as reported by Frushicheva and co-workers[6], and long molecular dynamics calculations have confirmed that the ligand charges in the two states are compatible and thus stable within the protein modeled using a classical force field. The state enthalpy is evaluated as an average across an ensemble of backbone conformations, each of which has a large ensemble of side chain packings, such that we define

$H = \langle H \rangle_{SC,BB}$ for a given state: the EL† complex, the EL complex, and apo state of the enzyme E.

The state entropy term defined in Eq. (1) can be further decomposed into sums over (i) contributions from the individual residues in the enzyme, as well as (ii) contributions from correlated motion between side chains of residues[40, 41], averaged over the backbone configurations

$$S \sim \sum_{i}^{N_{res}} \langle S^{(i)} \rangle_{SC,BB} - \sum_{i}^{N_{res}-1} \sum_{j=i+1}^{N_{res}} \langle I^{(i,j)} \rangle_{SC,BB} + ... \quad (2)$$

and similarly Eq. (2) can be used to define the entropy of EL†, EL, and E states. Thus, we see that the catalytic power of an enzyme as measured from k_{cat}/K_M, can ultimately be related to entropy contributions from individual residues, mutual information between residue pairs, or even higher order correlations, when defining the total entropy change.

COMPUTATIONAL METHODS

Generating backbone ensembles for the apo, EL and EL† states of KE07 and KE70. Although we mostly focused on the two end state sequences, i.e. the two designed enzymes and the final LDE rounds for KE07 and KE70, some results in the SI material also consider the intermediate rounds of LDE for each of the enzymes. The initial backbone structures and initial definition of the side chain rotameric state of the KE07 apo enzyme for the initial design and LDE rounds 4 and 6 were taken from the PDB database[42]. Apo state structures for rounds without PDB structures were generated using Modeller with the KE07 design as the backbone/side chain template. For KE70, the apo structure of the initial design was taken from the computational model reported elsewhere[2]. For round 2, the apo state structure was taken from the PDB (ID: 3NPX) and rounds 4, 5 and 6 variants were generated by Modeller using the KE70 design as template.

Modeller was used to generate the EL state structure using the apo state as the template for the original designs and all LDE rounds for KE07 and KE70. For the EL state of the KE07 and

KE70 designs, we used the docked structure definition of the ligand as reported elsewhere[2]. The ligand was then kept fixed in its modeled position for all subsequent backbone perturbations and MC-SCE calculations. The substrate geometry for the EL† state was kept the same as in the EL complex, and only TS charges were changed to reflect the transition state of the bound complex.

Using each of these PDB/modeled structures for the backbone in the apo and ligand bound states, we then used the backrub algorithm implemented in Rosetta[21] to run 50 independent simulations, each generating 10,000 trial moves using the C_α atoms as pivot residues, to generate uncorrelated backbone ensembles. From each simulation the lowest energy structure was saved and these 50 low energy backrub structures were selected, and divided into 5 backbone ensembles with 10 structures in each ensemble; this was done for all the rounds for both apo and ligand bound states. Since the backbone scaffolds for KE07 and KE70 are quite rigid, we believe the backbone variations we have generated are adequate.

Generating side chain ensembles for the apo, EL and EL† states of KE07 and KE70. We have recently developed a Monte Carlo Side Chain Ensemble method (MC-SCE)[43] to create large side chain ensembles to calculate the terms in Eq. (2). The MC-SCE method has been validated across a large number of proteins and protein complexes, in which it was found to be highly accurate when compared against high quality X-ray crystallography and NMR J-coupling data for side chain rotameric preferences[43]. The MC-SCE use a Rosenbluth chain growth algorithm to generate an ensemble of side chain packings for a given protein backbone. From the bare backbone conformation m, and for subsequent steps i, the side chain rotamer, r_k, for residue k is selected according to the following probability

$$p_i^{(m,r_k)} = \frac{P_{r_k}^{(PDB)} e^{-\beta E_i^{(m,r_k)}}}{\sum_{\{v_k\}} P_{v_k}^{(PDB)} e^{-\beta E_i^{(m,v_k)}}} \tag{3}$$

where $\{v_k\}$[44] are the possible side chain conformations for residue k, using the values reported in the recent backbone-dependent Dunbrack library[45], which we have augmented by allowing for dihedral angle variations that are Gaussian distributed about a given rotamer value and weighted by its probability of occurrence in the PDB, $P_{r_k}^{(PDB)}$. $E_i^{(m,r_k)}$ is the energy of interaction of side chain conformation r_k of residue k with the backbone and all protein side chains grown so far (step i), using the energy function described above, and all residues are grown with ideal bond lengths and

angles. Once the side chain of a residue is placed, the process is repeated until all the side chains are grown, thereby creating one complete protein structure. Each complete structure *m* is then assigned a weight *W(m)* in order to adjust for sampling bias due to the chain growth as well as to account for energetic solvent effects

$$W(m) = e^{-\beta F_{solv}^{(m)}} \prod_{i=1}^{N} \frac{\sum_{\{v_k\}} P_{v_k}^{(PDB)} e^{-\beta E_i^{(m,v_k)}}}{P_{v_k}^{(PDB)}} \tag{4}$$

For unsuccessful chain growths, the partially grown structure is considered dead and its weight is set to zero. This process is repeated in order to create ~20,000 side chain ensemble on the given backbone.

Since we use a total of 5 independent backbone ensembles, each comprised of 10 backbones, our ensemble for each state are comprised of a total of 1,000,000 fully grown structures. For each of the independent backbone ensembles we calculate the probability $p_{vk}^{(k)}$ of each rotameric state v_k using equation (5)

$$p_{v_k}^{(k)} = \frac{\sum_{m=1}^{M} W(m) \delta_{r_k, v_k}^{(m)}}{\sum_{m=1}^{M} W(m)} \tag{5}$$

where M=200,000 and the Kronecker delta is 1 if the side chain conformation r_k that was picked for the residue *k* in the m-th structure is v_k and 0 otherwise. The probabilities in Eq. (5) are then used to calculate side chain entropy (SCE) of each residue k using the Gibbs probabilistic definition, with SCE values in units of $k_B T$.

$$S_{SC}^{(k)} = \sum_{\{v_k\}} p_{v_k}^{(k)} \log p_{v_k}^{(k)} \tag{6}$$

We estimated the mean and standard deviation for the SCE values from the 5 independent backbone ensembles for the apo, EL and EL for each protein for each round.

Given our MC-SCE method, we can also calculate mutual information, $I^{(i,j)}$. It is defined as the amount of information residue *k* has about another residue *j* based on the amount of coupled side chain dihedral angle fluctuations. In units of $k_B T$, this can be written as

$$I_{SC}^{(k,j)} = \sum_{\{v_k\}} \sum_{\{v_j\}} p_{v_k,v_j}^{(k,j)} \log\left(\frac{p_{v_k,v_j}^{(k,j)}}{p_{v_k}^{(k)} p_{v_j}^{(j)}} \right) \quad (7)$$

where in analogy to Eq. (5)

$$p_{v_k,v_j}^{(k,j)} = \frac{\sum_{m=1}^{M} W(m) \delta_{r_k,v_k}^{(m)} \delta_{r_j,v_j}^{(m)}}{\sum_{m=1}^{M} W(m)} \quad (8)$$

Thus Eq. (7) can be further simplified to

$$I_{SC}^{(k,j)} = (S_{SC}^{(k)} + S_{SC}^{(j)}) - S_{SC}^{(k,j)} \quad (9)$$

in which the individual entropy $S_{SC}^{(k)}$ and joint entropy, $I_{SC}^{(k,j)}$, is calculated using the probabilistic definition of entropy via Eq. (6), and thus Eq. (9) can be interpreted as the degree of coupling of torsional motions of residues *k* and *j*.

In practice, a background error persists in mutual information calculations since two completely uncorrelated variables will never be zero given a finite simulation time. In order to correct for this, we modified the strategy used by Dubay and Geissler[37] to subtract out the erroneous extra mutual information that persists due to finite time scales. We first carry out our MC-SCE chain growth with the full energy function over all backbones in an ensemble, and using Eq. (8) we calculate the mutual information for the *N* structures obtained using the complete energy model, $I_{SC}^{(k,j)}$.

We then use our MC-SCE method to create structures where side chains for each residue are grown independent of the environment, i.e. clashes are ignored and the energy (and hence probability of chain growth) of each side chain conformer ν_k of residue k is given by

$$-\beta E_{uncorr}^{(r_k)} = \log\left(p_{v_k}^{(k)} \right) \quad (10)$$

where the energy in Eq. (10) used in the Rosenbluth sampling is replaced by the log of the probabilities $(p_{v_k}^{(k)})$ determined from Eq. (8) from the full energy MC-SCE simulation to calculate $I_{SC,uncorr}^{(k,j)}$

$I^{(k,j)}_{SC,uncorr}$ for *n* structures that lie beyond the energy cutoff. This value reflects the background error due to the chain growth process and can be cancelled out to yield the true mutual information value as given in Eq. (12).

$$I^{(k,j)}_{SC}(N,n) = I^{(k,j)}_{SC}(N,n) - I^{(k,j)}_{SC,uncorr}(N,n) \tag{11}$$

In this paper, all mutual information (MI) values reported are background corrected.

Reproducibility and Error Analysis. The reproducibility of SCE and MI values was tested on a randomly selected backbone ensemble of R7 and carried out 5 independent times. The data is shown in SI Table S1. SCE values are consistent and the background corrected MI values are reproducible within a reasonable error. The MI values without background correction is also included to give an estimate of the amount of spurious error possible in these calculations. Error bars shown in this paper are standard error of the mean calculated from the backbone ensembles (5 ensembles each for both apo and ligand bound states). As an example, to determine the error in side chain entropy for a set of residues {k}, variances resulting from backbone fluctuation (σ^2) as well as intrinsic error of MC-SCE method (σ'^2) were added up as given in Eq. (12).

$$\sigma^{\{k\}}_{SCE} = \left[\overbrace{\sum_{\{k\}} \left(\sigma^2_{apo,k} + \sigma^2_{lig,k} \right)}^{\text{Backbone variability}} + \overbrace{\sum_{\{k\}} \left(\sigma'^2_{apo,k} + \sigma'^2_{lig,k} \right)}^{\text{Intrinsic error}} \right]^{1/2} \tag{12}$$

Intrinsic error data was taken from the backbone ensemble used to test MI/SCE reproducibility above.

EXPERIMENTAL METHODS

The ligand 5-nitrobenzisoxazole was synthesized by following an earlier published method[46], and its improved version from the Hilvert laboratory[47]. The KE07 R7-2 plasmids were kindly provided by the David Baker laboratory at University of Washington, Seattle, WA, and variants studied in this work were generated by site-directed mutagenesis using a Quik Change II site-directed mutagenesis kit (Stratagene; Agilent Technologies, Santa Clara) using appropriate PCR primers (Table S2). After the mutagenesis PCR reactions, the mutated plasmids were transformed into XL-10 gold cells and the plasmids encoding individual mutations were isolated. The identity of the mutated plasmids were confirmed by sequencing the plasmid from both forward and reverse directions using T7

forward and T7 reverse primers at UC Berkeley Sequencing facility. The individual mutated plasmids were transformed into expression cell line BL21 (DE3) gold.

A single colony from the transformed cells containing individual variant was used to inoculate a starter culture of 20 mL LB medium supplemented with 50 μg/mL kanamycin and the resulting culture incubated with shaking overnight at 37°C. This starter culture was used to inoculate 500 mL LB medium with 50 μg/mL kanamycin and incubated for ~3h at 37°C until OD600 reached ~1.2. The culture was then induced with 1mM IPTG for overproduction and the culture was further grown with shaking at 37°C for 4h. The cells from the liquid culture were harvested and stored at -80°C until used for the isolation. In general, roughly 2 g of the wet cells were routinely obtained from 0.5L culture.

The harvested cells were thawed, re-suspended in 35 mL lysis buffer (25 mM Hepes, pH 7.25 containing 100 mM NaCl, 5% glycerol), lysed by sonication, centrifuged to remove insoluble debris and the soluble fraction loaded into pre-washed NI-NTA column (5mL resin, His-Pur, Thermo-Fisher). The NI-NTA resin with the bound proteins were washed first with 10 column volume of lysis buffer followed by 15 column volume of 20 mM NaPi, pH 7.4, 500 mM NaCl, 30 mM Imidazole to remove nonspecific and weakly bound proteins. The bound His-tagged fusion protein was then eluted from the NI-NTA resin with 20-25 mL of 500mM Imidazole buffer solution (20 mM NaPi pH 8.0, 500 mM NaCl, 500 mM Imidazole). The eluted fusion protein were extensively dialysed in lysis buffer, concentrated through Amicon filters (30,000 MWCO, Millipore), its concentration estimated by measuring the absorbances at 280 nm and stored at -80°C in smaller aliquots. This purification protocol yielded over 90% pure protein (assessed through the visible bands in SDS-PAGE) and routinely produced 18-23 mg of His-tagged KE07 proteins.

The enzymatic characterization of the KE07 R7 variants was performed similar to previously published work[42] with some modification in the Cary 50 spectrophotometer (Varian) that used a quarz cuvette. In short, the kinetic analysis were performed in 25 mM Hepes, pH 7.25, 100 mM NaCl, 5% glycerol with 5-nitrobenzisoxazole concentration ranging from 5-1500 μM with the co-solvent acetonitrile concentration equalized to 1.5% (v/v) in a micro-cuvette capable of monitoring reaction at 200 μL. A known amount of dry 5-nitroxybenzisoxazole was dissolved in acetonitrile to have 100mM substrate stock. From this stock a series of dilutions of the substrate were made in acetonitrile to achieve the concentration ranges in the kinetic assay. The reaction was initiated by the addition of small amount of the enzyme aliquot (final concentration from 0.2-1.0 μM in the assay) and the product formation was monitored spectrophotometrically at 380 nm ($\Delta\varepsilon$ =

15,800 M-1, cm-1). Steady-state parameters were obtained after fitting the data to the Michelis-Menten equation.

RESULTS

For KE07 (Figure 1b), the key intended active site residues include Glu101 as the catalytic base, Lys222 for stabilizing the developing negative charge on oxygen in the transition state, and Trp50 as a π-stacking residue to orient the 5-nitro-benzisoxazole ligand (Figure 1c). In addition, 10 other positions in the original scaffold (1THF) were changed to accommodate the engineered active site, culminating in a total of 13 designed residues for KE07. The initial design exhibited very poor activity (k_{cat}/K_m = 12 $M^{-1}s^{-1}$) but after 7 rounds of LDE, a two-order improvement in catalytic performance was obtained for KE07-R7[42]. Table S3 lists the KE07 designed residues and the sequence changes made during LDE, as well as the corresponding improvements in k_{cat} and K_M for each round.

Enzyme KE70 (Figure 1d) also utilized a TIM barrel scaffold but one that differed from KE07 (deoxyribose phosphate aldolase from E. coli, PDB 1JCL). KE70 was designed using a His17-Asp45 dyad as the catalytic base, Ser138 as the charge stabilizing residue and Tyr48 as the π-stacking residue (Figure 1e). In addition, 12 other positions were designed to support the incorporation of the new active site. In terms of catalytic performance, the original KE70 design was an order of magnitude better than KE07 (k_{cat}/K_m = 126 $M^{-1}s^{-1}$) and with LDE KE70 reached a peak performance in round 6 (KE70-R6) that led to a further 450 factor improvement over its starting sequence[48]. Table S4 summarizes the original design, the mutations from straight DE (i.e. random mutagenesis), and later rounds using “spiked” DE through recombination of new design features (R2, R4 and R6) and the corresponding improvements in k_{cat} and K_M for each round.

Nearly all of the LDE changes in KE07 were satellite residues in the undesigned regions of the scaffold, with only one designed residue being mutated in the first round of LDE (Asn224Asp). In stark contrast to the LDE results for KE07, the designed residues in KE70 were directly targeted for change such that the best R6 variant mutated 7 of the originally designed residues, some of which were in the active site. While this might imply that the KE07 design was robust, our MD and MC-SCE simulations found that the overall active site chemistry was quite different than that shown in Figure 1c. Although Lys222 was a designed residue whose role is to stabilize the charged ligand in the transition state, instead we found that the heavy atom distances for Lys222N_ζ to the ligand oxygen was greater than 5.0 Å in all KE07 enzyme constructs; this is consistent with previous

studies[4, 5] that showed that Lys222 is never in spatial proximity to the ligand to fulfill this role. Instead we find that Lys222 often forms a hydrogen bond with Ser48, as well as with residues Glu46 and Ile7 or its replacement in LDE R4 with Asp7; we find that catalytic activity is annihilated when we perform site mutagenesis at positions Ser48 and Lys222 (Table 1), as was true for mutation of Asp7 reported elsewhere[42]. This supports the reasoning of Khersonsky et al. that Asp7 serves to tether Lys222 so that it does not have unproductive interactions with the catalytic base[42], although we find a more extended network of Lys222 interactions. Hence, although Lys222 never fulfilled its intended design role, it is involved in interactions that nonetheless support the catalytic purpose of KE07[42].

Instead, we find that His201 is closest to the oxygen of the substrate heterocycle, with heavy atom distances between His201N_ε and the ligand oxygen found to be ~3.5-4.0 Å; Table 1 reports the experimental mutation at His201Ala and confirms that it destroys all enzyme activity. Furthermore the Gly202Arg mutation introduced in all rounds of LDE resulted in a very stable hydrogen bond between the Arg202-N_ζ and the nitro group of the ligand, and the designed Tyr128 forms a hydrogen bond with Arg202 that appears to further stabilize that interaction; in fact when Tyr128 is mutated to Phe, all enzyme activity is destroyed (Table 1). Similar “re-purposing” of other scaffold residues to aid in ligand positioning or charge stabilization has also been observed in crystal structures of another *de novo* designed Kemp eliminase, HG3.17 with a substrate analog[49]. Figure 2 shows the rotamer flexibility found in the greater network of the active site region of the best performing R7 variant for KE07, which stands in contrast to the static truncated active site assumed during the design process (Figure 1c). Further details pertaining to Figure 2 are given in Table S7.

We next consider an overall thermodynamic analysis of the Michaelis-Menten scheme and the enthalpy and entropy breakdowns for the relative free energy of stabilization of the apo state, EL reactive complex and the $EL^\dagger$ transition state complex (Table 2) for the designed enzymes and their best evolved variant KE07-R7 and KE70-R6. Note that for numerical calculations of free energy we ignore mutual information contributions due to the poor convergence of Eq. (2) where higher order correlations are clearly needed. Although we account for ligand solvation free energies by evaluating the ligand in our implicit solvent model, we are also missing explicit solvation or other types of solvent reorganization contributions that will stabilize each state differently. Furthermore, we model the transition state classically using altered partial charges that attempt to describe the electrostatics of bond-making and bond-breaking of the true quantum mechanical process. As such the absolute thermodynamic values for each state should be taken with caution, as we would require

these additional contributions to connect to the experimental k_{cat} and K_M numbers. The idea behind the free energy analysis is instead to show how the individual contributions of side chain entropy and enthalpy reproduce the overall trends in these quantities, and yield a fairly suggestive picture as to why the KE07-R7 and KE70-R6 enzymes proved to be better biocatalysts than their designed counterparts.

We find that the enthalpy change between the EL complex and the apo state of the enzyme, $\Delta H_{EL} = \langle H_{EL} \rangle - \langle H_E \rangle$ is destabilized in the best evolved KE07-R7 and KE70-R6 enzymes compared to the original designs, consistent with what has been reported previously using EVB calculations[6]. However, we find the same destabilization trend is also observed for the entropy as well, since both designed enzymes exhibit $-T\Delta S_{EL} = -T\left(\langle S_{EL} \rangle - \langle S_E \rangle\right) < 0$; this means that there is greater conformational flexibility when the enzyme binds the ligand relative to the apo state, thereby stabilizing the enzyme-substrate complex. However, the introduction of new mutations in successive rounds of LDE leading to KE07-R7 and KE70-R6 contributes to reduction in the favorable entropy of the EL state ($-T\Delta S_{EL} > 0$), and hence the entropy also contributes to destabilization of the EL complex in the best LDE enzymes.

We also evaluate the enthalpy and entropy of the $EL^{\dagger}$ complex and how that changes with respect to the EL state based on a linear response approximation. We first assume an adiabatic step in which the $EL^{\dagger}$ complex is averaged over the EL ensemble to isolate the enthalpy, and then a subsequent step to account for enthalpic and entropic contributions due to enzyme reorganization in response to the change in ligand charges by averaging over the $EL^{\dagger}$ ensemble.

$$\Delta H^{\dagger} = \langle \Delta H^{\dagger} \rangle_{EL} + \langle \Delta H^{\dagger} \rangle_{EL^{\dagger}} \tag{13}$$

$$\Delta S^{\dagger} = \langle \Delta S^{\dagger} \rangle_{EL^{\dagger}} \tag{14}$$

Based on the linear response approximation using Eqs. (13) and (14), we find a very small stabilization of the adiabatic enthalpy for KE07-R7 relative to the original design, consistent with previous EVB calculations[6]. By contrast the large number of active site modifications made on the KE70 enzyme is consistent with the fact that the adiabatic enthalpy barrier is nearly halved in the

KE70-R6 enzyme. However, by considering the reorganization terms as well, we find that there is transition state stabilization not only through the enthalpy, but that the entropy further lowers the catalytic barrier of the best enzymes relative to the original designs for both KE07 and KE70. Thus our thermodynamic calculations summarized in Table 2 supports the view that the active site of the original KE07 and KE70 enzymes were over-designed for the binding affinity of the EL state, whereas the LDE process created enzymes that unambiguously preferred the $EL^{\dagger}$ complex instead, especially for the KE70 enzyme.

Although the higher order terms in the entropy expansion in Eq. (2) may be directly related to k_{cat}/K_M, they can't currently be included for numerical calculations for free energies since higher order correlations are required for convergence of the total entropy. Nonetheless, we show that mutual information can yield even further insight as to why the evolved Kemp eliminases are better enzymes by focusing on residues with the largest mutual information with other residues; more specifically, such a position is defined as a "network hub" when it has a large ΔI as measured by |Z-scores| > 2 with at least 25 other residues throughout the scaffold. While this definition is somewhat arbitrary, it does quantify the residues with the strongest correlations with a large number of other residues, i.e. those with highest MI are always found using other definitions. One of the important features of the network hubs is that they are all mutually coupled, i.e. they each count as one of their connections all of the other hub residues. We shall see that network hubs are often identified as active site residues as well as mutational hot spots during the directed evolution process (Tables S5 and S6).

Figures 3 and 4 show how the network hubs are distributed over the scaffold in the apo state, EL state, as well as the $EL^{\dagger}$ state, for the original designs and the best KE70-R6 and KE07-R7 enzymes, respectively. As evident from Figure 3 and tabulated in Table 3, the designed KE70 enzyme has high MI in the EL state and low MI in the apo and transition state. Furthermore we identify the His17-Asp45 dyad as 2 network hubs whose motions are strongly correlated with residues in KE70 that were subsequently mutated during the LDE process (23, 29, 48, 74, 166; see Tables S5). However, by the end of LDE the strongly correlated network in the EL state has been destroyed in favor of high MI in the apo state and transition state instead (Table 3). For KE07, there are no active site residues that are identified as network hubs in the designed enzyme for any of the states (Table 4). However, 7 out of the 13 LDE mutations were classified as a network hub at some point during the LDE process (Table S6), so that by the end of LDE the best evolved KE07-R7 enzyme exhibits network hubs involving active site residues 7, 50, 128, 201, 202, and 222 in the

apo and/or $EL^{\dagger}$ states. In turn the active site residues are highly correlated with *other* network hub residues, some of which are located far away from active site (> 10 Å) for the R7 variant of KE07.

To test the robustness of whether these other network hub residues are catalytically important due to their connection to the active site residues, we experimentally mutated network hubs for KE07-R7 (Table 4). The identified networks hubs included and were mutated as follows: Arg16Gln, Asn25Ser, Leu52Ala, Met62Ala, His84Tyr, Lys132Asn, and finally Ile199 to Ser, Phe, and Ala (Table 1). In all cases activity was diminished, with k_{cat}/K_M values anywhere between 10% to 78% of the KE07-R7 result, highlighting that residues located far away from the active site can also affect catalytic activity. We also performed two types of control experiments, in which a residue not identified to be a network hub is mutated (Lys162Ala, Leu170Ala and Glu185Ala) or in one case replacing a network hub residue with another residue that was also found to be a network hub and correlated to the active site (Lys132Met). We found in all four control experiments that catalytic activity was unaffected, even though one position Leu170 is within close proximity to the active site and substrate. This result clearly illustrates that residues with high mutual information are critical in the improved enzymatic activity of the KE07-R7 variant and by extension to the KE70 enzyme as well.

CONCLUSIONS

Given our current limitations in developing robust enzyme designs, laboratory directed evolution provides an attractive addition to rational computational design approaches since it is highly flexible in application to different biocatalysis reactions. Nonetheless although often highly successful, LDE is an opaque process because it offers no complete rationale as to why the mutations were successful, and therefore stands outside our ability to systematically reach novel catalysis outcomes. To bridge this design gap, we have used side chain entropy and mutual information metrics applied to two different *de novo* enzymes and their LDE variants to better understand how conformational flexibility influences catalysis, which is central to many prominent proposals about the origin of enzyme activity[50-52]. Our analysis showed that by the end of the LDE process that changes in entropy, as well as enthalpy, helped to destabilize the EL complex in favor of stabilization of the $EL^{\dagger}$ complex for both KE07-R7 and KE70-R6 when compared to the designed enzymes. Furthermore, we identified new active site players in KE07-R7, and using site mutagenesis showed that residues with large mutual information are catalytically important in KE07-R7 even though they may be remote from the active site.

There are two prominent but competing proposals as to what are the most important considerations in optimizing enzyme performance. Warshel and co-workers have emphasized that electrostatic pre-organization is the primary strategy by which enzymes achieve their remarkable catalytic activity compared to the uncatalyzed reaction[6, 53]. To recapitulate that argument, the electrostatic environment of the enzyme active site is structurally optimized in the apo state such that it is pre-organized to preferentially bind the transition state over reactants or products, thereby lowering $\Delta G_T^{\dagger}$ relative to that of the uncatalyzed reaction that must fold in the cost of reorganization factors (polarization) that raise the catalytic barrier. The other proposal is that conformational motion can also be key to catalytic performance by lowering $\Delta G_T^{\dagger}$, where equilibrium statistical fluctuations[23-25], dynamical coupling[26] and maximizing the reactive flux through the transition state surface[27] have emerged as potentially important dynamical aspects of the success of natural enzymes.

We believe that our results presented here on side chain entropy and mutual information are consistent with both the dynamical picture and the electrostatic pre-organization principle long advocated by Warshel and co-workers[6, 53]. For KE07-R7, network hubs in the apo state including Tyr128, His201 and Arg202 formed direct electrostatic interactions with the substrate, or the remote residues were charged or polar residues (Arg16, Asp25 and Lys132) whose long-ranged electrostatic effects clearly played some role in lowering $\Delta G_T^{\dagger}$ given that our experimental results showed reduced activity when these residues were mutated.

Furthermore we note a very interesting observation that there are high mutual information hubs in the EL state with much fewer MI hubs in the apo and $EL^{\dagger}$ transition state complex in both the KE07 and KE70 designs. This we believe could be a signature of the problem of over-design of the EL state using the Rosetta strategy, and that LDE intervened to create new residue correlations in the apo and $EL^{\dagger}$ transition state complex in the most improved enzyme variants. While highly speculative, it may be evidence for pre-organization signatures in the apo state, and a network of interactions that favor the $EL^{\dagger}$ complex instead of the EL complex. At this point we can only quantify the importance of these changes in high MI sites through experimental mutagenesis.

Given that the active site residues of both KE07 and KE70 proved to have high mutual information and were strongly networked to other residues that also have extensive networks of strong side chain couplings, we believe that it should be possible to use computation to propose new sequence mutations that will improve the catalytic activity of other Kemp Eliminase enzymes for which LDE has not been performed. Furthermore, our method needs to be generalized to enzymes with substrates that are larger and more flexible than 5-nitroxybenzisoxazole, and estimating the enthalpy and entropy contribution from water would require explicit treatment of water, currently an area of research in our lab.

ACKNOWLEDGEMENTS. This work was supported by the Laboratory Directed Research and Development Program of Lawrence Berkeley National Laboratory under U.S. Department of Energy Contract No. DE-AC02-05CH11231.

REFERENCES

1. L. Jiang, E. a. Althoff, F. R. Clemente, L. Doyle, D. Röthlisberger, A. Zanghellini, J. L. Gallaher, J. L. Betker, F. Tanaka, C. F. Barbas, D. Hilvert, K. N. Houk, B. L. Stoddard and D. Baker, *Science* 2008, **319**, 1387-1391.
2. D. Röthlisberger, O. Khersonsky, A. M. Wollacott, L. Jiang, J. DeChancie, J. Betker, J. L. Gallaher, E. a. Althoff, A. Zanghellini, O. Dym, S. Albeck, K. N. Houk, D. S. Tawfik and D. Baker, *Nature*, 2008, **453**, 190-195.
3. M. Casey and D. Kemp, *J. Org. Chem.*, 1973, **58**, 33-34.
4. A. N. Alexandrova, D. Röthlisberger, D. Baker and W. L. Jorgensen, *J. Amer. Chem. Soc.*, 2008, **130**, 15907-15915.
5. G. Kiss, D. Röthlisberger, D. Baker and K. N. Houk, *Protein science*, 2010, **19**, 1760-1773.
6. M. P. Frushicheva, J. Cao, Z. T. Chu and A. Warshel, *Proc Natl Acad Sci U S A*, 2010, **107**, 16869-16874.
7. a. Labas, E. Szabo, L. Mones and M. Fuxreiter, *Biochim. Biophys. Acta*, 2013, **1834**, 908-917.
8. C. Jackel, P. Kast and D. Hilvert, *Annu Rev Biophys*, 2008, **37**, 153-173.
9. D. N. Bolon, C. A. Voigt and S. L. Mayo, *Curr Opin Chem Biol*, 2002, **6**, 125-129.
10. S. V. Taylor, P. Kast and D. Hilvert, *Angew Chem Int Ed Engl*, 2001, **40**, 3310-3335.
11. F. H. Arnold, *Acc Chem Res*, 1998, **31**, 125-131.
12. P. A. Romero and F. H. Arnold, *Nat Rev Mol Cell Biol*, 2009, **10**, 866-876.
13. D. N. Bolon and S. L. Mayo, *Proc Natl Acad Sci U S A*, 2001, **98**, 14274-14279.
14. J. R. Cherry, M. H. Lamsa, P. Schneider, J. Vind, A. Svendsen, A. Jones and A. H. Pedersen, *Nature Biotechnol.*, 1999, **17**, 379-384.
15. C. A. Voigt, S. L. Mayo, F. H. Arnold and Z.-G. Wang, *Proc Natl Acad Sci USA*, 2001, **98**, 3778-3783.
16. S. Wu, J. P. Acevedo and M. T. Reetz, *Proceedings of the National Academy of Sciences of the United States of America*, 2010, **107**, 2775-2780.

17. R. J. Fox, S. C. Davis, E. C. Mundorff, L. M. Newman, V. Gavrilovic, S. K. Ma, L. M. Chung, C. Ching, S. Tam, S. Muley, J. Grate, J. Gruber, J. C. Whitman, R. A. Sheldon and G. W. Huisman, *Nature Biotechnol*, 2007, **25**, 338-344.
18. C. A. Voigt, S. L. Mayo, F. H. Arnold and Z. G. Wang, *Proc Natl Acad Sci U S A*, 2001, **98**, 3778-3783.
19. C. T. Saunders and D. Baker, *J. Mol. Bio.*, 2002, **322**, 891-901.
20. S. J. Fleishman, S. D. Khare, N. Koga and D. Baker, *Protein Sci*, 2011, **20**, 753-757.
21. A. J. Smith, R. Müller, M. D. Toscano, P. Kast, H. W. Hellinga, D. Hilvert and K. N. Houk, *J. Amer. Chem. Soc.*, 2008, **130**, 15361-15373.
22. A. Fersht, *Enzyme Structure and Mechanism*, WH Freeman, New York, New York, 1985.
23. J. Klinman, *Dynamics in Enzyme Catalysis*, Springer Berlin Heidelberg, Berlin, Heidelberg, 2013.
24. K. F. Wong, T. Selzer, S. J. Benkovic and S. Hammes-Schiffer, *Proc Natl Acad Sci U S A*, 2005, **102**, 6807-6812.
25. D. D. Boehr, R. Nussinov and P. E. Wright, *Nature Chem. Bio.*, 2009, **5**, 789-796.
26. G. Bhabha, J. Lee, D. C. Ekiert, J. Gam, I. a. Wilson, H. J. Dyson, S. J. Benkovic and P. E. Wright, *Science*, 2011, **332**, 234-238.
27. N. Boekelheide, R. Salomón-Ferrer and T. F. M. III, *Proc Natl Acad Sci USA*, 2011, **108**, 16159.
28. K. Henzler-Wildman and D. Kern, *Nature*, 2007, **450**, 964-972.
29. K. a. Henzler-Wildman, M. Lei, V. Thai, S. J. Kerns, M. Karplus and D. Kern, *Nature*, 2007, **450**, 913-916.
30. G. J. Bartlett, C. T. Porter, N. Borkakoti and J. M. Thornton, *J. Mol. Bio.*, 2002, **324**, 105-121.
31. J. S. Fraser, H. van den Bedem, A. J. Samelson, P. T. Lang, J. M. Holton, N. Echols and T. Alber, *Proc Natl Acad Sci U S A*, 2011, **108**, 16247-16252.
32. K. K. Frederick, M. S. Marlow, K. G. Valentine and a. J. Wand, *Nature*, 2007, **448**, 325-329.
33. L. M. Tuttle, H. J. Dyson and P. E. Wright, *Biochemistry*, 2013, **52**, 3464-3477.
34. J. S. Fraser, M. W. Clarkson, S. C. Degnan, R. Erion, D. Kern and T. Alber, *Nature*, 2009, **462**, 669-673.
35. K. Gunasekaran, B. Ma and R. Nussinov, *Proteins*, 2004, **57**, 433-443.
36. J. Monod, J. Wyman and J. P. Changeux, *J. Mol. Bio.*, 1965, **12**, 88-118.
37. K. H. Dubay, J. P. Bothma and P. L. Geissler, *PLoS Comp. Bio.*, 2011, **7**, e1002168.
38. M. S. Lin, N. L. Fawzi and T. Head-Gordon, *Structure* 2007, **15**, 727-740.
39. M. S. Lin and T. Head-gordon, *J. Comp. Chem.*, 2011, **32**, 709-717.
40. C. L. McClendon, G. Friedland, D. L. Mobley, H. Amirkhani and M. P. Jacobson, *J Chem Theory Comput*, 2009, **5**, 2486-2502.
41. B. J. Killian, J. Yundenfreund Kravitz and M. K. Gilson, *J Chem Phys*, 2007, **127**, 024107.
42. O. Khersonsky, D. Röthlisberger, O. Dym, S. Albeck, C. J. Jackson, D. Baker and D. S. Tawfik, *J. Mol. Bio.*, 2010, **396**, 1025-1042.
43. A. Bhowmick and T. Head-Gordon, *Structure*, 2015, **23**, 44-55.
44. V. Hornak, R. Abel, A. Okur, B. Strockbine, A. Roitberg and C. Simmerling, *Proteins*, 2006, **65**, 712-725.
45. M. V. Shapovalov and R. L. Dunbrack, *Structure* 2011, **19**, 844-858.
46. H. Lindemann and H. Thiele, *Justus Liebigs Ann. Chem.*, 1926, **449**, 63-81.
47. F. Hollfelder, A. J. Kirby, D. S. Tawfik, K. Kikuchi and D. Hilvert, *J. Amer. Chem. Soc.*, 2000, **122**, 1022-1029.

48. O. Khersonsky, D. Röthlisberger, A. M. Wollacott, P. Murphy, O. Dym, S. Albeck, G. Kiss, K. N. Houk, D. Baker and D. S. Tawfik, *J. Mol. Bio.*, 2011, **407**, 391-412.
49. R. Blomberg, H. Kries, D. M. Pinkas, P. R. E. Mittl, M. G. Grütter, H. K. Privett, S. L. Mayo and D. Hilvert, *Nature*, 2013, **503**, 418-421.
50. J. Villà, M. Štrajbl, T. M. Glennon, Y. Y. Sham, Z. T. Chu and A. Warshel, *Proc Natl Acad Sci U S A*, 2000, **97**, 11899-11904.
51. M. J. Stone, *Acc. Chem. Res.*, 2001, **34**, 379-388.
52. A. J. Wand, *Nature Struct. Bio.*, 2001, **8**, 926-931.
53. M. P. Frushicheva, J. Cao and A. Warshel, *Biochemistry*, 2011, **50**, 3849-3858.

Table 1. *Experimental validation of effect of mutating network hubs in R7 on catalytic activity.* Steady-state measurements were recorded in 20 mM Hepes at pH 7.25 containing 100 mM NaCl, 5% glycerol at 20°C. The substrate (5-nitroxybenzisoxazole) was dissolved in acetonitrile and the enzymatic assay contained final concentration of acetonitrile at 1.5% (v/v).

KE07 Variant	k_{cat}, s^{-1}	K_M, mM	K_{cat}/K_M, $M^{-1}s^{-1}$	% Activity relative to R7
R7	0.81(0.01)	407(15)	1990(79)	100.0
Active Site Residues				
R7, S48N	0.1	689.7	145	7.3
R7, Y128F	-	-	-	~0
R7, H201A[a]	-	-	108(11)	5.4
R7, H201K[a]	0.562	5411.4	104	5.2
R7, K222A[a]	-	-	40 (6)	2.0
Distance Residues with high MI				
R7, R16Q	0.57(0.02)	589(46)	968(83)	49
R7, N25S	0.58(0.03)	479(57)	1221(157)	61
R7, L52A	0.51	514	992	50
R7, M62A	0.64	542	1181	59
R7, H84Y	0.77	497	1549	78
R7, K132N	0.75	560	1339	67
R7, I199S	0.33	771	428	22
R7, I199F	0.26	564	461	23
R7, I199A	0.23	1467	155	7.8
Controls				
R7, K132M	0.72	352	2045	116
R7, K162A	0.72	419	1718	86
R7, L170A	0.65(0.01)	338(19)	1929(115)	98
R7, E185A	0.89(0.02)	430(19)	2065(98)	104

[a] These variants did not exhibit substrate saturation and only sub-saturating substrate concentration data points were used to estimate k_{cat}/K_M.

Table 2. *Evaluation of the free energy under the Michaelis-Menten scheme for KE07 and KE70.* Calculated enthalpy and entropy differences between apo, EL and EL† states and their summed free energies, all in kcal/mole. We use a linear response approximation to evaluate the energy and entropy contributions for the transition state that involves the addition of an adiabatic step followed by enzyme reorganization (see text) in order to define the total free energy change. Note that we

ignore mutual information contributions due to the poor convergence of the total entropy in Eq. (2), and we can't reliably account for explicit solvent free energy contributions, and hence we can't make direct or quantitative contact with k_{cat} and K_M values. We can only describe the qualitative trends in side chain entropy and enthalpy as shown.

State Function	KE07	KE07-R7		KE70	KE70-R6
EL Stabilization					
ΔH_{EL}	-9.9	-3.6		-13.5	5.7
$-T\Delta S_{EL}$	-6.3	4.5		-4.4	0.7
ΔG_{EL} [a]	-33.7	-16.4		-35.4	-11.0
EL† Barrier (Adiabatic)					
$\langle \Delta H^{\dagger} \rangle_{EL}$	11.6	10.5		15.7	8.7
$\langle -T\Delta S^{\dagger} \rangle_{EL}$	0	0		0	0
EL† Barrier (Reorganization)					
$\langle \Delta H^{\dagger} \rangle_{EL^{\dagger}}$	-2.0	-3.3		1.5	-3.0
$\langle -T\Delta S^{\dagger} \rangle_{EL^{\dagger}}$	-0.7	-3.8		3.6	1.7
EL† Barrier (Total)					
$\Delta G^{\dagger}$	8.9	3.4		20.8	7.4

[a] Includes the ligand solvation free energy, calculated from our model to be 17.5 kcal/mole.

Table 3. *Residues that were determined to be network hubs with high mutual information for KE70*. Residues colored red were designed into the scaffold of 1JCL and residues colored blue were mutated during the course of LDE; the only exceptions are residues 43, 48, 74, and 166 that were both a designed and mutated residue.

Round	Highest MI in Apo state	Highest MI in EL complex	Highest MI in EL† complex
Design	27, 64, 143	6, 11, 14, 17, 23, 24, 29, 38, 45, 48, 58, 67, 70, 74, 83, 90, 100, 104, 115, 117, 121, 142, 147, 153, 154, 166, 167, 170, 173, 184, 186, 188, 191, 193, 216, 217, 221, 247	28
R6	11, 18, 25, 33, 35, 45, 50, 52, 56, 59, 64, 67, 70, 83, 90, 115, 118, 148, 154, 170, 174, 198, 223, 247	10, 15, 17, 22, 58, 76, 123, 148, 165, 232	6, 14, 25, 28, 43, 58, 73, 95, 97, 100, 107, 109, 115, 120, 123, 124, 136, 141, 154, 160, 166, 173, 185, 186, 193, 196, 204, 247, 249

Table 4. *Residues that were determined to be network hubs with high mutual information for KE07*. Residues colored red were designed into the scaffold of 1THF and residues colored blue were mutated during the course of LDE. The bold faced residues identified as network hubs in R7 were subjected to mutagenesis to confirm that they reduced enzyme activity.

Round	Highest MI in Apo state	Highest MI in EL complex	Highest MI in EL† complex
Design	4, 10, 63, 66, 71, 87, 118, 212	16, 19, 58, 68, 85, 86, 139, 163, 174, 175, 185, 230, 232, 235	19, 51, 58, 64, 68, 91, 123, 161
R7	12, **16**, **25**, 42, **52**, 74, 94, 95, 128, **132**, 133, 149, 201, 202, 209, 222, 230	5, 19, **62**, 63, 71, 73, **84**, 87, 91, 92, 118, 148, 155, 159, **199**, 244, 247	7, 12, 37, 42, 50, 52, 58, 68, **84**, 92, 137, 148, 159, 182, 201, 208, 222, 230, 238

FIGURE CAPTIONS

Figure 1. *The Kemp elimination KE07 and KE70 designs*. (a) The one-step reaction scheme involving the abstraction of hydrogen from 5-nitro benzisoxazole by a catalytic base. Shown is the transition state that has a partial negative charge on the substrate oxygen with cleavage of the O-N bond. (b) KE07 involved residues mutated from the original scaffold (red) as well as mutations introduced by LDE shown in blue. (c) Relative orientation of the key catalytic residues with respect to the ligand in the ideal active site of KE07. (d) KE70 involved residues mutated from the original scaffold (red) as well as mutations made during laboratory DE shown in blue. Additional design mutations via a recombination DE strategy are shown in green (e) Relative orientation of the key catalytic residues with respect to the ligand in the ideal active site of KE70.

Figure 2. *The Kemp elimination KE07 active site in the best R7 variant*. The percentage represents the occupation of each rotamer as determined from the side chain ensemble of KE07-R7. We note that no one has reported on the importance of either His201 nor Tyr128 for the active site chemistry in KE07, which has been confirmed by experimental site mutagenesis in Table 1.

Figure 3. *Change in high mutual information hubs for the apo state, EL state, and EL† state for (a) designed KE70 and (b) the KE70-R6 variant.* The spheres represent residues that are high mutual information hubs (centered at the Cα position). We have uploaded an interactive visualizer of these hubs on our website at http://thglab.berkeley.edu

Figure 4. *Change in high mutual information hubs for the apo state, EL state, and EL† state for (a) designed KE07 and (b) the KE07-R7 variant.* The spheres represent residues that are high mutual information hubs (centered at the Cα position). We have uploaded an interactive visualizer of these hubs on our website at http://thglab.berkeley.edu

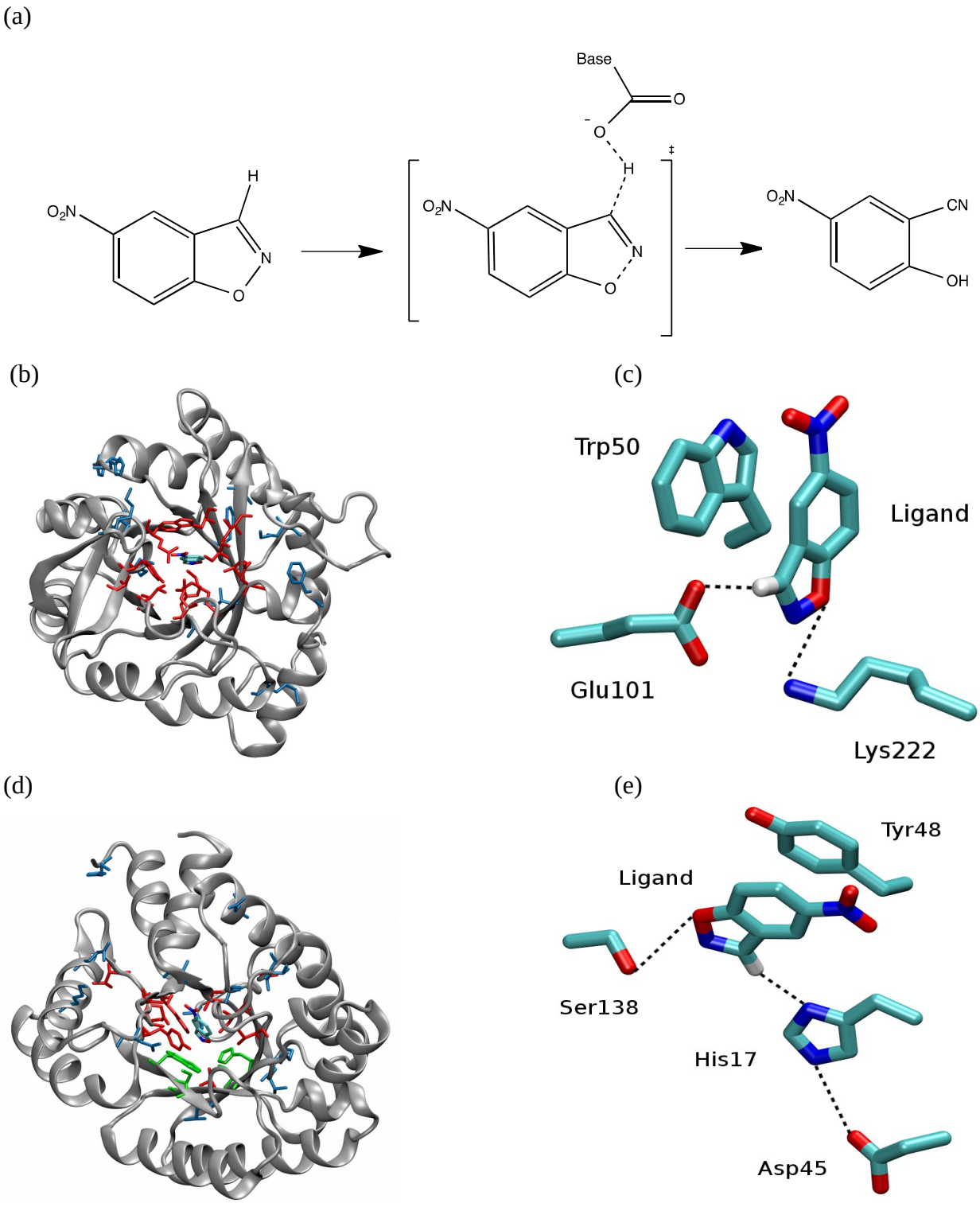


Figure 1. Bhowmick and co-workers

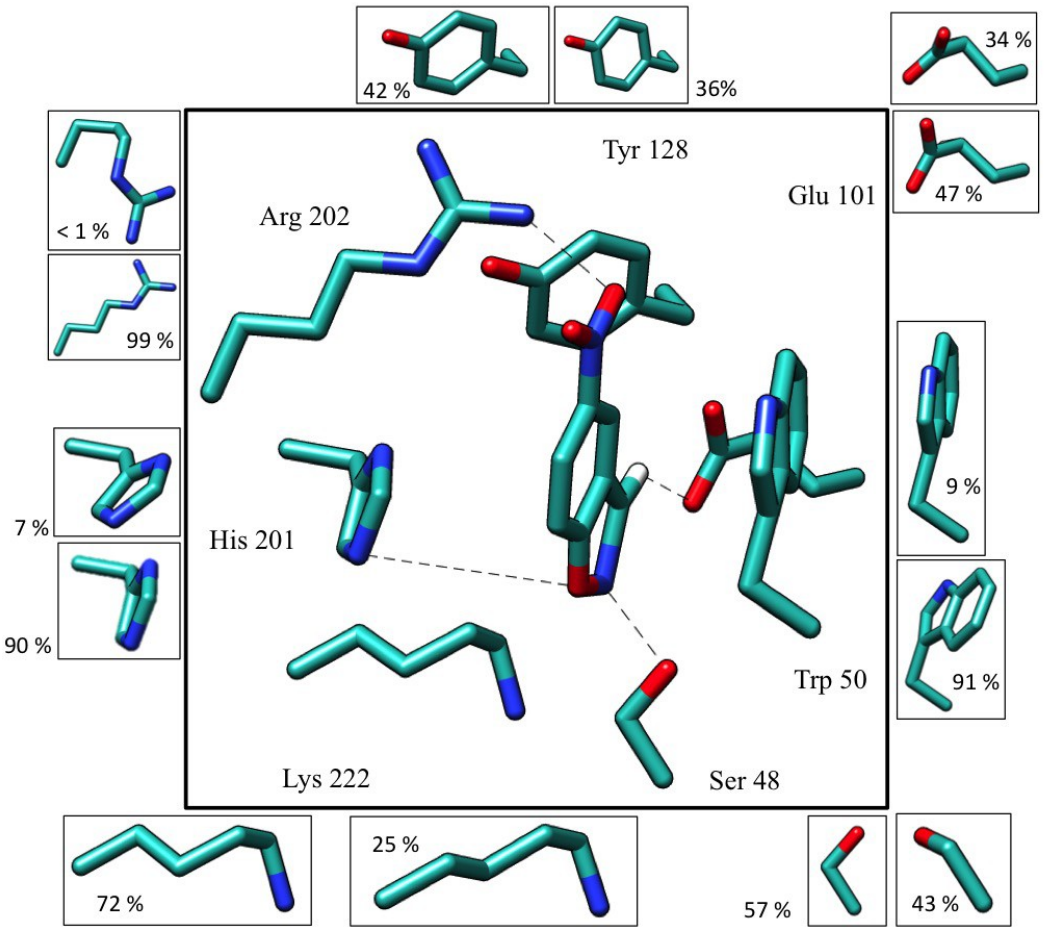


Figure 2. Bhowmick and co-workers

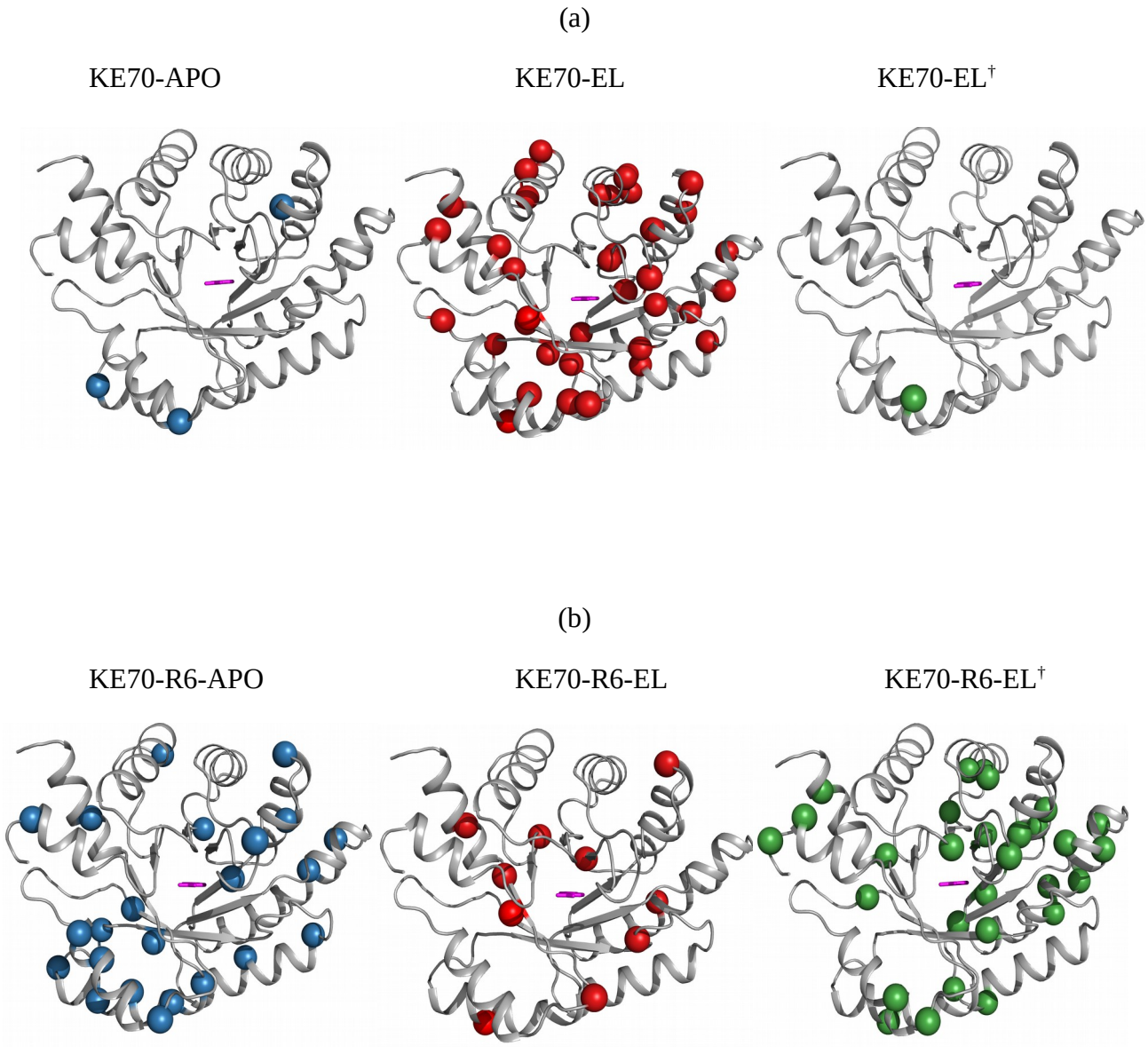
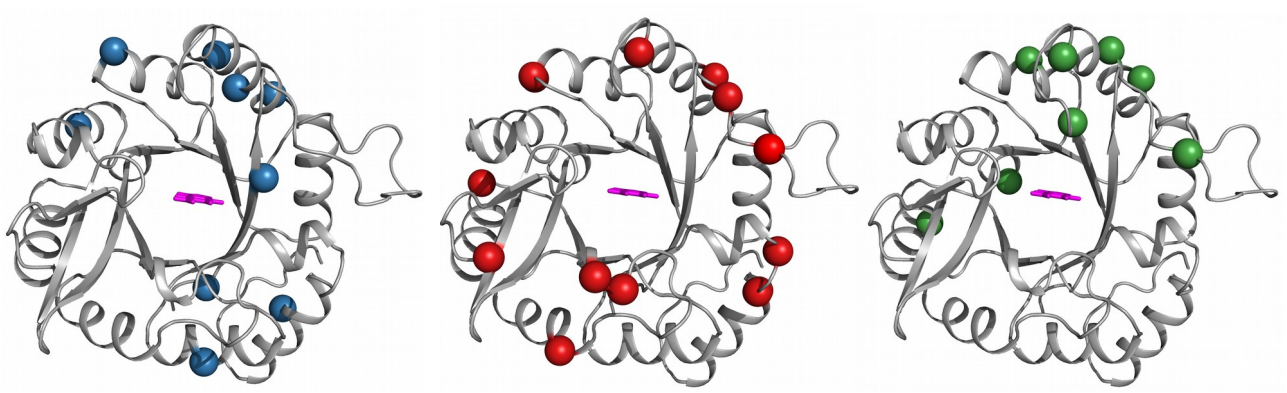


Figure 3. Bhowmick and co-workers

(a)

KE07-APO KE07-EL KE07-EL†



(b)

KE07-R7-APO KE07-R7-EL KE07-R7-EL†

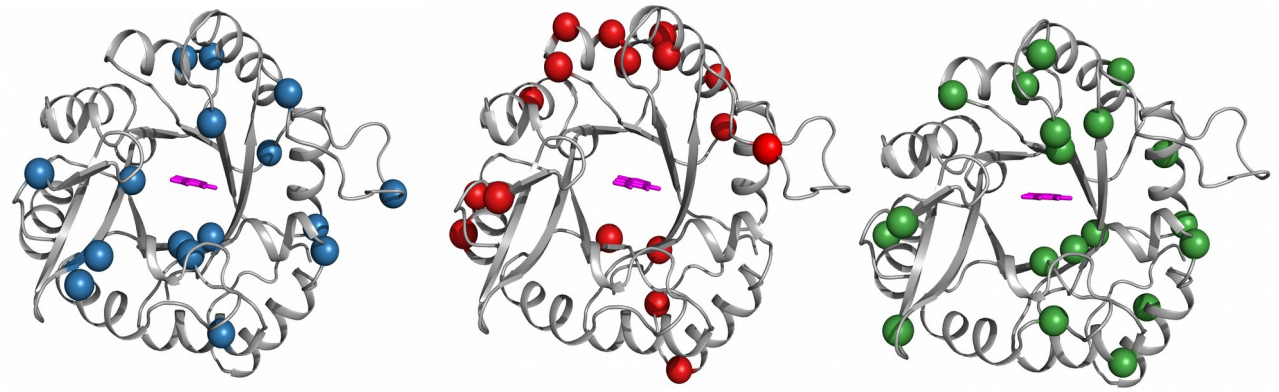


Figure 4. Bhowmick and co-workers