Chapter III

Towards the computational design of a Kemp elimination enzyme and investigation of an inactive design

A majority of the experimental work described in this chapter was performed in the Mayo laboratory. Dr. Leonard Thomas of the Caltech Molecular Observatory solved and refined the crystal structure of HG-1 and Gert Kiss of the Houk lab (UCLA) carried out the molecular dynamics simulations.

Abstract

A general method for the *de novo* design of enzymatic activity has long been a major goal in computational protein design. Previous *de novo* enzyme design efforts placed little, if any, emphasis on the analysis of inactive designs, and the highly active designed enzymes that were recently reported were identified through the synthesis and screening of a large number of designs. Here, we describe an iterative method for the computational design of enzymes that combines computational enzyme design techniques with molecular dynamics simulations and crystallographic analysis of inactive designs. We used these methods to design and analyze an inactive design for the catalysis of the Kemp elimination, which helped us to understand the causes of our design's inactivity, including excessive solvent exposure and active site flexibility. We then applied this information to adjust our design procedure for future rounds of design.

Introduction

In early incarnations of computational protein design methodology, the strategy for software development was put forth in terms of the so called "protein design cycle" (Figure 1-1) where the experimental evaluation of an initial design is used to inform the adjustment of the design process for further rounds of designs.[1,2] These steps would ideally be continued iteratively until the protein sequences predicted by the algorithm have the desired characteristics. However, there is little evidence that this strategy has actually been used outside of force-field parameterization purposes,[1,3-5] and most inactive proteins from subsequent design efforts were likely discarded without comment or further investigation into the cause of the inactivity. This is unfortunate because little information can be gained from such an approach to protein design. Without detailed analysis of failed designs, flaws in the design procedure cannot be identified and adjusted to produce designed proteins with the desired characteristics.

Despite the lack of the type of systematic approach described above, the field of computational enzyme design has seen some success in the past ten years. The first successful computational introduction of *de novo* enzymatic activity into an inert scaffold was a protozyme, which catalyzed the hydrolysis of an activated ester with a very small rate enhancement (k_{cat}/k_{uncat} of about 10^2).[6] Additional incremental advances were made with the design of metalloenzymes and computationally designed changes in enzyme specificity.[7-9] In 2008, a breakthrough in *de novo* computational enzyme design came when the Rosetta protein design software was used to introduce catalytic activity for two distinct chemical reactions into a variety of inert scaffolds.[10,11] The reaction rates of these designed enzymes are impressive, but still well below those of natural enzymes.

Directed evolution was subsequently used to substantially increase the k_{cat}/K_m and the rate enhancement of one of the active designs.[11]

The synthesis and experimental evaluation of each potential enzyme is expensive both in terms of time and money. In the case of the enzymes designed by Röthlisberger *et al.*, over 70 individual designs, which were predicted to be active by their protein design methodology, were screened in order to identify 8 active enzymes.[11] While their methodology was successful in terms of producing active enzymes, the necessity of a "shotgun" approach suggests an incomplete understanding of some details of the enzymatic system and/or inaccurate modeling by the protein design algorithm.

Here, we focused on the development of a single design to rigorously test our understanding of enzymatic catalysis and the applicability of the protein design cycle to *de novo* enzyme design problems. To this end, we used our protein design software, ORBIT, in an attempt to design a catalyst for a model chemical system, the Kemp elimination (Figure 1-2). These methods resulted in a design called HG-1, which showed no Kemp elimination activity. In collaboration with the Houk lab at UCLA and the Molecular Observatory at Caltech, molecular dynamics simulations and crystallographic techniques were used to identify the possible sources of the inactivity, and a plan for future redesigns of the protein to create an active Kemp elimination enzyme was outlined.

Materials and Methods

Transitions state (TS) structure

Our TS structure was previously calculated *ab initio*[12] using the partial atomic charges used listed in Table 3-1. To assist with defining geometric constraints in the calculations, a pseudoatom named PSA with no charge and a negligible radius was added 0.01 Å from H3 in the plane of the TS. The final TS structure is shown in Figure 3-1 along with the atom names.

Scaffold selection

For the initial design, the xylanase from the thermophilic fungus *T. aurantiacus* (TAX) was chosen as the scaffold. TAX has a $(\alpha/\beta)_8$-barrel fold with a large, solvent-exposed active site. The crystal structure (PDB code: 1GOR) is 1.7 Å resolution with xylobiose bound in the active site.[13] The structure from the PDB was used without minimization and hydrogens were added with Molprobity.[14]

Active site search

All non-Pro positions within 5 Å of xylobiose in 1GOR were designated as design positions (residues 50, 83, 84, 87, 90, 130, 172, 207, 209, 267, and 275). Catalytic positions were defined as all of the carboxylate residues within 5 Å of xylobiose in 1GOR (E46, E131, and E237). All design positions were allowed to sample all conformations of Gly, His, Phe, Trp, Ser, Thr, and Tyr. Catalytic positions were allowed to sample all conformations of Glu and Asp. A backbone-independent conformer library was used to approximate sidechain flexibility.[15] As described in Chapter II, a library of

TS poses was generated in the active site by targeted ligand placement in which a defined set of ligand structures is built with respect to each conformation of the general base (Asp/Glu) at each catalytic position. The contact geometries that represent this set of ligand poses are listed in Table 3-2. After the generation of ligand poses, interaction energies were calculated, and TS-sidechain interaction energies were then biased to favor interactions that satisfy the geometric constraints in Table 3-3. An additional energy-biasing step was carried out to favor the hydrogen bond interactions to the base that were required (Table 3-4). The energy calculation and energy biasing steps are described in detail in Chapter II.

Active site repacking

For the repacking calculation, the initial TS position was taken from the results of the active site search. From this initial position, the TS structure was translated ± 0.2 Å in x, y, and z in 0.1 Å steps and rotated 5° in each direction in 2.5° steps. The geometric constraints from Tables 3-3 and 3-4 were applied to enforce the contacts between the TS and each of the three catalytic residues identified in the active site search (Y90, E237, and W275) as well as between E237, and H209. Positions 83 and 239, which are positioned such that they could be members of a potential hydrogen bond network with E237 were allowed to sample all conformations of all residues except for Cys and Pro. All other positions that clashed with the TS or catalytic residues including 21, 46, 87, 89, and 267 were allowed to sample all conformations of any non-polar residue.

Lazaridis-Karplus occlusion-based solvation[16] was applied with a scale factor of 1.0 for nonpolar burial and nonpolar exposure and a scale factor of 0.6 for polar burial.

Other standard ORBIT parameters were applied as in Lassila *et al.,* and a backbone-independent conformer library was used to represent sidechain flexibility.[15] As in the active site search, sidechain-TS interaction energies were biased to favor those contacts that satisfy the geometries in Tables 3-3 and 3-4. Sequence optimization was carried out with FASTER,[17,18] and a Monte Carlo-based algorithm[19,20] was used to sample sequences around the minimum energy conformation identified by FASTER (FMEC).

Hydrophobic active site repacking

This repacking calculation was similar to the previous active site repacking described above. All polar residues within the active site that were not directly contacting the base were designed and restricted to sampling conformations of hydrophobic residues only (Ala, Val, Leu, Ile, Phe, Tyr, Trp, and Met). Residues 46, 47, 84, 87, 89, 130, 131, 207, and 267 were allowed to sample all conformations of Ala, Val, Leu, Ile, Phe, Tyr, and Trp. In addition, residues 83, 21, and 50 were also allowed to sample conformations of Met.

Gene synthesis and cloning

The gene for TAX was back-translated from the protein sequence using the codon usage bias of *E. coli* in DNA 2.0.[21] The DNA sequence for a Factor Xa cleavage site and six-histidine purification tag were added to the 3′ end of the gene. Overlapping oligonucleotides spanning the gene sequence and flanking primers were designed using the Assembly PCR Oligo Maker web server.[22] The basic melting temperature calculation method was used along with a 50 mM cation concentration, a DNA concentration of 0.5

μM, maximum oligo length of 40, a 55°C annealing temperature, and a 40°C acceptable overlap T_m. The resulting oligonucleotides were 31-41 base pairs in length and the full-length gene was constructed by recursive PCR of the overlapping oligonucleotides using a method based on the one described by Stemmer *et al.*[23]

A primer mix was made by combining 2 μL each of the oligonucleotide primers, which had been diluted to 10 μM in H_2O. The 50 μL gene assembly reactions contained 5 μL 10x KOD buffer (Applied Biosystems), 1 mM $MgCl_2$ (Applied Biosystems), 0.2 mM dNTP mix (Novagen), 2.5 U KOD hot start (Applied Biosystems), and 10 μL of the primer mix. Gene amplification was carried out in 50 μL reactions containing 1 μL of the assembly reaction, 5 μL 10x KOD buffer (Applied Biosystems), 1 mM $MgCl_2$ (Applied Biosystems), 0.2 mM dNTP mix (Novagen), 2.5 U KOD hot start (Applied Biosystems), and 1 μM of the forward and reverse flanking primers. The reactions were carried out on Mastercycler Personal Thermocycler (Eppendorf) using the temperature program described in Table 3-5.

After amplification, the reactions were purified with a QIAquik PCR Purification Kit (Qiagen) then digested with *Bam*HI and *Nde*I (New England Biolabs). The digested genes were then run on a 1% agarose gel, extracted and purified using a QIAquik Gel Extraction Kit (Qiagen), and ligated into similarly digested pET11a vector (Novagen).

Site-directed mutagenesis

Mutagenesis primers were designed using the rules outlined in the Site-Directed Mutagenesis Kit (Stratagene). Primer melting temperatures were calculated using Equation 3-1:

$$T_m = 81.5 + 0.41(\%GC) - \frac{675}{N} - \%mismatch \quad . \qquad \text{[Equation 3-1]}$$

Mutagenesis primers were phosphorylated in 50 μL reactions containing 5 μL 10x T4-ligase buffer (New England Biolabs), 4 μg primer, and 10 units T4-polynucleotide kinase (New England Biolabs). Phosphorylation reactions were incubated at 37°C for 2 hours. Mutants were created using a protocol modified from the QuickChange Multi Site-Directed Mutagenesis Kit (Stratagene). Each 50 μL site-directed mutagenesis reactions contained 50 ng template, 120 ng of each phosphorylated mutagenesis primer, 3 μL each 10x Pfu Turbo buffer, 3 μL 10 x Taq ligase buffer (New England Biolabs), 2 mM dNTP mix (Stratagene), 2.5 U Pfu Turbo (Stratagene), and 40 U Taq DNA ligase (New England Biolabs). PCR was carried out on Mastercycler Personal Thermocycler (Eppendorf) using the temperature program described in Table 3-5.

Protein expression and purification

TAX and HG-1 were expressed in BL-21 (DE3) *E. coli* cells in LB with ampicillin. 30 mL starter cultures were grown overnight at 37°C and then used to innoculate 1L cultures. The cells were grown with shaking at 37°C until OD_{600} ~ 0.3 and were then grown at 25°C until OD_{600} ~ 0.6. Expression was induced with 1 mM isopropyl β-D-1-thiogalactopyranoside (IPTG) and carried out at 25°C for 12-18 hours. Cells were harvested by centrifugation and were resuspended in 30 mL buffer 1 (20 mM Tris pH 7.4, 300 mM NaCl, 10 mM imidazole). 10 mM $MgCl_2$, ribonuclease A, and deoxyribonuclease I (Stratagene) and the cells were lysed mechanically with an Emulsiflex-C5 (Avestin). The soluble fraction was incubated for ~ 1 hour with 3 mL Ni-

NTA resin (Qiagen) that had been equilibrated with 10 bed volumes of buffer 1. The resin was washed in a gravity column with 10 bed volumes of buffer 2 (20 mM Tris pH 7.4, 300 mM NaCl, 20 mM imidazole) and HG-1 was eluted with 3 bed volumes of buffer 3 (20 mM Tris pH 7.4, 300 mM NaCl, 250 mM imidazole). The eluate was concentrated using Amicon 10,000 MWCO centrifugal concentrators (Millipore) and the buffer was exchanged to 50 mM sodium citrate, 150 mM NaCl pH 5.5 using PD-10 desalting columns (GE).

Protein characterization

The molecular weights of all proteins were confirmed by electrospray mass spectrometry at the Protein/Peptide MicroAnalytical Laboratory (PPMAL) at Caltech. The secondary structure and stability of the proteins were characterized by circular dichrosim (CD) spectroscopy with an Aviv 62DS spectrometer equipped with a thermoelectric temperature controller. All CD experiments were carried out in a 1 mm cuvette with 10 μM protein in 50 mM MES pH 5.5, 100 mM NaCl, or 50 mM potassium phosphate, 100 mM NaCl pH 7.5. Wavelength scans were performed at 25°C with 3 sec of averaging at each point. Thermal denaturation was monitored at 222 nm from 4°C to 98°C with 1°C temperature steps, 2 min of temperature equilibration time, and 30 sec of averaging. Apparent midpoints of thermal denaturation (T_m) were obtained using the equation of Minor and Kim.[24] T_m values should be considered approximate and not actual thermodynamic parameters, as the denaturation was not reversible in any case tested.

Substrate synthesis and purification

5-nitrobenzisoxazole was synthesized as described.[25] Briefly, 3.5 g benzisoxazole (Alfa Aesar) was added to 15 mL concentrated H_2SO_4 cooled in an ice water bath. 2 mL of concentrated HNO_3 was added dropwise to the mixture. Nitrification was carried out for 30 min with stirring at 0°C. The reaction mixture was poured into 100 mL ice water and the yellow precipitate was filtered. The solid was recrystalized three times from 60 mL of warm absolute ethanol yielding long, colorless needles. m.p.: 126 °C (lit.[25] m.p. 126-127°C).

IR ($CDCl_3$): ν =1620 (arom);1520, 1352 (NO_2) cm^{-1}

1H-NMR ($CDCl_3$): δ = 7.75 (d, 1H); 8.5 (dd, 1H); 8.75 (d, 1H); 8.9 (d, 1H) ppm

MS: m/z = 163.1 (100%, M^-), 133.0 (47%, M^- - NO), 117.9 (7%. M^- - NO_2)

Nuclear magnetic resonance (NMR) analysis was performed on a Varian Mercury 300 MHz machine and IR spectra were recorded using a Perkin-Elmer Spectrum BX spectrometer. Electrospray mass spectrometry was performed at PPMAL (Caltech).

Kinetic measurements

The production of the phenolate product (Figure 1-2) was monitored at 405 nm using a Shimadzu UV-1601 spectrophotometer equipped with a temperature-controlled cell holder. A 40 mM stock of 5-nitrobensoxazole was made in acetonitrile and stored at -20°C. Assays were carried out with 250 μM protein in 40 mM buffer, 100 mM NaCl at 20°C and 37°C.

Crystallization conditions

Crystallization conditions were set up by hand in 24-well hanging drop plates using 2.0 × 2.0 μL drops and Index, Crystal 1, and Crystal 2 screens (Hampton Research) as well as a variety of custom screens using conditions similar to the crystallization conditions for 1GOR.[13] The protein sample contained 10 mg/mL HG-1 in 50 mM sodium citrate pH 5.5, 100 mM NaCl. A single large crystal (~ 0.5 mm) with multiple protrusions was found in buffer containing 0.8 M posassium-sodium tartrate, 0.1 M Tris pH 8.5, 0.5% PEG-MME 5000. A smaller single crystal (~ 0.1 mm) was identified in 100 mM sodium phosphate pH 6.5, 12% PEG-600.

X-ray diffraction data collection and refinement

The spikes projecting off the 0.5 mm crystal were broken off for diffraction screening. The best diffraction was obtained from a crystal cryoprotected in the reservoir solution (0.8 M posassium-sodium tartrate, 0.1 M Tris pH 8.5, 0.5% PEG-MME 5000) with 15% ethylene glycol.

Data were collected using a MicroMax-007HF X-ray generator with a RAXIS IV++ detector (Rigaku Corp.) All data were processed using CrystalClear (Rigaku Corp.) and Mosflm.[26] The indexed and scaled data were further evaluated using CCP4.[27] Molecular replacement was carried out using Phaser version 1.3.3[28] and 1GOR as the starting model.[13] Further refinement was performed with Coot[29] and Refmac.[30]

Molecular dynamics (MD) simulations

MD simulations were carried out for 20 ns for each enzyme-substrate complex at

NPT conditions (constant number of particles, pressure, and temperature) with a pressure of 1 bar and temperature of 300 K. The TIP3P explicit solvent model[31] was used to solvate the protein in an octahedral-shaped volume, ensuring a solvent layer of at least 10 Å from any point of the protein surface. Each face of the octahedral box is connected to a mirror image of itself, which allows the system to be treated with periodic boundary conditions: when a water molecule "escapes" the original octahedral box, it re-enters from the opposite face of that same box. This ensures equilibration through diffusion, while the number of particles is kept constant. The temperature is regulated and evenly distributed through a Langevin equilibration scheme. The benzisoxazole substrate was parameterized from quantum mechanical calculations in order to be treated correctly by the AMBER force field.[32] Prior to production MD, the geometry of the X-ray-based structure was optimized and slowly heated to 300 K over a time period of 300 ps at NVT conditions, and then equilibrated at NPT conditions for 2 ns.

Results and Discussion

Design procedure

TS theory suggests that natural enzymes achieve their large rate accelerations through preferential stabilization of the TS.[33] Thus, our computational enzyme design strategy is based on a TS or high-energy intermediate structure and the active site residues are chosen for their ability to stabilize the TS as well as participate in the reaction chemistry.[15]

As mentioned in Chapter I, our computational protein design software was optimized for stabilizing small proteins and required adaptation to the unique requirements of enzymatic catalysis.[15] Specifically, the interactions that are predicted to stabilize the TS were to be enforced through geometric requirements and the positioning of the TS was varied systematically within the binding pocket during the sequence optimization. The details of this design methodology are described in Chapter II.

The catalytic contacts that were predicted to stabilize the Kemp elimination (KE) TS are shown in Figure 3-2. These contacts were inspired by the catalytic residues observed in the catalytic antibody 34E4 (Figure 1-3). A carboxylate-containing residue was chosen as a general base in analogy to the Glu that was determined to be the general base in the catalytic antibody 34E4.[34] In addition, a carboxylate base avoids problems with the protonation state variations of His and the large number of rotameric states of Lys. A π-stacking residue was required to facilitate electron delocalization within the TS. A hydroxyl group was added to stabilize the negative charge that develops on the benzisoxazole oxygen in the TS.[12] No analogous hydrogen bond donor is observed in the catalytic antibody, although water is postulated to fill this role.[34] During the course of this study, a similar active site arrangement was used by Röthlisberger *et al.* to obtain active KE enzymes.[11]

In addition to the direct contacts to the TS, a supporting contact to the general base was also required to stabilize its conformation. As shown in Figure 3-3, this contact is a hydrogen bond to the general base from a His (Hie and Hid are His residues, which are protonated at Nε or Nδ, respectively), Glu, or Asp residue. This contact was required

through a second, separate geometry-biasing step. The geometric definitions for the three catalytic contacts and the additional base contact are shown in Table 3-2.

Active site search

The active site search resulted in multiple possible active site configurations that satisfied all four of the required contacts. The active sites were assessed visually and active site 1-2 was chosen for repacking (Figure 3-4A). In 1-2, the general base is a wild-type Glu at position 237 that is in a conformation similar to that seen in the crystal structure. The base is held in place by a wild-type His at position 209 that preserves a contact seen in the crystal structure of the wild-type TAX. A wild-type Trp at position 275 makes a stacking contact to the TS and is in its crystallographic conformation. Finally, the hydrogen bond contact is fulfilled by a Tyr at position 90.

Active site repacking

The active site repacking calculation resulted in a seven-fold mutant called HG-1 (Figure 3-4B). This design features an extensive hydrogen bond network supporting the conformation of the base, E237, which is similar to the hydrogen bond network surrounding E237 in the wild-type protein. In the crystal structure of TAX, His209 makes a bridging contact between Glu237 and Asp239, indicating that it is in its fully protonated state which would also serve to stabilize the negative charge on both carboxylates. As a result, the pKa of Glu237 is probably quite low. This bridging contact is part of a larger hydrogen bond network that includes the indole of Trp275. One difference between the hydrogen bond network in the design and the one in TAX is

the D127N mutation, which should help to ensure that H209 is not positively charged, thus stabilizing the protonated form of E237 and increasing its pKa. The pKa of E237 was predicted to be 7.1 by PROPKA,[35] which may be sufficient for proton abstraction from the substrate.

Other mutations in the active site include W267A, W87G, A21M, E46M, and S89F (Figure 3-4B), which were introduced to make room in the pocket for substrate binding and to pack tightly around the substrate once it is bound.

HG-1 activity assays

HG-1 showed no KE activity over background under any of the conditions tested (pH 5.0-9.0, 20-37°C) (Figure 3-5). Analysis by circular dichroism (CD) spectroscopy showed that the protein had similar secondary structure as the wild-type scaffold, indicating that the inactivity was not due to global unfolding or misfolding of the protein (Figure 3-6A). The T_m for HG-1, also determined by CD, was about 10°C lower than the wild-type scaffold at pH 5.5 (Figure 3-6B). At pH 7.5, an additional 10°C decrease in melting temperature was seen compared to wild-type TAX at pH 5.5. However, the HG-1 was fully folded at all of the temperatures and pHs tested at which KE activity was assayed. Single and double mutants of HG-1 were also tested including N239D, N239D/M46A, and M46A. No rate acceleration was seen for any of these variants above the scaffold-catalyzed reaction.

HG-1 crystal structure

In an effort to understand the source of the inactivity of HG-1, we crystallized and solved the crystal structure of HG-1 to a resolution of 2.0 Å in collaboration with the Molecular Observatory at Caltech (Table 3-6). The overall RMSD of HG-1 with respect to the scaffold is 0.36 Å, confirming that the overall fold has been maintained (Figure 3-7A).

The structure fits well to the electron density as shown in Figure 3-7B. One concern with using molecular replacement (MR) to calculate the phases is potential bias from the template structure 1GOR. As shown in Figure 3-8, the electron density around the mutated positions is well-defined and shows no signs of the wild-type amino acid, indicating little bias.

The active site residues of HG-1, including the base, the hydrogen bond donor, and most of the hydrogen bond network, overlay well with the conformations predicted by ORBIT (Figure 3-9). The largest deviations of the active site residues from the design are in the π-stacking residue (W275), the associated Arg (R276) and N172, which acts as a hydrogen bond bridge between the indole of W275 and H209 (Figure 3-9B). W275 and R276 are both rotated out of the active site with respect to the design, removing potential binding interactions with the substrate and exposing the active site to solvent. In addition, N172 is rotated 90° from the predicted conformation, but it is unclear if this deviation is the cause for the conformation of W275.

The crystal structure revealed the presence of six ordered water molecules in the active site of HG-1 (Figure 3-10), five of which must be displaced before substrate binding could occur. The presence of these waters could result in a large desolvation

penalty, which could prevent substrate binding as well as a reduction the pKa of the base through stabilization of the anionic form.

Overall, the crystal structure indicates that the active site configuration predicted by ORBIT is correct with the exception of N172, W275, and R276. The flexibility of the active site, indicated by the deviation of W275 and the associated R276, could result in the loss of favorable binding interactions. The movement of these residues could also expose the active site to solvent, further decreasing the energetic benefit of substrate binding. Unfortunately, the crystal structure could not unambiguously identify the cause of the protein's inactivity, but these two factors could be significant impediments to activity.

MD simulations

The hypothesis of active site flexibility and solvent exposure being the cause of the inactivity was supported by MD simulations carried out by Gert Kiss in the Houk lab at UCLA. The design model of HG-1 with the substrate bound in the active site was used as the starting structure of the simulations. Figure 3-11 shows that the substrate-protein complex exists predominantly in two states. In state 1, the substrate is stably bound and maintains a substrate-base distance of about 3 Å. In state 2, a major reorientation of the substrate occurs and the substrate exits the active site. Closer analysis of the active site residues during the transition between state 1 and state 2 indicates that within 2 ns, R276 flips away from the active site, followed by W275 at 5 ns. The loss of these binding interactions seems to directly contribute to the reorientation and subsequent expulsion of the substrate from the active site around 7 ns. In addition, solvent molecules moving

rapidly into and out of the active site preferentially interact with the base and other polar residues causing significant destabilization of the bound substrate, actually pushing it from the active site. These results suggesting that the active site is very flexible near W275 and R276 and is occupied by a large number of solvent molecules agree with findings from the crystal structure of HG-1.

Design of a more hydrophobic active site

The more hydrophobic design of HG-1 resulted in the sequence HG-1h, an eight-fold mutant with respect to HG-1 (N47F, K50M, M64A, H83V, T84F, F89L, N130F, and E131A). Figure 3-12 shows that the overall conformation of the active site in HG-1h, including the TS and catalytic residues, is very similar to that in HG-1. The pKa of E239 in HG-1h was predicted to be about 1.0 pH unit higher than in HG-1. MD analysis of this design shows that as in HG-1, the active site is very flexible and allows the substrate to exit the active site during the simulation. The active site is still very solvent exposed, and solvent molecules still enter the active site over the course of the simulation to interact with the remaining polar residues and crowd the substrate out of the active site. Figure 3-13A shows representative configurations of HG-1h before and after the MD simulation. In the final configuration, the substrate-base distance is over 10 Å and the substrate-base distances shown in Figure 3-13B and 3-13C indicate that the contact is not maintained for any significant length of time. These results suggest that our redesign is not likely to be active for many of the same reasons as HG-1.

Conclusions

While we were unsuccessful in designing an active KE enzyme, we were able to solve the crystal structure of one of our inactive designs, allowing us to gain further insight into the cause of the inactivity. Unfortunately, neither the crystal structure nor the MD simulations of HG-1 were able to unambiguously determine the cause of HG-1's inactivity. However, both analyses indicated that the active site is very solvent exposed, flexible, and occupied by a large number of water molecules. While it is not certain that these factors are the cause of the inactivity, any one of them could have a negative effect on substrate binding as well as on the base pKa and all three issues could be addressed by choosing an active site that is more deeply buried within a protein scaffold. The natural active site of TAX, which was the focus of the designs presented here, was too solvent exposed and flexible to be amenable to this reaction. Future redesigns of this scaffold must locate the active site away from the natural binding pocket in a location that is more deeply buried, perhaps in the barrel of the scaffold. Other scaffolds with less solvent-exposed active sites should also be explored. A more deeply buried active site could effectively reduce the flexibility of the binding residues and shield the active site from solvent, though it remains to be seen if the resolution of these factors alone will result in an active KE enzyme.

Acknowledgements

We thank Paul Cheong from the Houk lab at UCLA for providing the coordinates and charges for the KE TS. Gert Kiss from the Houk lab at UCLA carried out the MD simulations and created Figure 3-11. Dr. Leonard Thomas of the Caltech Molecular

Observatory solved and refined the crystal structure of HG-1. We also thank Dr. Jonathan Lassila and Dr. Roberto Chica for discussions about enzymes and enzyme kinetics.

References

1. Dahiyat, B. I.; Mayo, S. L., Protein design automation. *Protein Sci.* **1996**, *5*, 895-903.

2. Dahiyat, B. I.; Mayo, S. L., De novo protein design: fully automated sequence selection. *Science* **1997**, *278*, 82-87.

3. Dahiyat, B. I.; Mayo, S. L., Probing the role of packing specificity in protein design. *Proc. Natl. Acad. Sci. USA* **1997**, *94*, 10172-10177.

4. Dahiyat, B. I.; Gordon, B.; Mayo, S. L., Automated design of the surface positions of protein helices. *Protein Sci.* **1997**, *6*, 1333-1337.

5. Zollars, E. S.; Marshall, S. A.; Mayo, S. L., Simple electrostatic model improves designed protein sequences. *Protein Sci.* **2006**, *15*, 2014-2018.

6. Bolon, D. N.; Mayo, S. L., Enzyme-like proteins by computational protein design. *Proc. Natl. Acad. Sci. USA* **2001**, *98*, 14274-14279.

7. Kaplan, J.; DeGrado, W. F., *De novo* design of catalytic proteins. *Proc. Natl. Acad. Sci. USA* **2004**, *101*, 11566-11570.

8. Ashworth, J.; Havranek, J. J.; Duarte, C. M.; Sussman, D.; Monnat, R. J.; Stoddard, B. L.; Baker, D., Computational redesign of endonuclease DNA binding and cleavage specificity. *Nature* **2006**, *441*, 656-659.

9. Chen, C. Y.; Georgiev, I.; Anderson, A. C.; Donald, B. R., Computational structure-based redesign of enzyme activity. *Proc. Natl. Acad. Sci. USA* **2009**.

10. Jiang, L.; Althoff, E. A.; Clemente, F. R.; Doyle, L.; Röthlisberger, D.; Zanghellini, A.; Gallaher, J. L.; Betker, J. L.; Tanaka, F.; Barbas, C. F.; Hilvert, D.; Houk, K. N.; Stoddard, B. L.; Baker, D., De novo computational design of retro-aldol enzymes. *Science* **2008**, *319*, 1387-1391.

11. Rothlisberger, D.; Khersonsky, O.; Wollacott, A.; Jiang, L.; Dechancie, J.; Betker, J.; Gallaher, J. L.; Althoff, E. A.; Zanghellini, A.; Dym, O.; Albeck, S.; Houk, K. N.; Tawfik, D.; Baker, D., Kemp elimination catalysts by computational enzyme design. *Nature* **2008**, *453*, 190-195.

12. Na, J.; Houk, K. N.; Hilvert, D., Transition state of the base-promoted ring-opening of isoxazoles. Theoretical prediction of catalytic functionalities and design of haptens for antibody production. *J. Am. Chem. Soc.* **1996**, *118*, 6462-6471.

13. Lo Leggio, L.; Kalogiannis, S.; Eckert, K.; Teixeira, S. C.; Bhat, M. K.; Andrei, C.; Pickersgill, R. W.; Larsen, S., Substrate specificity and subsite mobility in *T. aurantiacus* xylanase 10A. *FEBS Lett.* **2001**, *509*, 303-308.

14. Davis, I. W.; Murray, L. W.; Richardson, J. S.; Richardson, D. C., MolProbity: all-atom contacts and structure validation for proteins and nucleic acids. *Nucleic Acids Res.* **2007**, *35*, W375-W383.

15. Lassila, J. K.; Privett, H. K.; Allen, B. D.; Mayo, S. L., Combinatorial methods for small-molecule placement in computational enzyme design. *Proc. Natl. Acad. Sci. USA* **2006**, *103*, 16710-16715.

16. Lazaridis, T.; Karplus, M., Effective energy function for proteins in solution. *Proteins* **1999**, *35*, 133-152.

17. Desmet, J.; Spriet, J.; Lasters, I., Fast and accurate side-chain topology and energy refinement (FASTER) as a new method for protein structure optimization. *Proteins* **2002**, *48*, 31-43.

18. Allen, B. D.; Mayo, S. L., Dramatic performance enhancements for the FASTER optimization algorithm. *J. Comput. Chem.* **2006**, *27*, 1071-1075.

19. Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., Equation of state calculations by fast computing machines. *J. Chem. Phys.* **1953**, *21*, 1087-1092.

20. Voigt, C. A.; Gordon, D. B.; Mayo, S. L., Trading accuracy for speed: A quantitative comparison of search algorithms in protein sequence design. *J. Mol. Biol.* **2000**, *299*, 789-803.

21. Villalobos, A.; Ness, J. E.; Gustafsson, C.; Minshull, J.; Govindarajan, S., Gene Designer: a synthetic biology tool for constructing artificial DNA segments. *BMC Bioinformatics* **2006**, *7*, 285-292.

22. Rydzanicz, R.; Zhao, X. S.; Johnson, P. E., Assembly PCR oligo maker: a tool for designing oligodeoxynucleotides for constructing long DNA molecules for RNA production. *Nucleic Acids Res.* **2005**, *33*, W521-525.

23. Stemmer, W. P.; Crameri, A.; Ha, K. D.; Brennan, T. M.; Heyneker, H. L., Single-step assembly of a gene and entire plasmid from large numbers of oligodeoxyribonucleotides. *Gene* **1995**, *164*, 49-53.

24. Minor, D. L.; Kim, P. S., Measurement of the bold β-sheet-forming propensities of amino acids. *Nature* **1994**, *367*, 660-663.

25. Lindemann, H.; Thiele, H., The chemistry of benzene-α, β-isoxazoles. *Justus Liebigs Annalen Der Chemie* **1926**, *76*, 63-81.

26. Leslie, A. G. W., Recent changes to the MOSFLM package for processing film and image plate data. *Joint CCP4 + ESF-EAMCB Newsletter on Prot. Crystallography* **1992**, *26*.

27. The CCP4 suite: programs for protein crystallography. *Acta Crystallogr., Sect. D* **1994**, *50*, 760-763.

28. McCoy, A. J.; Grosse-Kunstleve, R. W.; Storoni, L. C.; Read, R. J., Likelihood-enhanced fast translation functions. *Acta Crystallogr., Sect. D* **2005**, *61*, 458-464.

29. Emsley, P.; Cowtan, K., Coot: model-building tools for molecular graphics. *Acta Crystallogr., Sect. D* **2004**, *61*, 458-464.

30. Murshudov, G. N.; Vagin, A. A.; Dodson, E. J., Refinement of macromolecular structures by the maximum-likelihood method. *Acta Crystallogr., Sect. D* **1997**, *53*, 240-255.

31. Jorgensen, W. L.; Chandrasekhar, J.; Madura, J. D.; Impey, R. W.; Klein, M. L., Comparison of simple potential functions for simulating liquid water. *J. Chem. Phys.* **1983**, *79*, 926-935.

32. Pearlman, D. A.; Case, D. A.; Caldwell, J. W.; Ross, W. S.; Cheatham, T. E.; DeBolt, S.; Ferguson, D.; Seibel, G.; Kollman, P., AMBER, a package of computer programs for applying molecular mechanics, normal mode analysis, molecular dynamics and free energy calculations to simulate the structural and energetic properties of molecules. *Comput. Phys. Commun.* **1995**, *91*, 1-41.

33. Lienhard, G. E., Enzymatic catalysis and transition-state theory. *Science* **1973**, *180*, 149-154.

34. Debler, E. W.; Ito, S.; Seebeck, F. P.; Heine, A.; Hilvert, D.; Wilson, I. A., Structural origins of efficient proton abstraction from carbon by a catalytic antibody. *Proc. Natl. Acad. Sci. USA* **2005**, *102*, 4984-4989.

35. Li, H.; Robertson, A. D.; Jensen, J. H., Very fast empirical prediction and interpretation of protein pKa values. *Proteins* **2005**, *61*, 704-721.

Table 3-1. Partial atomic charges for 5-nitrobenzisoxazole (5-NBZ).

atom name	charge
O1	-0.586
N2	-0.212
C3	0.041
C4	-0.252
C5	-0.148
C6	0.099
N6	0.517
O61	-0.484
O62	-0.506
C7	-0.168
C8	-0.269
C9	0.516
H3	0.490
H5	0.275
H7	0.251
H8	0.212

Table 3-2. Variation of contact geometry for targeted ligand placement. The TS was placed at every indicated geometry for every rotamer of Asp or Glu in the design calculation. Because Asp and Glu both have two equivalent, but distinctly named, oxygens in their sidechains, a separate set of geometries is described for each oxygen.

Placement: H3 from Aspartate (OD1)

type	atom1	atom2	atom3	atom4	min	max	step
DISTANCE	OD1	H3			1.1	1.5	0.2
ANGLE	OD2	OD1	H3		70.0	110.0	20.0
TORSION	CG	OD2	OD1	H3	160.0	200.0	20.0
ANGLE	OD1	H3	PSA		70.0	110.0	20.0
TORSION	OD2	OD1	H3	PSA	0.0	360.0	20.0
TORSION	OD1	H3	PSA	C3	160.0	200.0	20.0

Placement: H3 from Aspartate (OD2)

type	atom1	atom2	atom3	atom4	min	max	step
DISTANCE	OD2	H3			1.1	1.5	0.2
ANGLE	OD1	OD2	H3		70.0	110.0	20.0
TORSION	CG	OD1	OD2	H3	160.0	200.0	20.0
ANGLE	OD2	H3	PSA		70.0	110.0	20.0
TORSION	OD1	OD2	H3	PSA	0.0	360.0	20.0
TORSION	OD2	H3	PSA	C3	160.0	200.0	20.0

Placement: H3 from Glutamate (OE1)

type	atom1	atom2	atom3	atom4	min	max	step
DISTANCE	OE1	H3			1.1	1.5	0.2
ANGLE	OE2	OE1	H3		70.0	110.0	20.0
TORSION	CD	OE2	OE1	H3	160.0	200.0	20.0
ANGLE	OE1	H3	PSA		70.0	110.0	20.0
TORSION	OE2	OE1	H3	PSA	0.0	360.0	20.0
TORSION	OE1	H3	PSA	C3	160.0	200.0	20.0

Placement: H3 from Glutamate (OE2)

type	atom1	atom2	atom3	atom4	min	max	step
DISTANCE	OE2	H3			1.1	1.5	0.2
ANGLE	OE1	OE2	H3		70.0	110.0	20.0
TORSION	CD	OE1	OE2	H3	160.0	200.0	20.0
ANGLE	OE2	H3	PSA		70.0	110.0	20.0
TORSION	OE1	OE2	H3	PSA	0.0	360.0	20.0
TORSION	OE2	H3	PSA	C3	160.0	200.0	20.0

Table 3-3. Geometric constraints contacts between the active site residues and the transition state in the active site search. Three contacts are required: Asp/Glu, Phe/Trp and Ser/Thr/Tyr. Distance measurements are given in Ångströms. Angle, torsion, and plane measurements are given in degrees.

Contact: Asp/Glu to H3

residue	type	atom1	atom2	atom3	atom4	min	max
Asp	DISTANCE	OD1	H3			1.0	1.6
	ANGLE	OD2	OD1	H3		69.0	111.0
	TORSION	CG	OD2	OD1	H3	159.0	201.0
	ANGLE	OD1	H3	PSA		69.0	111.0
	TORSION	OD2	OD1	H3	PSA	0.0	360.0
	TORSION	OD1	H3	PSA	C3	159.0	201.0
Asp	DISTANCE	OD2	H3			1.0	1.6
	ANGLE	OD1	OD2	H3		69.0	111.0
	TORSION	CG	OD1	OD2	H3	159.0	201.0
	ANGLE	OD2	H3	PSA		69.0	111.0
	TORSION	OD1	OD2	H3	PSA	0.0	360.0
	TORSION	OD2	H3	PSA	C3	159.0	201.0
Glu	DISTANCE	OE1	H3			1.0	1.6
	ANGLE	OE2	OE1	H3		69.0	111.0
	TORSION	CD	OE2	OE1	H3	159.0	201.0
	ANGLE	OE1	H3	PSA		69.0	111.0
	TORSION	OE2	OE1	H3	PSA	0.0	360.0
	TORSION	OE1	H3	PSA	C3	159.0	201.0
Glu	DISTANCE	OE2	H3			1.0	1.6
	ANGLE	OE1	OE2	H3		69.0	111.0
	TORSION	CD	OE1	OE2	H3	159.0	201.0
	ANGLE	OE2	H3	PSA		69.0	111.0
	TORSION	OE1	OE2	H3	PSA	0.0	360.0
	TORSION	OE2	H3	PSA	C3	159.0	201.0

Contact: Ser/Thr/Tyr to O1

residue	type	atom1	atom2	atom3	atom4	min	max
Ser	DISTANCE	OG	O1			2.6	4.0
	ANGLE	OG	HG	O1		150.0	180.0
	ANGLE	HG	O1	N2		100.0	160.0
	TORSION	HG	O2	N2	C3	120.0	240.0
Thr	DISTANCE	OG1	O1			2.6	4.0
	ANGLE	OG1	HG1	O1		150.0	180.0
	ANGLE	HG1	O1	N2		100.0	160.0
	TORSION	HG1	O2	N2	C3	120.0	240.0
Tyr	DISTANCE	OH	O1			2.6	4.0
	ANGLE	OH	HH	O1		150.0	180.0
	ANGLE	HH	O1	N2		100.0	160.0
	TORSION	HH	O2	N2	C3	120.0	240.0

Contact: Phe/Trp to PS2

residue	type	atom1	atom2	atom3	atom4	atom5	atom6	min	max
Phe	PSEUDO_ATOM	PS1, equidistant between CE1 and CD2							
	PSEUDO_ATOM	PS2, equidistant between C4 and C9							
	DISTANCE	PS1	PS2					3.0	4.0
	PLANE	CG	CE1	CE2	C5	C8	N2	0.0	40.0
Trp	PSEUDO_ATOM	PS1, equidistant between CE2 and CD2							
	PSEUDO_ATOM	PS2, equidistant between C4 and C9							
	DISTANCE	PS1	PS2					3.0	4.0
	PLANE	CD1	CE3	CH2	C5	C8	N2	0.0	40.0

Table 3-4. Geometric constraints for additional base contact. A single hydrogen bond contact was required to one of the carboxylate oxygens of the general base. Distances are given in Ångströms. Angles are given in degrees.

Contact: His to OD1/OD2

residue	type	atom1	atom2	atom3	min	max
Hie	DISTANCE	NE2	OD1		2.6	4.0
	ANGLE	NE2	HE2	OD1	130.0	180.0
	ANGLE	HE2	OD1	CG	120.0	180.0
Hie	DISTANCE	NE2	OD2		2.6	4.0
	ANGLE	NE2	HE2	OD2	130.0	180.0
	ANGLE	HE2	OD2	CG	120.0	180.0
Hid	DISTANCE	ND1	OD1		2.6	4.0
	ANGLE	ND1	HD1	OD1	130.0	180.0
	ANGLE	HD1	OD1	CG	120.0	180.0
Hid	DISTANCE	ND1	OD2		2.6	4.0
	ANGLE	ND1	HD1	OD2	130.0	180.0
	ANGLE	HD1	OD2	CG	120.0	180.0

Contact: His to OE1/OE2

residue	type	atom1	atom2	atom3	min	max
Hie	DISTANCE	NE2	OE1		2.6	4.0
	ANGLE	NE2	HE2	OE1	130.0	180.0
	ANGLE	HE2	OE1	CD	120.0	180.0
Hie	DISTANCE	NE2	OE2		2.6	4.0
	ANGLE	NE2	HE2	OE2	130.0	180.0
	ANGLE	HE2	OE2	CD	120.0	180.0
Hid	DISTANCE	ND1	OE1		2.6	4.0
	ANGLE	ND1	HD1	OE1	130.0	180.0
	ANGLE	HD1	OD1	CD	120.0	180.0
Hid	DISTANCE	ND1	OE2		2.6	4.0
	ANGLE	ND1	HD1	OE2	130.0	180.0
	ANGLE	HD1	OE2	CD	120.0	180.0

Table 3-5. Thermocycler temperature programs for gene construction and mutagenesis reactions.

gene assembly		
Temp (°C)	**Time (min)**	**# cycles**
95	1	
95	1	25x
55	1	
68	14	
4	hold	

amplification		
Temp (°C)	**Time (min)**	**# cycles**
95	1	
95	1	23x
55	1	
68	14	
4	hold	

site-directed mutagenesis		
Temp (°C)	**Time (min)**	**# cycles**
95	1	
95	1	30x
55	1	
68	14	
4	hold	

Table 3-6. Crystallographic statistics for HG-1.

Data collection	
Space group	$P2_12_12_1$
Cell dimensions	
a,b,c (Å)	48.3, 72.5, 74.6
α,β,γ (deg)	90.0, 90.0, 90.0
Resolution (Å) *	2.0 (2.1 - 2.0)
Number of reflections	48807
unique reflections	18096
Rmerge (%) *	2.5 (4.4)
I/σI *	30.7 (21.1)
Completeness (%) *	99.2 (98.2)
Redundancy *	2.7 (2.6)

Refinement	
Resolution (Å)	29.5 - 2.0
Number of reflections	
working set	17163
test set	924
R_{work}/R_{free}	16.4 / 23.0 %
No. atoms	
protein	2326
solvent	187
B-factors	
protein	13.9
water	19.7
R.m.s deviations	
bond lengths (Å)	0.027
bond angles	1.938
Ramachandran plot	
favored (%)	91.0
allowed (%)	8.6
generously allowed (%)	0.4
disallowed (%)	0

* Last shell is shown in parentheses

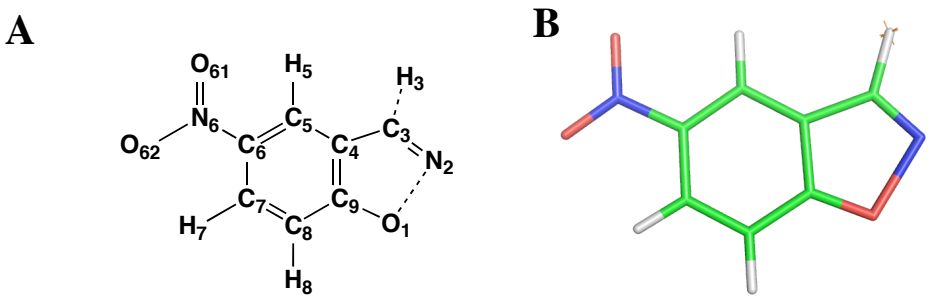

Figure 3-1. 5-NBX transition state. (A) 5-NBX atom names. The bonds being broken are shown as dotted lines. (B) Stick representation of the transition state structure. The pseudoatom is shown as an orange star near H3.

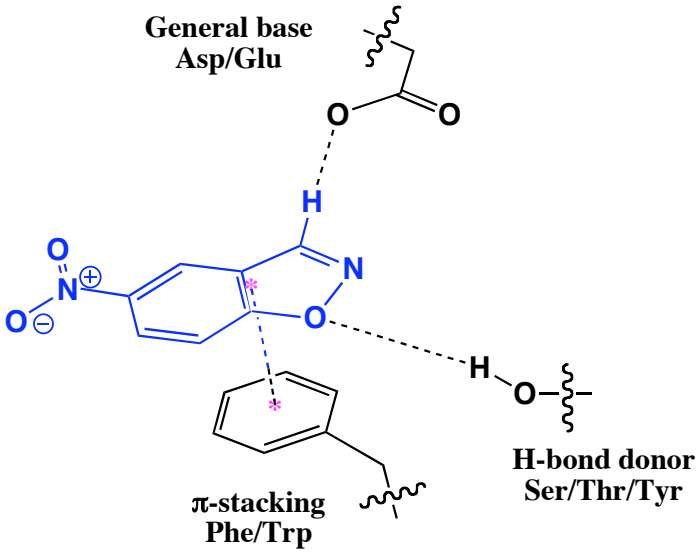

Figure 3-2. Kemp elimination ideal active site. The enzyme design calculations required either an Asp or Glu as a general base, a Ser, Thr, or Tyr as a H-bond donor to the phenolic oxygen, and a π-stacking residue above or below the plane of the TS structure (blue). The pseudoatoms that were used to define the π-stacking contact are shown in pink.

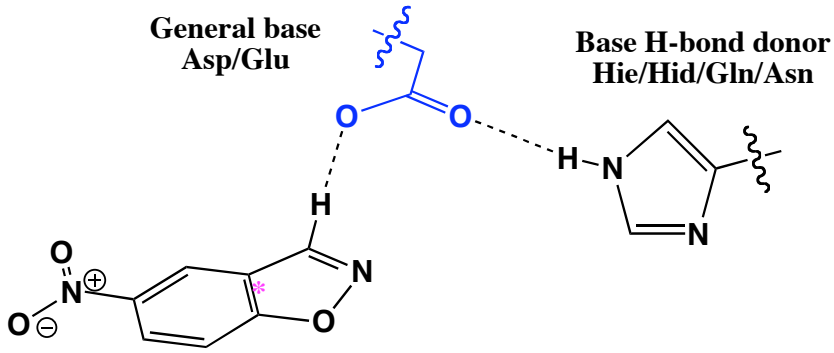

Figure 3-3. Additional hydrogen bond contact. An additional contact (Hie/Hid/Gln/Asn) was required to the base (shown in blue).

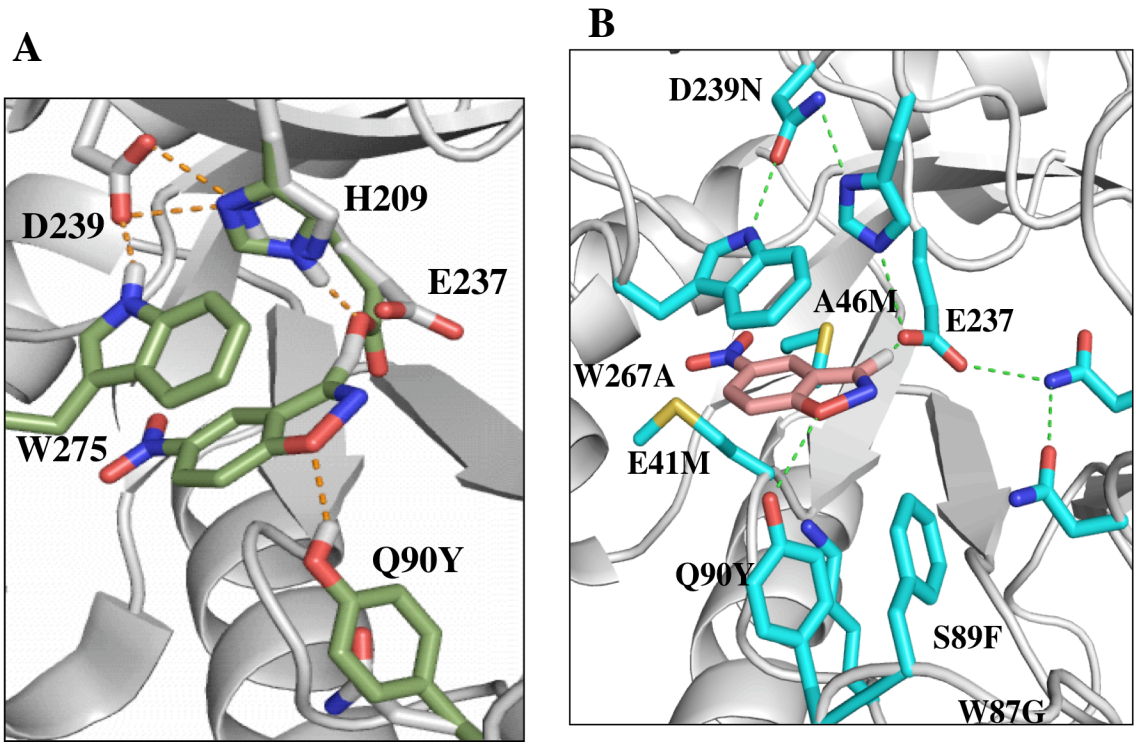

Figure 3-4. Predicted active site structure of HG-1. **(A)** Active site 2-1, identified through the active site search. The catalytic residues and TS are shown in green overlaid with the wild-type crystal structure in grey. **(B)** Repacked active site. The TS model is shown in pink and all seven mutations are identified. The hydrogen bond network supporting the conformation of the base is indicated with green dotted lines.

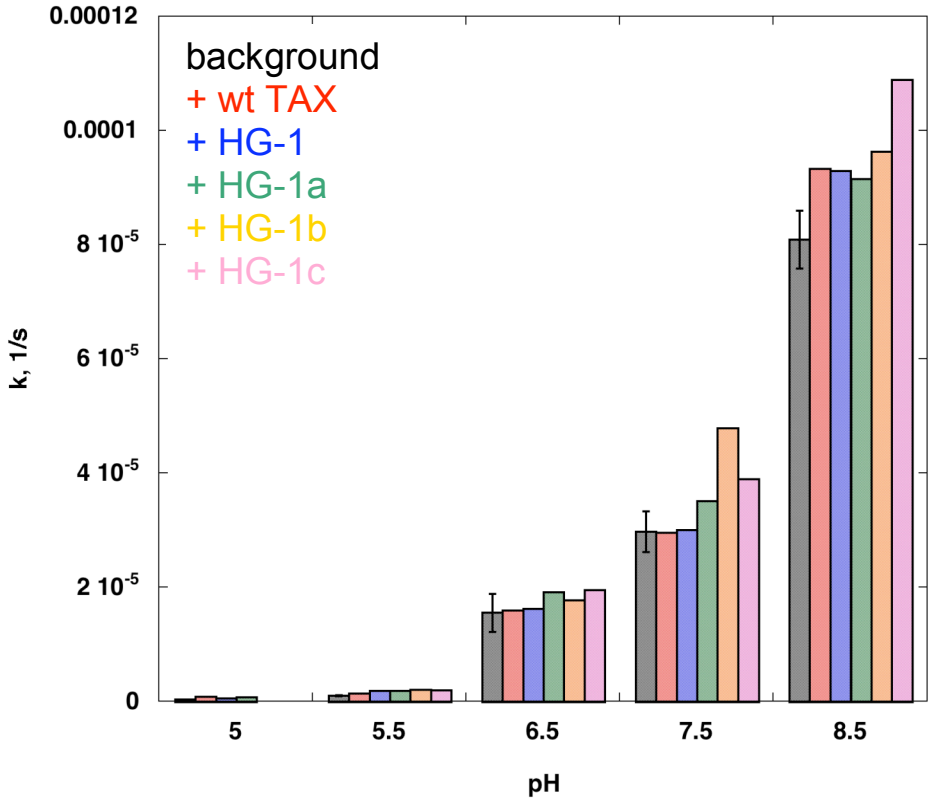

Figure 3-5. First-order rate constants for KE reaction. The rate constants for the reactions including 200 μM wild-type TAX, HG-1, HG-1a (HG-1/N239D), HG-1b (HG-1/N239D/ M46A), HG-1c (HG-1/M46A), and the buffer-catalyzed background reaction are given. No significant rate acceleration was seen above that for the scaffold (wild-type TAX)-catalyzed reaction for any of the variants assayed under any of the conditions tested. Sodium citrate was used for pH 5.0 and 5.5, sodium phosphate was used for pH 6.5 and 7.5, and sodium borate was used for pH 8.5.

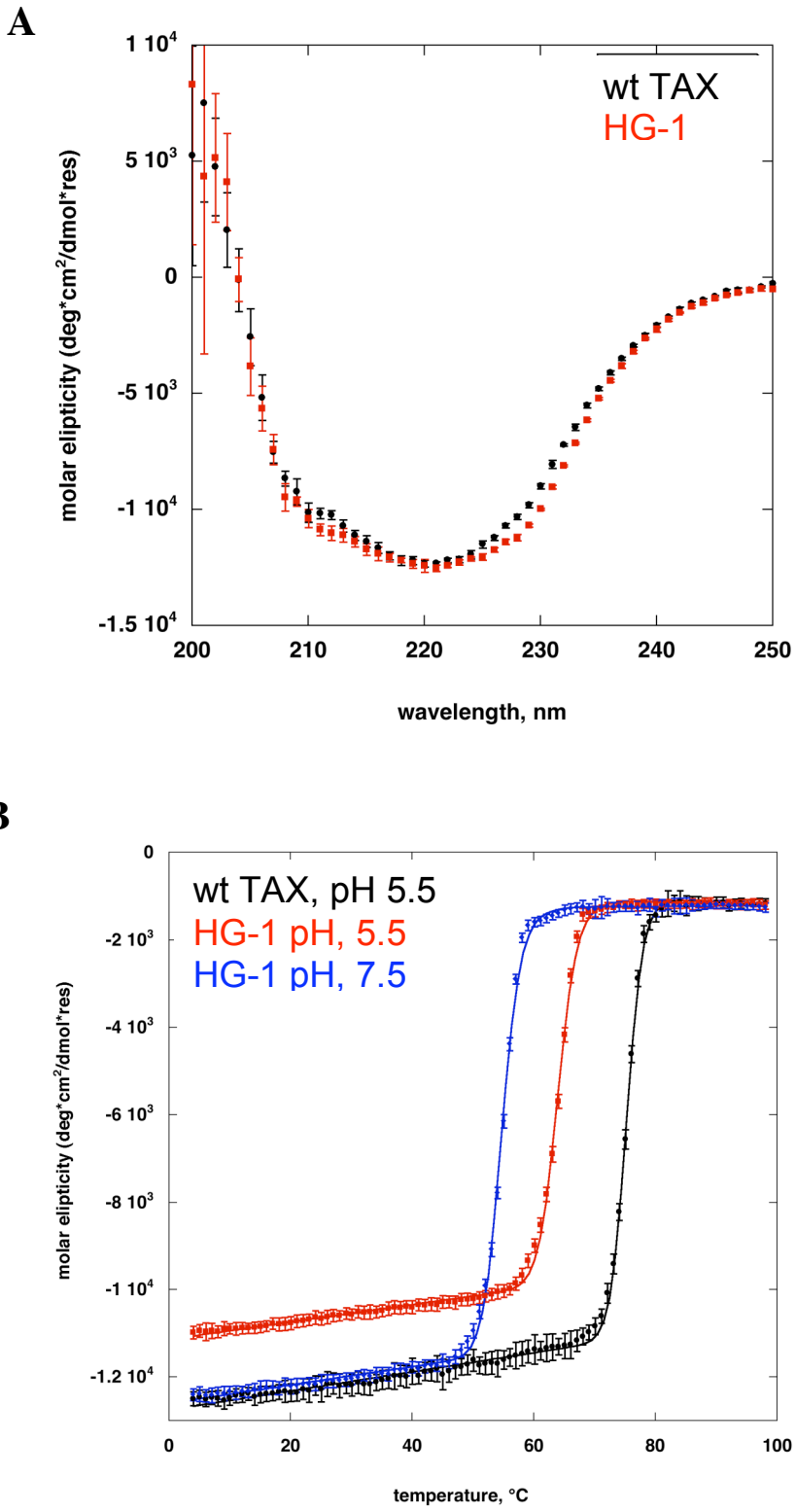

Figure 3-6. CD analysis of HG-1. **(A)** Far-UV wavelength scan indicating that HG-1 is folded. **(B)** Thermal denaturation curve indicating that HG-1 is significantly destabilized compared to TAX.

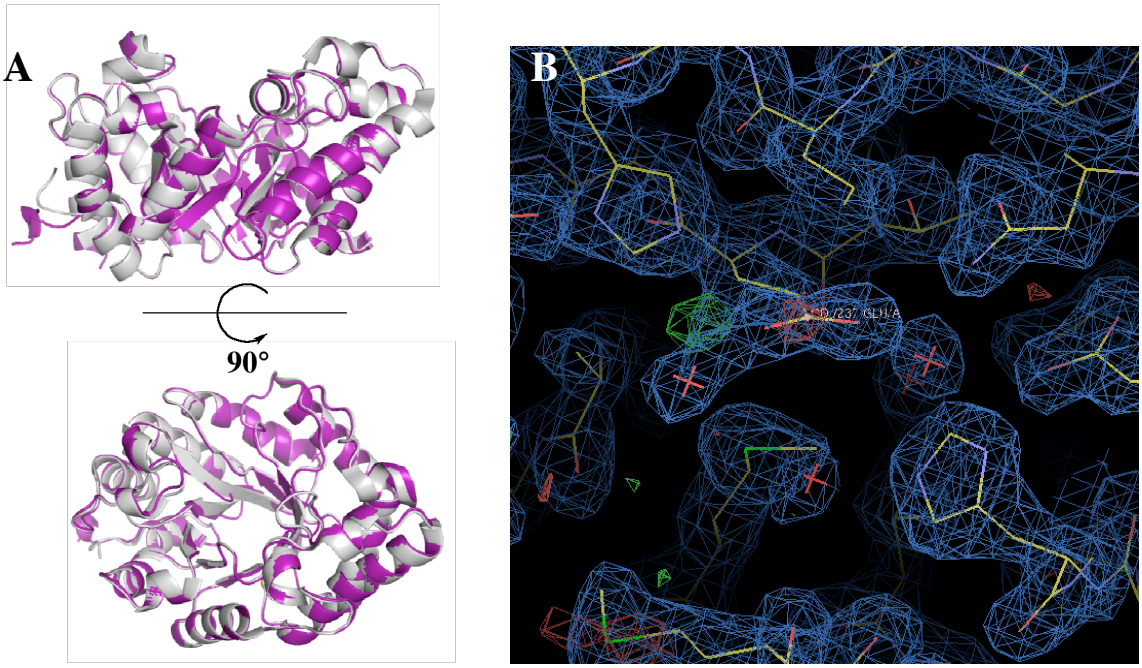

Figure 3-7. Crystal structure of HG-1. **(A)** Overlay of backbone structure of HG-1 (magenta) and wild-type TAX structure 1GOR (grey). **(B)** Electron density in active site of HG-1.

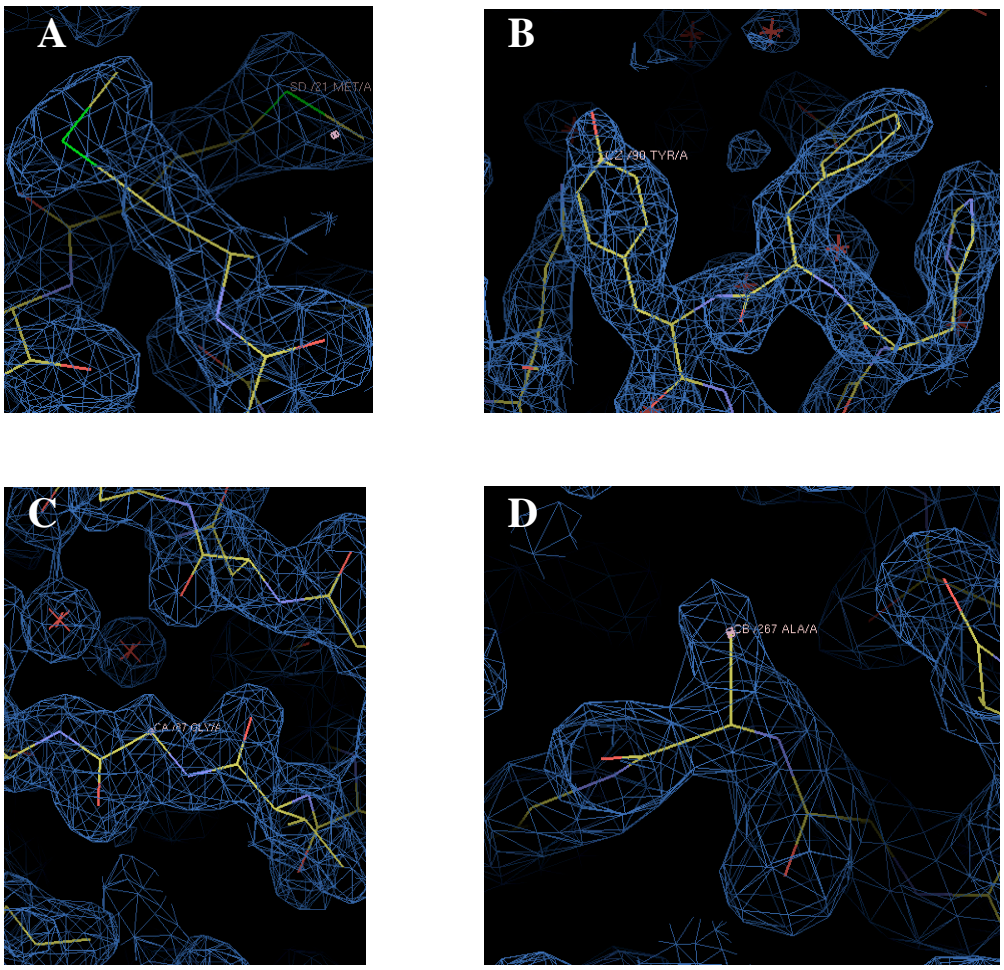

Figure 3-8. Electron density of HG-1 mutation sites. **(A)** A24M/E46M. **(B)** S89F/Q90Y. **(C)** W87G. **(D)** W267A.

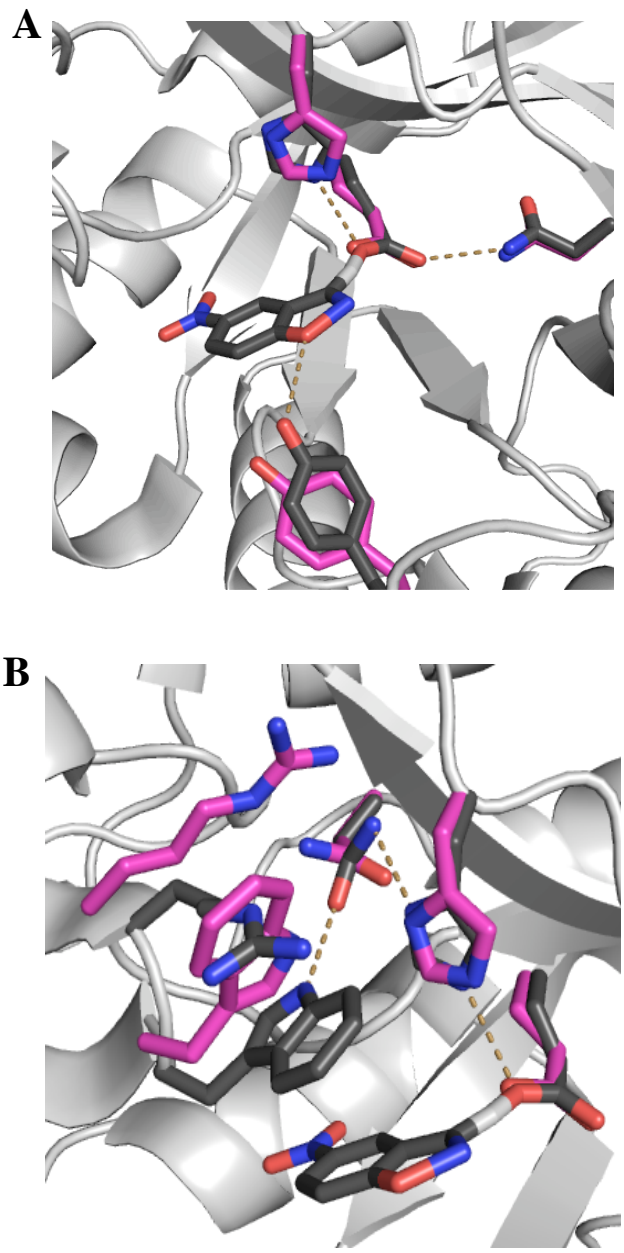

Figure 3-9. Overlay of HG-1 crystal structure active site with design model. The structure predicted by ORBIT (grey) is very similar to the actual conformations seen in the crystal structure (magenta). Predicted hydrogen bonds are represented as yellow dotted lines. **(A)** E237 and the supporting hydrogen bond network and Y90. **(B)** E237 and supporting hydrogen bond network including H209, N239, and W275.

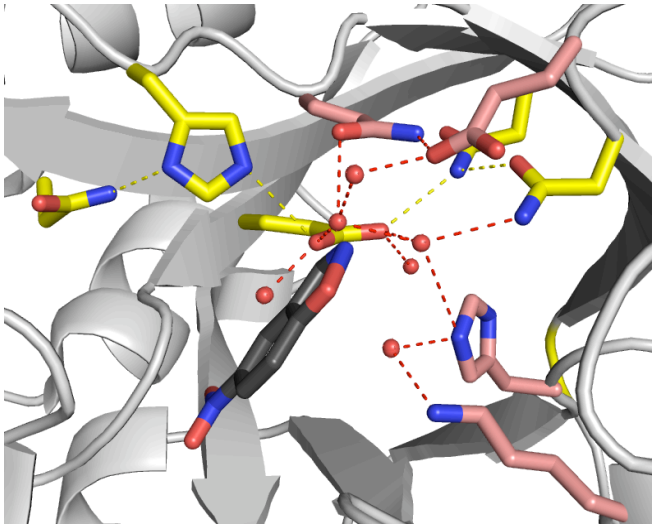

Figure 3-10. Ordered water molecules near E237 in the active site. Residues that participate in the hydrogen bond network to E237 are shown in yellow, additional polar residues in the active site are shown in pink, and water molecules are shown as red spheres. The modeled TS is shown in grey.

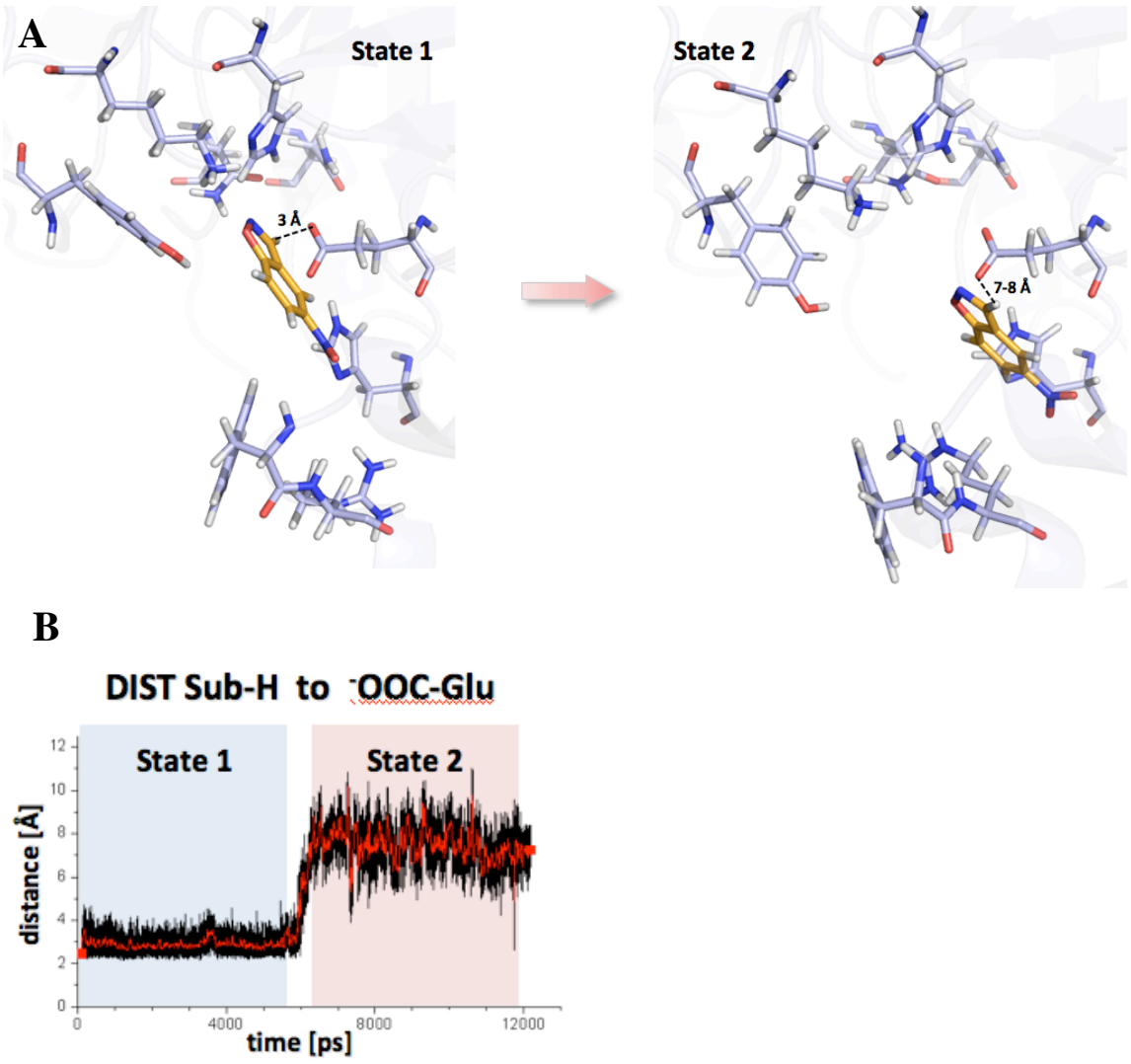

Figure 3-11. MD analysis of HG-1. **(A)** Representative structures of HG-1 in state 1 where the substrate is bound in the active site and in state 2 after the substrate has exited the active site. **(B)** The substrate-base distance over the course of the simulation. States 1 and 2 are indicated.

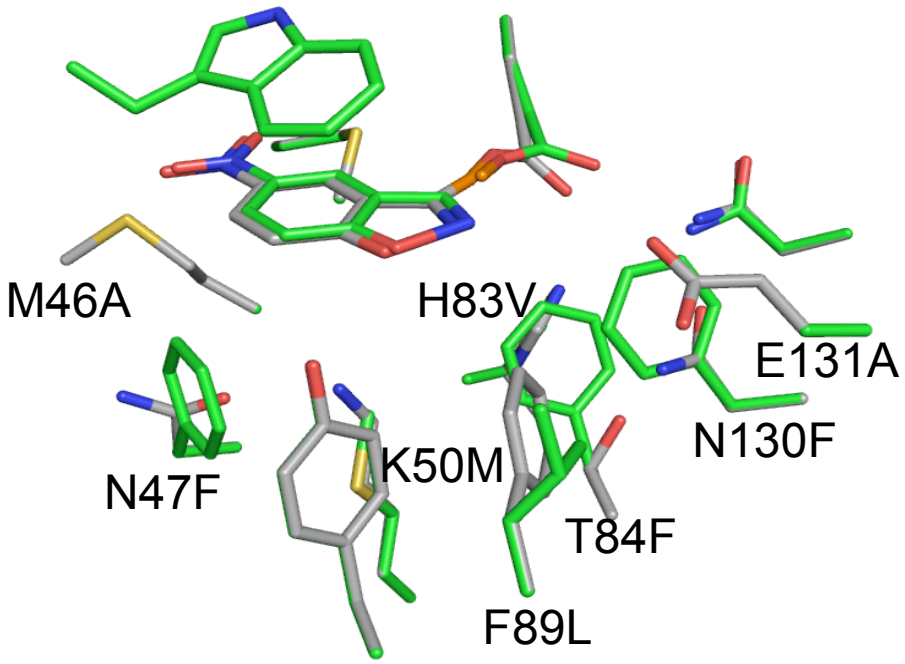

Figure 3-12. Active site of HG-1h. The active site residues of HG-1h (green) are shown overlaid with the HG-1 active site (grey). Mutations with respect to HG-1 are indicated.

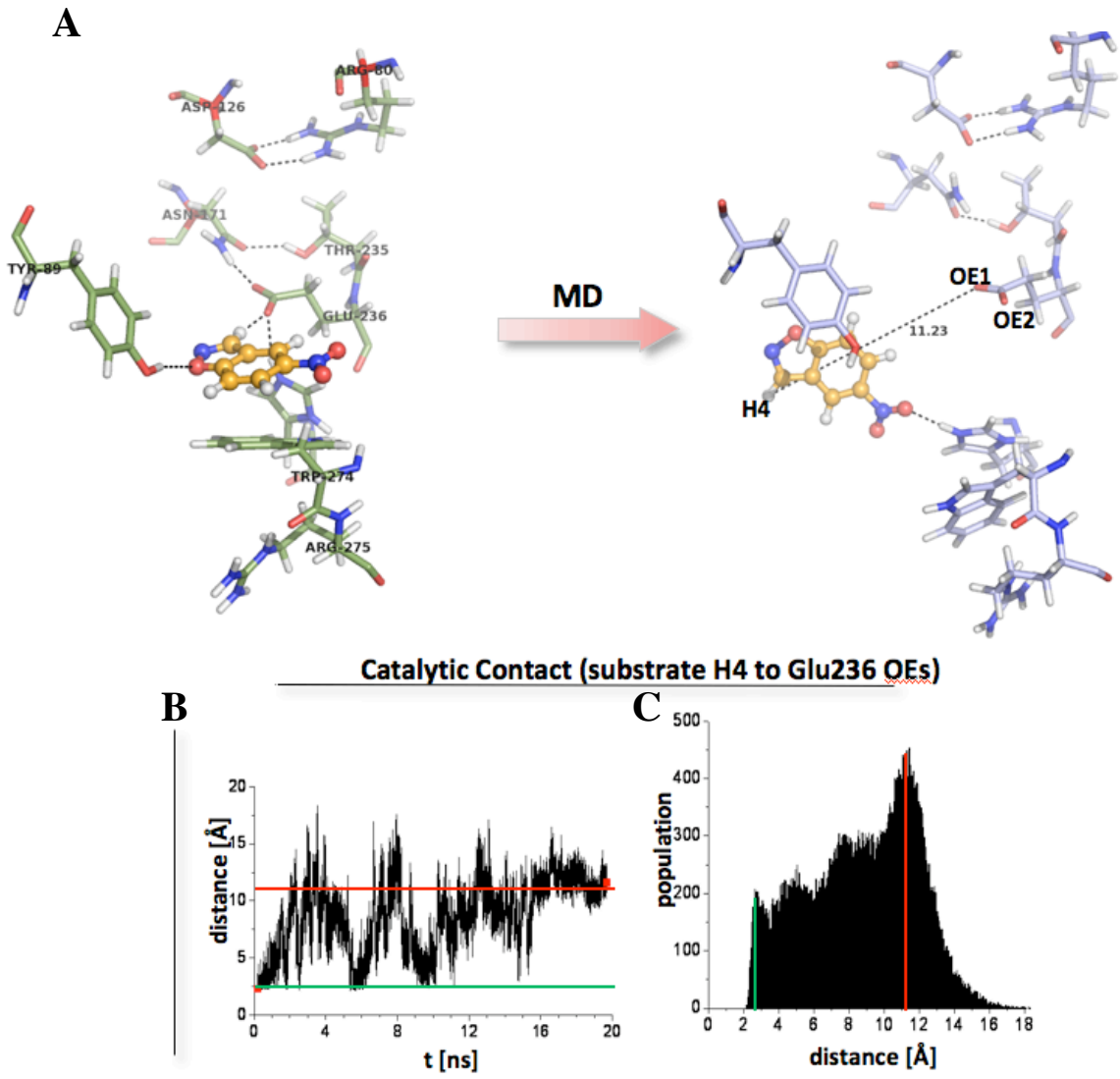

Figure 3-13. MD analysis of HG-1h. **(A)** Representative structures before (green) and after the MD simulation (blue). **(B)** The substrate-base distance over the course of the simulation. **(C)** The distribution of the substrate-base distances for the entire simulation. The substrate-base distance for the initial configuration is indicated by a green line and the substrate-base distance for the final configuration is indicated by a red line.