QM GUIDED COMPUTATIONAL ENZYME DESIGN

By

BEIDI LU

A thesis submitted to the

Graduate School-New Brunswick

Rutgers, The State University of New Jersey

in partial fulfillment of the requirements

for the degree of

Master of Science

Graduate Program in Chemistry and Chemical Biology

written under the direction of

Sagar Khare

and approved by

_____

_____

_____

New Brunswick, New Jersey

October. 2014

ABSTRACT OF THE THESIS

QM guided computational enzyme design

by BEIDI LU

Thesis Director:

Sagar Khare

PT3 is a redesigned adenosine deaminase that could catalyze the hydrolysis of organophosphate. In this present work, we evaluate the impact of previous designed mutations by investigating the mechanism of PT3. We started from the high-resolution crystal structure and truncated the active site residues as the QM model. Then the potential surface of the reaction was discovered by locating the transition states and intermediates along the reaction path with QM method B3LYP. The results showed a similar energy profile compared to natural phosphotriesterase and the nucleophilic attack turned out to be the rate-limiting step. The impact of a single mutation V218F that leaded to 20-fold increase in the catalytic rate k_{cat} was rationalized by including this residue in QM model and a 5.0 kcal/mol difference of the reaction barrier was discovered. Then with the rationalized model, we performed a low-level QM calculation with key bond lengths fixed at a value from high-level QM results. The barrier difference of V218F changed to 3.6kcal/mol, which was still consistent with experimental results while the computation time was cut half. With this fast computational setting, we are able to analyze more mutations of their impact on the reaction barrier quantum mechanically.

Table of contents

List of illustrations

1. Aims and significance

Computational enzyme design usually begins with a quantum mechanics (QM) calculated transition state model of the rate-determining step of the reaction under consideration[1]. The model, called theozyme[4], is docked into the active site of every scaffold enzyme from a structural database, and the design algorithm then iteratively searches the conformational space of the ligand and protein side chains to minimize the energy of variant active site sequences. Current design software likes Rosetta[3] use energy functions that do not accurately model the electronic interactions in active site region; however, these interactions are known to be critical for enzyme activity. As a result, success rates and the catalytic efficacy of designed enzymes are low and laborious laboratory directed evolution is needed to increase the efficiency of designed enzymes. For example, in a previously computationally designed organophosphate hydrolase called PT3, 2500-fold increase in catalytic efficiency was induced by accruing 7 mutations[2]. In particular, a single mutation (V218F) that leads to 20-fold increase in the catalytic rate k_{cat} was identified in directed evolution screens; this substitution is not particularly favorable according to the (molecular-mechanics type) Rosetta energy function. We hypothesize that a QM approach that models the effect of mutations on the electronic environment of the active site is needed to rationalize such existing mutational data and to identify further activity-enhancing substitutions in active site. The central goal of this proposal is to develop a QM-based approach for modeling the impact of amino acid substitutions in designed active sites.

The proposed approach will allow the in silico evolution of designed enzyme activities thereby decreasing the need for experimental directed evolution, and will provide a fundamental understanding of the origins of catalysis in computationally designed enzymes.

To achieve this goal I propose the following three specific aims:

Aim 1: To rationalize the impact of activity-enhancing mutations in PT3 using QM and QM/MM simulations

I will focus on rationalizing the impact of the V218F and other activity-enhancing mutations to benchmark our QM approach. First I will create a model of the active site with first shell residues including the mutated residue (position 218), and use QM to calculate each transition state (TS) and intermediate along the reaction path for 218V and 218F variants. The energy barrier difference in the rate-limiting step will be directly estimated from these calculations and compared with experimental trends. The impact of protein environment and dynamics on the catalytic efficiency and reaction barriers will be probed by performing QM/MM simulations of wild type and variant enzymes. Rationalization of the existing mutational data will set the stage for the use of this model to further improve the PT3 activity to degrade organophosphorus nerve agents and pesticides[5].

Aim 2: Develop a rapid computational method to calculate the impact of mutations on the reaction barrier of PT3

To make it practical to interrogate the impact of a large number of mutations on the catalytic efficiency using QM calculation, we will attempt to shorten the time-scale of discovering the reaction energy path for each mutation using a hierarchical approach[6]. Starting from the optimized structure mentioned in Aim 1, we will first replace the residue with desired mutation and re-optimize the structure at lower-level QM theory such as HF/6-31g(d). Then we can use a high-level theory like B3LYP/6-31g(d) to get the single point energy from the newly optimized structure. By carefully tuning the basis set and the level of method we will decrease the running time of the simulation. We will benchmark our approach with the reaction path and barrier heights calculated in Aim 1.

Aim 3: Use the fast screening method to identify mutations that increase cyanuric acid hydrolysis efficiency of guanine deaminase.

With this fast screening method built upon the model of PT3, it is necessary to test its feasibility in a blind test on an unrelated system. Our lab is working on designing new enzymes for cyanuric acid hydrolysis and our collaborator has identified weak activity in an enzyme, guanine deaminase that is structurally related to PT3. To suggest mutations that will help increase activity, we plan to first calculate the reaction path for cyanuric acid hydrolysis by creating a model of the first-shell residues in the

enzyme using high-level QM theory. Then we will switch to low-level methods to screen the mutations around the active site and test them in experiments. A more efficient cyanuric acid hydrolase developed using our studies will help develop more efficient approaches to degrade s-triazine pollutant compounds.

2. Preliminary Results and Progress

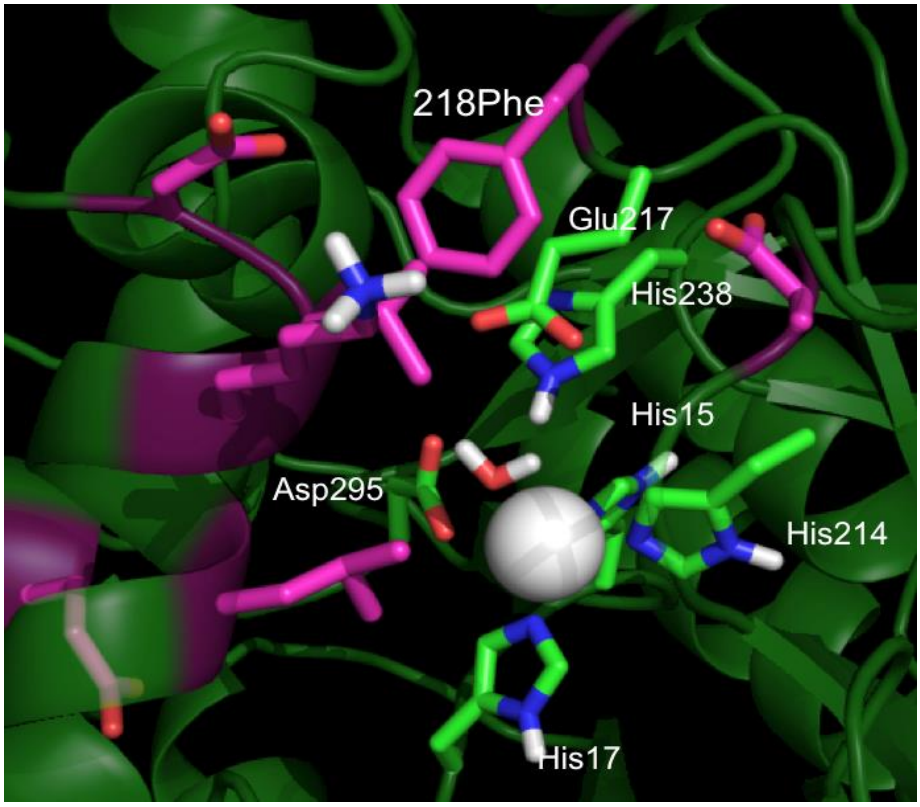


Figure 1. Active site of PT3

Methods:

QM calculations: Unless specifically mentioned, all calculations are performed using the DFT functional B3LYP. Geometry optimization is carried out with a 6-31G(d) basis set for C, H, O, P, N elements and a LANL2DZ basis set for Zn. On the basis of these geometries, single-point calculations were obtained at the same level. To get the charges for the MD parameterization, electrostatic potential fit for the optimized structure were calculated at the same level using a CPCM method with dielectric constant chosen to be 4, which is the standard value to model the protein surroundings. All calculations were performed using the Gaussian 09 program package.

MD simulations: We use ff12SB force field in Amber, which is a standard FF for protein simulation. As for the active site involving Zn, we make custom parameters files. The bond angle values and charges are taken from an apo-state active site QM calculation. The bond force constants involving Zn are taken from ZAFF (a Zinc Amber FF[9]). The angle force constants around Zn are arbitrary set to 999kcal/mol to maintain the penta-coordinate geometry. The starting structure is from apo crystal structure 3T1G and the relative position of substrate is obtained from Rosetta Docking results. The whole system is first relaxed, then heated and annealed to 300K; then after 500ps of equilibrium, it is ready for production run.

Preparation of active site model:

The starting model comes from the crystal structure (PDB 3T1G). The model contains the Zn ion and its first shell residues including His15, His17, His214, Asp295, Glu217, His238, residue218 and the hydroxide (Figure 1). The ligands are truncated such that in principle only the side chains are kept. The histidines were thus modeled as 2-methyl-imidazoles and the aspartate by acetate. Considering that residue 217 and 218 may have some conformation change compared to apo state in crystal, all the side chain atoms of these residues are retained in the model (Figure 2). The overall charge of system is set to zero. To mimic the constraints placed by the surrounding environment of the enzyme, the outer layer carbon atoms of each residue are fixed. The substrate is DECP in the investigation. Then the reaction path is achieved by optimizing the truncated active site.

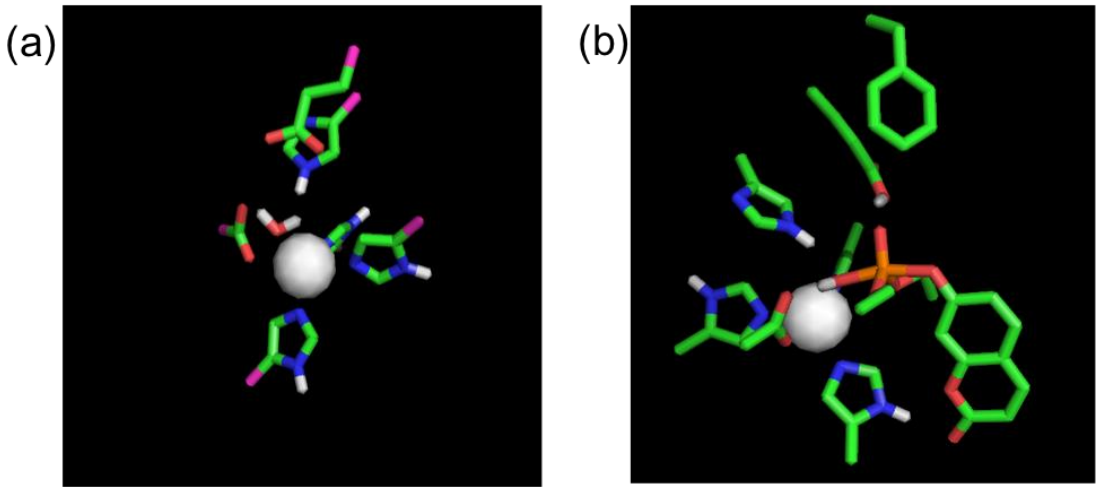
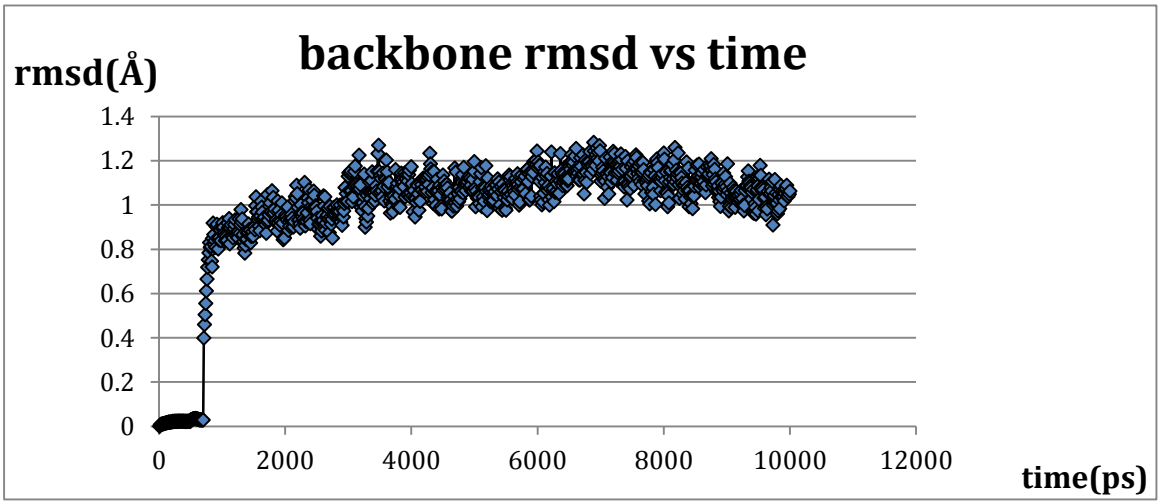


Figure 2. Active site QM model of (a) apo-state, (b) transition-state

2.1.1. Obtaining a structure of the Michaelis Complex using MD simulation:

The starting structure for MC (michaelis complex) is from a MD simulation. To get the parameters for the MD run, the apo-state was first optimized. As we know, the His238 in the WT deaminase enzyme originally acts as a base. And the activity of the WT goes to the ground level by knocking out this HIS. Therefore, we maintained its tautomeric geometry as it is in the crystal structure. The charge of apo-state was then calculated from an RESP (Restrained ElectroStatic Potential) fit of this structure. The side chain atoms directly use the RESP fit charge while the extra charge was smeared out to the backbone to make the overall charge 0. The MD simulation is stable for more that 10ns (Figure 3) with the coumarin ring and hydroxide being on two sides of the P atom, which is a chemically reasonable MC structure for nucleophilic attack.



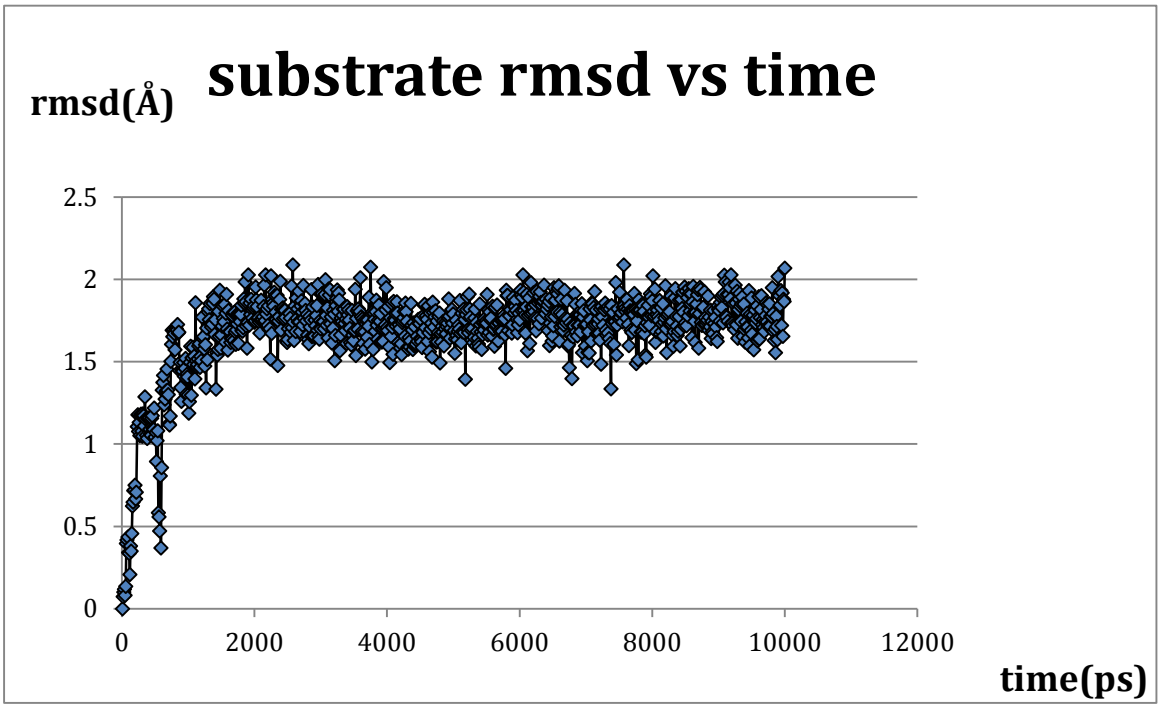


Figure 3. RMSD of (a) backbone and (b) substrate vs time in10ns MD simulations

2.1.2: Identifying a mechanistic pathway for hydrolysis using DFT simulations of an active-site model:

We hypothesized that the reaction follows a S_N2 mechanism in which water molecular catalyzed by Zn attacks the organophosphate and the leaving-group then departures. Previous evidence of enzyme catalyzed organophosphate hydrolysis[10] shows that these reactions follow a classic S_N2 mechanism. To identify the geometry of the transition state and energetics, we performed QM simulations for the active site. We defined a reaction co-ordinate that was the difference between the length of the bonds being broken and formed. i.e. the P-OAr and P-OH bond, respectively. We restrained the system to sample various discrete values of the breaking and forming bonds from 1.9 to 2.5 Angstroms. We find that the energies of optimized structures show a

parabolic behavior with the saddle point corresponding to the transition states. In agreement with previous calculations on the related enzyme PT3, we observed two transition states: TS1 (for nucleophilic attack) and TS2 (for leaving group departure). Frequency analysis showed one large imaginary vibration mode corresponding to the bonds breakage or formation at each TS. We performed identical simulations for a model in which position 218 corresponded to phenyl-methylene group (218F) instead of an isopropyl group (218V). The alpha carbon was maintained at the same position for both residues. The TS1 and TS2 were identified using the same sweeping method. The final optimized geometry of apo-state, MC, TS1, intermediate, TS2 and product are shown in Figure 4. And the energy profiles of these structures are given in Figure 5.

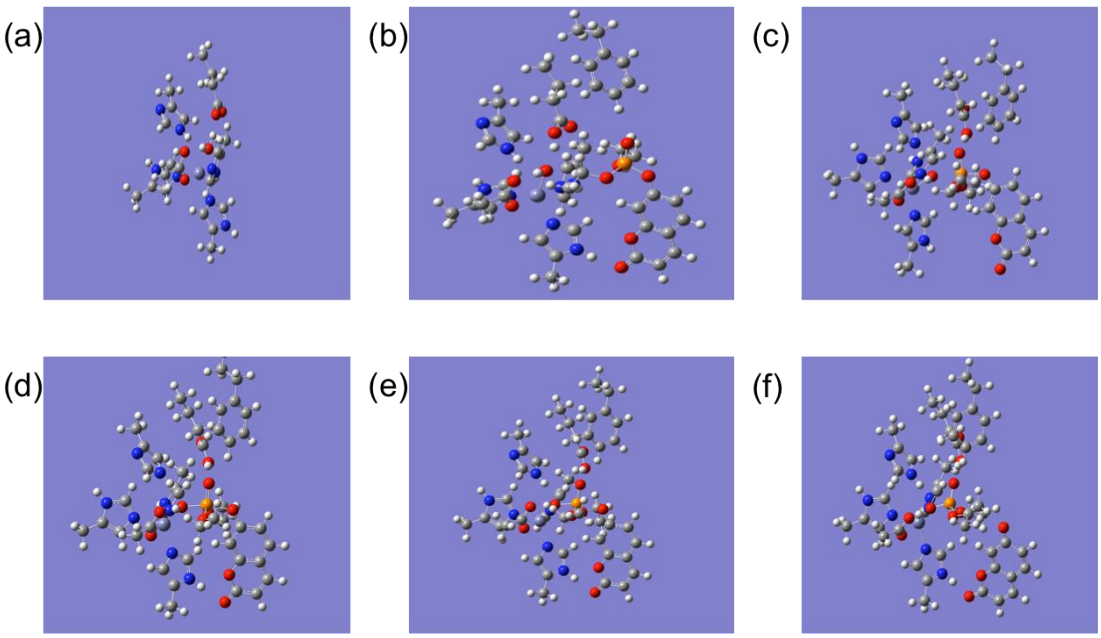


Figure 4. Optimized geometries of apo state(a), michaelis complex(b), TS1(c),

intermediate(d), TS2(e) and product(f) for 218F mutant

	MC	TS1	Int	TS2	Pro
P-OH(218F)	4.997	2.170	1.800	1.714	1.575
P-OAr(218F)	1.646	1.749	1.836	2.200	4.530
P-OH(218V)	5.039	2.170	1.806	1.711	1.570
P-Oar(218V)	1.648	1.738	1.818	2.200	4.557

Table 1. Key bond lengths for the 218F and 218V variants along the reaction path

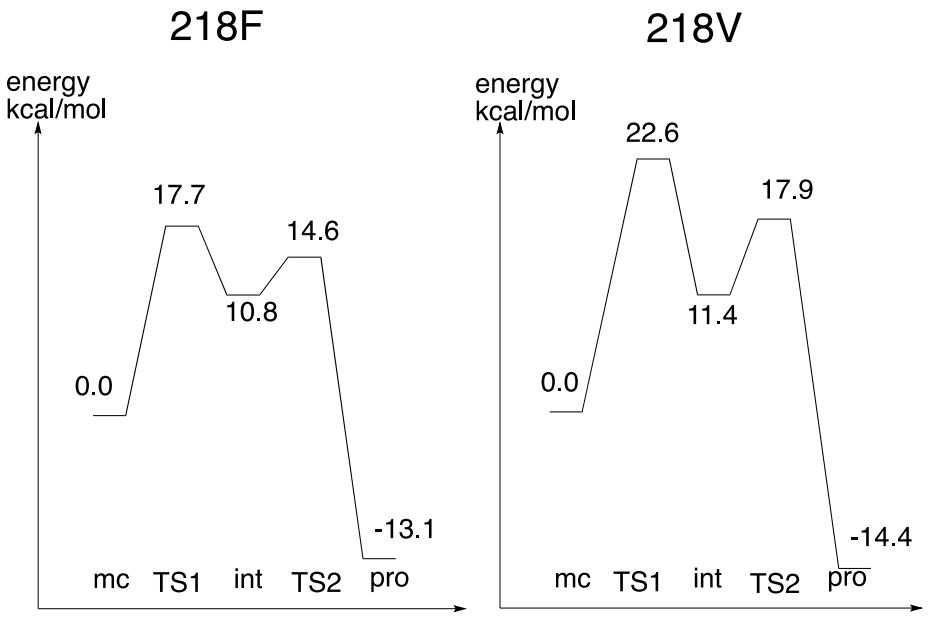


Figure 5. Energy profiles of 218F and 218V variant along the reaction path

We find that the nucleophilic attack step is the rate-limiting step, which is in agreement with previous DFT calculations. Coumarin is a good leaving group (pKa =

7.7), between the two variant, 218F is about 5kcal/mol lower than 218V, which is in good agreement with the catalytic rate experimental data. From the optimized structure, we can observe that the V218F has increased the hydrophobic bulk around the catalytic 217Glu side chain, which probably modulates its pKa and reactivity. This is also consistent with previous computation of pKa at Glu217, which increases 1.6ΔpH unit than PT3.0. The bond lengths along the reaction path (Table 1) closely agree with the previous calculations of paraoxon hydrolysis in a similar Zinc-metallo enzyme using QM[10].

Thus, our results show that this mechanism is consistent with experimental data showing that 218F is more catalytically efficient that 218V (or other substitutions at this position).

2.1.3. Discover why PT3 shows distinctive reactivity for two structurally similar substrates

In the original paper[2], both DECP and paraoxon substrates have been tested. In spite of their highly resembled structures and similar leaving group stability, their activities are totally different. Only DECP is reactive and paraoxon proved to be an inhibiter of this reaction. So I replaced DECP ligand with paraoxon and did the geometry optimization using the same setting again. The energy path turned out to be much different with TS2 being the rate-limiting step and barrier height lower than DECP.(Table 2)

E(kcal/mol)	MC	TS1	INT	TS2	PRO

DECP	0.0	19.5	14.0	14.4	-13.7
Paraoxon	0.0	14.6	12.7	18.4	-13.9

Table 2. Energy profiles for PT3.1 with DECP and paraoxon as substrates separately

The most probable guess will be that the MC between these two ligands has changed a lot. As we can see in the MC of DECP, there is a clear Hbond between HIS14 and the carbonyl group. The O-H bond distance in the previous bound state MD simulations also suggest that there is either a direct Hbond or a water facilitated Hbond.(Figure 6) While coumarin being replaced by nitro-phenol in MC of paraoxon, that interaction shouldn't occur since nitro group is not a good electron donor. Also in the original paper[2], the relatively small size and multiple binding modes are identified as the main reason for low activity. So it is critical to find the binding mode for paraoxon through a MD simulation.

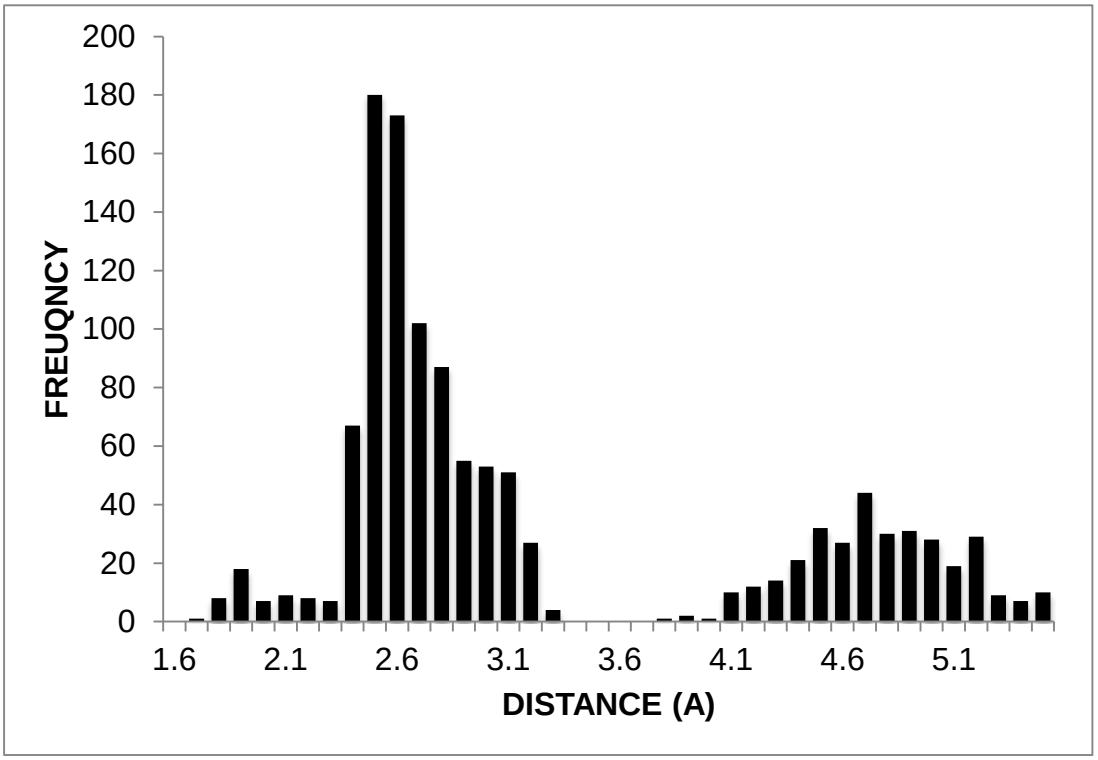


Figure 6. Histogram of O-H distance is bound state MD simulation of DECP substrate

2.1.4. Using a reduced basis set to recapitulate reaction profiles for mutations at position 218

The DFT calculations above yield a reaction pathway that agrees with experimental data, however these calculations are prohibitively expensive to be amenable to large-scale mutational screening. To decrease the computational cost of the simulations, we explored the use of Hartree-Fock methods by using HF/6-31g(d) instead of DFT to re-optimize these structures starting from geometries optimized by DFT. To facilitate the TS search using Gaussian, we sampled in the neighborhood of the structures (MC, intermediate and the two TS structures) identified by DFT. The obtained energy profiles (Figure 7) show a similar trend even though the value of reaction barrier

between 218F and 218V is different (energy difference is 3.6kcal/mol instead of 5.0kcal/mol). This lack of agreement in absolute values is however associated with a significant gain in computational efficiency: computation time was around 16 hrs (on 48 processers) compared to 30 hrs using DFT.

With this fast screening method in hand, we are now able to try to several mutations in the vicinity of the ligand shell. Then we can test their efficiency using direct evolution by a fluorescence absorbance assay[2].

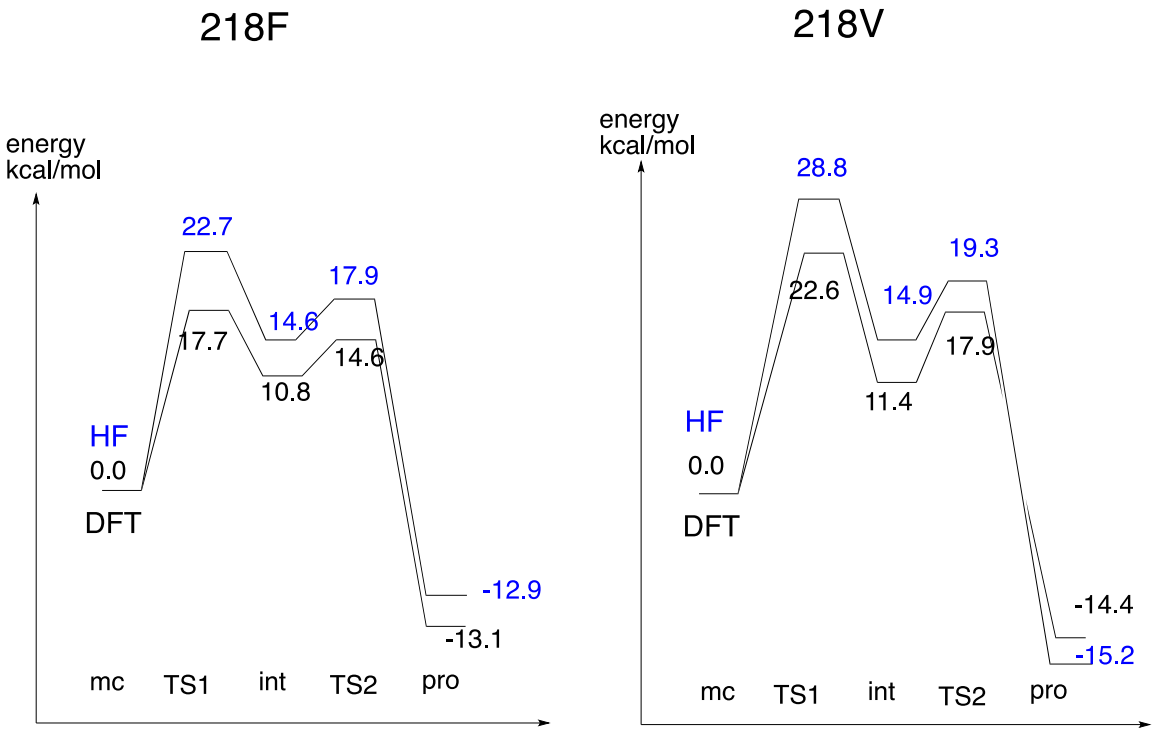


Figure 7. Energy profiles for PT3 reaction for 218F and 218V variants using different QM level methods: DFT: B3LYP/6-31g(d), HF: HF/6-31g(d)

3. Proposed Research

From the preliminary data, it is a good sign that the active site model corresponds to the experimental data pretty well. Then next move would be to take the dynamic effect into consideration, so we could capture details like the loop movement, which single point calculations cannot provide. As for the fast screening approach, we need to test more QM methods and basis sets to find the most efficient setting between accuracy and time.

Aim 1.

1) Rationalize our mechanistic hypothesis by performing QM/MM simulations on all PT3 variants and PT3.1 F218V mutant

There are extensive studies showing the importance of including entropic contributions from dynamic simulations and we could observe a significant energy barrier decrease compared to a single point calculation[7,8]. So a more theoretical rigorous way to investigating the reaction path and estimating energy barriers would be running a QM/MM simulation. The QM region will be the same as we treat before. Considering its relatively large size (>120 atoms), *ab initio* or DFT approaches would be not practical, so following previous work (ref) we propose using a dual-level(DL) approach, i.e. low-level(LL) semi-empirical AM1-QM/MM potential for the free energy simulations and a high-level(HL) B3LYP/6-31d(g) for the QM subsystem in

the gas phase. And the total energy will be:

$$E_{tot}(DL) = E_{QM(HL)} + \Delta E_{QM(LL)/MM} + E_{MM}$$

It is safe to assume that HL free energy can be directly calculated using structures at minimal-energy path (MEP) from LL simulation[8]. And the MEP will be determined by optimizing the structure of solvated enzyme-substrate complex as a function of reaction coordinate z, which is defined as the difference between the distances of breaking and forming bonds. So the free energy is obtained as:

$$W^{DL}(z) = W^{LL}(z) + [E_{QM(HL)}(z^{MEP}) - E_{QM(LL)}(z^{MEP})] + \Delta W_0$$

We will then perform umbrella sampling through a total of 20 separate simulations, spanning the entire reaction coordinate, each of which will be equilibrated for at least 20ps followed by an additional 100ps sampling.

By this way, we could get a more accurate estimate of energy barrier of all PT3 variants; it will be a strong backup for our current theory and model. Also, we could do a clustering of backbone alpha carbons in the QM region from the simulation. It provides another way to account for the backbone change from MC to TS. We can apply this new restraint to improve the active site model.

2) Discover why PT3 shows distinctive reactivity for two structurally similar substrates

It is shown that if paraoxon is able to bind PT3 in a similar fashion as DECP does, the reaction should proceed without any difficulty. So first I plan to identify the MC binding mode by MD simulation. The custom parameters and input files will be

prepared the same way as described above. We could cluster the trajectories that have a similar structure and use MMPBSA to calculate the binding affinity. Then we run a QM/MM simulation with the same method and the same reaction coordinate in 1) and get the energy barrier. We will then know how to correct the active site model. And this will give us a better insight of PT3 mechanism and we can design PT3 to catalyze substrates that don't have activity before.

Aim 2.

1) Tune the QM methods and basis sets to increase efficiency

The energy barrier difference between 218V and 218F decrease from 5.0kcal/mol to 3.6kcal/mol after we switch from B3LYP/6-31g(d) to HF/6-31g(d). Although it is still a big gap when converting to rate constant ratio, it makes this metric less distinguishable. And 16 hrs on 48 processors for one job is not satisfactory enough. To improve the results, we could try doing a high-level energy calculation such as B3LYP/6-31g(d,p) from the low-level optimized structure. And to reduce time, we could try semi-empirical methods such as AM1, PM3 and PM6 to optimize the structure. By trying all kinds of reasonable combinations, we could find a more efficient method that satisfies the experiment data at the same time. And if we could reduce the time scale to some extent, we could even include second or third ligand-shell in the active site model. That will vastly improve the utility of this method.

2) Screen mutations that lower the energy barrier of PT3 reaction

Due to the size limitation of current screening method, all mutation residues should be right next to the first-shell. And as we discussed above, the key residue for this PT3 is the GLu217, which will deprotonate the catalytic water in the MC, making it suitable for nucleophilic attack. And the reason why V218F mutation increases the reaction rate so much is that residue218 is sitting on top of Glu217, and when it mutates to a bigger and more hydrophobic Phe, it basically cut off all other interactions with Glu217, making Glu217 pka increase and more prone to deprotonate the water. So right now the key of design lies in increasing the hydrophobicity around the residue217. Other than 218, the other residue that is also close to the Glu(~4.5A) is Thr266, and there is enough space between these two residues. So we reason that by increasing the hydrophobicity of Thr266, we could further eliminate the contact between Glu217 and solvent water, thus increasing the pka and PT3 activity. So we will screen all the possible big hydrophobic residues at position 218 and 266and test those with lower energy barrier in experiments by direct evolution.

Aim3.

Test the feasibility of fast screening method in a new system

With the new screening method at hand, it is critical to perform a blind test to ensure its feasibility. Our lab is working on designing new enzymes for cyanuric acid hydrolysis and our collaborator has identified weak activity in an enzyme, guanine deaminase, which has a very similar active site structure (Figure 8). The mechanism

of how guanine deaminase catalyzes the hydrolysis of CAH is unknown. So we will perform a similar reaction path study like the way we did for PT3. First, we will find the MC structure with MD simulations and then optimize the truncated active site using QM. After we locate the TS of the rate-liming step, we could use low-level method to screen the surrounding residues with fixed breaking and forming bond length. Finally we will use direct evolution to test those designs that have lower energy barrier. A more efficient cyanuric acid hydrolase developed using our studies will help develop more efficient approaches to degrade s-triazine pollutant compounds.

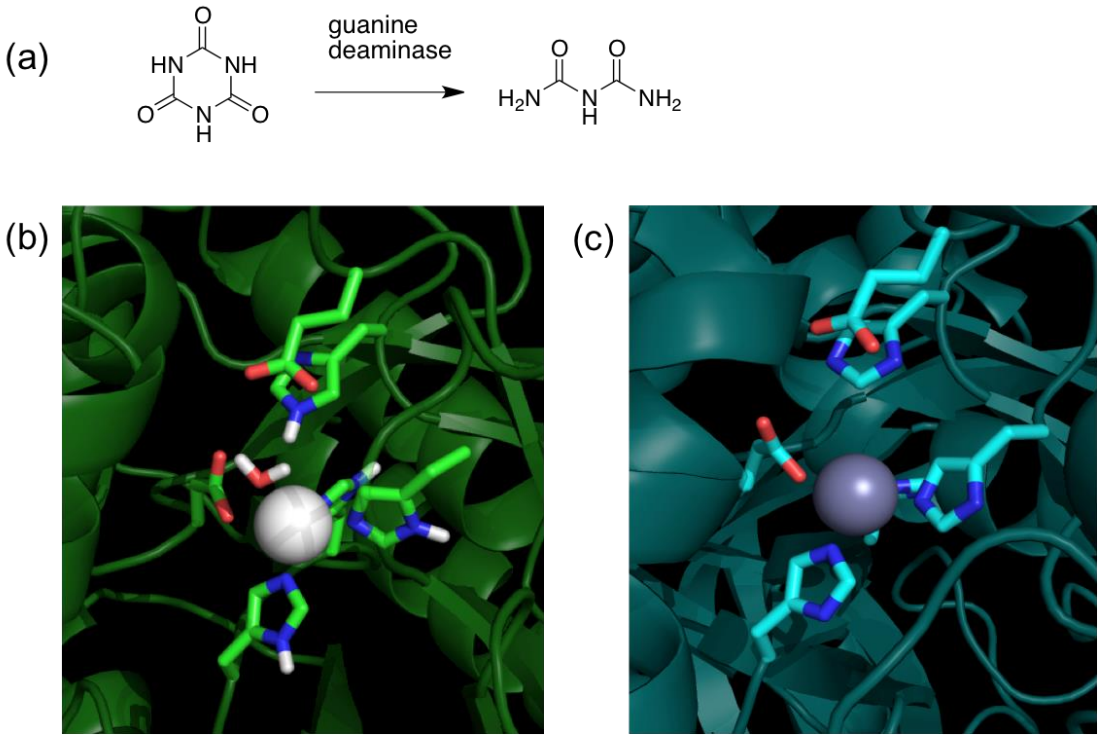


Figure 8. (a) Reaction of cyanuric acid hydrolysis. (b) Active site structure of PT3. (c) Active site structure of guanine deaminase

Acknowledgment:

I'd like to acknowledge all the work done by previous publications. I also want to thank my advisor Sagar Khare and my lab members for all the help in this project.

Reference:

1. Donald, Hilvert. "Design of Protein Catalysts." Annu. Rev. Biochem. (2013). 82:447–70
2. Khare SD, Kipnis Y, Greisen PJ, Takeuchi R, Ashani Y, Goldsmith M, Song Y, Gallaher JL, Silman I, Leader H, Sussman JL, Stoddard BL, Tawfik DS, Baker D. Nature Chem. Biol. (2012): 8:294–300.
3. Richter, Florian, et al. "De novo enzyme design using Rosetta3." PLoS One 6.5 (2011): e19230.
4. Tantillo, Dean J., Chen Jiangang, and Kendall N. Houk. "Theozymes and compuzymes: theoretical models for biological catalysis." Current opinion in chemical biology 2.6 (1998): 743-750.
5. Gupta, R.D. et al. Directed evolution of hydrolases for prevention of G-type nerve agent intoxication. Nat. Chem. Biol. (2011): 7, 120–125.
6. Hediger, Martin R., et al. "A computational methodology to screen activities of enzyme variants." PloS one 7.12 (2012): e49849.
7. Gao, Jiali, et al. "Mechanisms and free energies of enzymatic reactions."Chemical reviews 106.8 (2006): 3188-3209.
8. Wong, Kin-Yiu, and Jiali Gao. "The reaction mechanism of paraoxon hydrolysis by phosphotriesterase from combined QM/MM simulations." Biochemistry46.46 (2007): 13352-13369.
9. Peters, Martin. J. Chem. Theory Comput. (2010) 6.9: 2935-2947.
10. Chen, Shi-Lu, Wei-Hai Fang, and Fahmi Himo. "Theoretical study of the phosphotriesterase reaction mechanism." *The Journal of Physical Chemistry B* 111.6 (2007): 1253-1255.