A Computational Methodology to Screen Activities of Enzyme Variants

Martin R. Hediger[1], Luca De Vico[1], Allan Svendsen[2], Werner Besenmatter[2], Jan H. Jensen[1]*

1 Department of Chemistry, University of Copenhagen, Copenhagen, Denmark, **2** Novozymes A/S, Bagsværd, Denmark

Abstract

We present a fast computational method to efficiently screen enzyme activity. In the presented method, the effect of mutations on the barrier height of an enzyme-catalysed reaction can be computed within 24 hours on roughly 10 processors. The methodology is based on the PM6 and MOZYME methods as implemented in MOPAC2009, and is tested on the first step of the amide hydrolysis reaction catalyzed by the *Candida Antarctica* lipase B (CalB) enzyme. The barrier heights are estimated using adiabatic mapping and shown to give barrier heights to within 3 kcal/mol of B3LYP/6-31G(d)//RHF/3-21G results for a small model system. Relatively strict convergence criteria (0.5 kcal/(molÅ)), long NDDO cutoff distances within the MOZYME method (15 Å) and single point evaluations using conventional PM6 are needed for reliable results. The generation of mutant structures and subsequent setup of the semiempirical calculations are automated so that the effect on barrier heights can be estimated for hundreds of mutants in a matter of weeks using high performance computing.

Introduction

Current computational studies of enzyme activity as measured by the activation free energy generally restrict their focus to the wild type enzyme and, perhaps, one or two mutants described with a comparatively high [1–3] or moderately high [4–6] level of theory. The agreements with experimental results are often impressive and these studies can provide valuable insight into the catalytic mechanism [7]. However, the computational demands of these methods makes it difficult to apply them to the actual design of new enzymatic catalysts where the activity of hundreds of mutants has to be evaluated. This paper describes a computational method that makes this practically possible.

In order to make the method computationally feasible, relatively approximate treatments of the wave function, structural model, dynamics and reaction path are used. Given this and the automated setup of calculations, some inaccurate results will be unavoidable. However, the intend of the method is similar to experimental high through-put screens of enzyme activity where, for example, negative results may result from issues unrelated to the intrinsic activity of the enzyme such as imperfections in the activity assay, low expression yield, protein aggregation, etc. Just like its experimental counterpart our technique is intended to identify *potentially* interesting mutants for further study.

In this paper we develop and test the technique on the *Candida Antarctica* lipase B (CalB) enzyme. CalB catalyses the hydrolysis of lipophilic esters and shows only very low amidase activity. While we use the method to test the effect of a few mutations on the first step in the hydrolysis of a simple amide by CalB (Fig. 1), the main point of this study is the developement of a general, efficient and robust computational method that can be used on systems similar to this.

Methods

In this paper we focus on estimating k_{cat} rather than k_{cat}/K_M because most industrial uses of enzyme catalysts work at high substrate concentration where k_{cat} is most critical for product formation. Therefore, like in most computational studies of enzyme catalysis, substrate binding-affinity is not considered. The inclusion of protein dynamics is not considered here. The most common way of estimating the effect of protein dynamics on barrier height in QM/MM studies is to compute the barrier height starting from different snapshots from a molecular dynamics simulation. This way of treating protein dynamics can also be done with our method, but was not done in this study mainly for reasons of efficiency but also because it has not been conclusively demonstrated that this in fact increases the accuracy of the predicted barrier. For example, Friesner and co-workers [8] have predicted several barrier heights within a few kcal/mol of experiment without inclusion of such dynamic effects. Furthermore, when estimating relatively small changes in barrier heights due to mutations it is not clear that dynamic effects can be predicted precisely enough from averaging over a few snapshots. However, we hope to study this issue in future studies.

Another approximation is the use of gas phase energy evaluations to estimate the barrier. Exploratory calculations revealed that it is not possible to do COSMO [9] calculations with PM6 [10] for systems as large as this using MOPAC2009 [11]. While it is possible to perform COSMO calculations with MOZYME [12], our work shows that it is not clear that
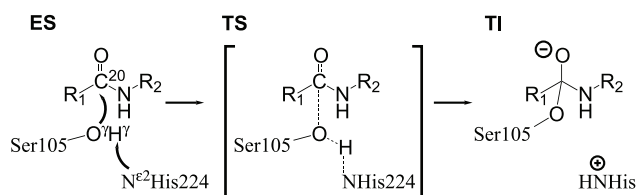
Figure 1. Formation of tetrahedral intermediate mechanism. Concerted nucleophilic attack of O^γ of S105 and abstraction of proton H^γ by $N^{\epsilon 2}$ of H224 and development of formal negative charge on substrate oxygen. The enzyme substrate complex **ES** (left) is transformed into tetrahedral intermediate, **TI**. R_1: $-CH_2CH_3$, R_2: $-CH_2C_6H_5$.
doi:10.1371/journal.pone.0049849.g001

MOZYME energies are sufficiently accurate to estimate relatively small differences in barrier height.

As will be discussed in more detail in the results section, a computational technique aimed at the study of activity in enzymes requires the molecular models to include a significant part of the enzyme. These models are in general too large to be treated with *abinitio* methods. The full quantum mechanical treatment of a large molecular model is however possible when using semiempirical (SE) methods in combination with linear scaling techniques. A range of semiempirical methods is therefore evaluated and discussed. In particular, the AM1 [13], PM3 [14] and RM1 [15] methods as implemented in the GAMESS [16] program and the PM6 method as implemented in the MOPAC2009 program are evaluated. In the evaluation of the semiempirical methods, single point energy calculations are carried out at the B3LYP/6-31G(d) level of theory (as implemented in GAMESS). Electronic energies and enthalpy of formation, $\Delta_f H$, are not corrected for zero point energy (ZPE).

Since the semiempirical methods use a predefined (Slater type) basis set (minimal basis for AM1, PM3 and RM1, augmented by d-orbitals on main-group atoms in PM6) and core approximation [17], a quantum chemical geometry optimization is mainly configured by the setting of the gradient convergence criterion (GCC). When using localized molecular orbitals (LMOs) provided through the MOZYME method in MOPAC, it is in addition possible to adjust the distance at which the neglect of diatomic differential overlap (NDDO) approximations [18] are discarded and replaced by point charge interactions. Initially, the MOZYME method generates a Lewis structure of the molecule which is used to calculate the initial density matrix for the self-consistent field (SCF) procedure. The implications of using MOZYME LMOs are further discussed below.

This work is considered only with the estimation of the barrier of the reaction of Fig. 1, whereas binding effects and solvation effects are not considered explicitly. The description of a robust and efficient technique for the estimation of said reaction barrier is the purpose of this publication.

The computer scripts for generating the molecular models are available online, the URL is provided in Text S1.

Results and Discussion

Evaluation of SE methods

To assess which computational method is best suited for use in a screening approach, the first step is the evaluation of the accuracy of the various methods in predicting the geometry of the transition state (**TS**) of the reaction in Fig. 1.

The method evaluation is done in a small model representing the active site (**AS**) of the enzyme, consisting of 54 atoms, (**1**), Fig. 2.
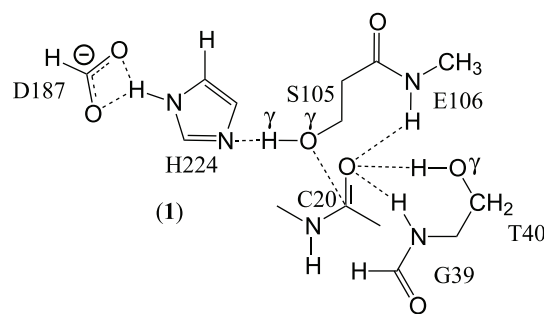

Figure 2. Molecular system used in method evaluation. The system carries overall charge of -1. The oxyanion hole is formed by backbone amide of G39, T40 and E106 and O^γ of T40.
doi:10.1371/journal.pone.0049849.g002

The geometries of the *TS* obtained from the SE methods are compared to the *Hartree-Fock* (HF) geometry, Fig. 3.

The molecular structure of (**1**) is generated by extracting the coordinates of the atoms of the residues G39, T40, S105, E106, D187 and H224 from the crystal structure of CalB (PDB ID 1LBS [19]). In order to reduce computational effort, only fragments of the amino acids are included. From G39, the carbonyl group and the backbone amide is included, from T40 C^α, C^β and O^γ are included. From S105, the backbone nitrogen is discarded, from E106 only the backbone nitrogen is included, the rest of the amino acid is replaced by a methyl group. D187 is represented by formic acid and from H224 only the imidazole moiety is included. All open valences are completed by the addition of hydrogens. The substrate methylacetamide ($CH_3NHCOCH_3$) is introduced by replacing the bound inhibitor molecule from the crystal structure.

The **TS** is located by providing a suitable guess structure as a starting point followed by carrying out Newton-Raphson optimization. In the guess structure, the distance between O^γ of S105 and C20 of the substrate is 1.80 Å and the distance between O^γ and H^γ is 1.1 Å. The **TS** is located with HF and after confirmation of the **TS** nature by vibrational analysis, it is used as a guess structure for the calculations with the SE methods. For every SE method, the nature of the **TS** is verified by carrying out vibrational analysis. In all optimizations of **TS**, no constraints are applied. To verify that the **TS** indeed connects the enzyme-substrate complex (**ES**) and the tetrahedral intermediate (**TI**), intrinsic reaction
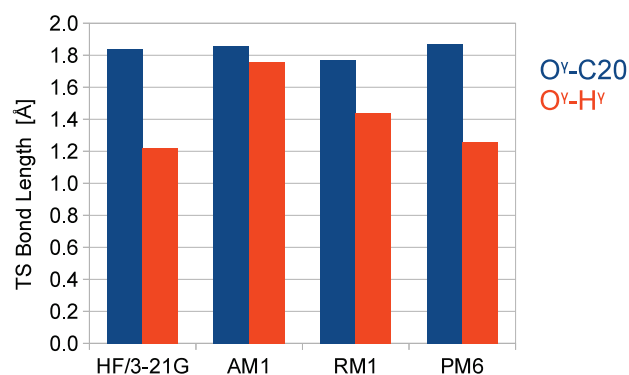
Figure 3. TS geometries of SE methods. Comparison of TS bond length r between S105 O^γ and H^γ (in Å, PM3 values not reported, see text) in (**1**). HF: $r^{O^\gamma/H^\gamma} = 1.22$; PM6: $r^{O^\gamma/H^\gamma} = 1.26$; AM1: $r^{O^\gamma/H^\gamma} = 1.76$, i.e. H^γ completely transfered to imidazole ring; RM1: $r^{O^\gamma/H^\gamma} = 1.44$. RMSD of alignment (in Å): HF/RM1 = 0.554; HF/AM1 = 0.538; PM6/HF = 0.224.
doi:10.1371/journal.pone.0049849.g003

coordinate (IRC) calculations are carried out. The stationary end points, i.e. **ES** and **TI**, of the IRC calculation are optimized without any constraints at the same level of theory as used in the **TS** search and density functional theory (DFT) single point energy calculations are performed on the optimized stationary points. In all geometry optimizations, the gradient convergence criterion is set to 0.5 mHa/Bohr using GAMESS and 0.5 kcal/(molÅ) using MOPAC.

Using the distances $O^\gamma/C20$ and O^γ/H^γ in the HF **TS** and the RMSD between the HF and the SE **TS** structures as a measure of comparison between different methods, it is observed that the geometry obtained from PM6 is in best agreement with the HF reference, Fig. 3.

It is observed that the major difference between the methods is in the position of the H^γ proton. The distance between the nucleophilic O^γ and C20 of the substrate is very similar in all cases.

The IRC calculations show that all methods, except PM3, are able to locate a **TS** which corresponds to a concerted mechanism of nucleophilic attack and proton abstraction. The PM3 method produces a stepwise mechanism where a deprotonated serine is formed, carrying a formal negative charge. In this species, O^γ of the serine is hydrogen bonding to the amide proton of the substrate and significant rearrangement of the molecular structure is observed (RMSD of alignment between HF and PM3: 1.66 Å).

It is observed that the energy difference for the geometries obtained by PM6 is in very close agreement to the HF reference geometry, Table 1.

It is interesting to note that the energy difference based on the **TS** geometry obtained from AM1 is also very close to the HF value, however, the corresponding structure is qualitatively different. Using AM1, the **TS** is characterized by a deprotonated serine, whereas in HF and PM6 the proton is partially bonded between the serine and the histidine. The lower barrier from the RM1 based geometries is explained by a minor increase of the energy of the reactant relative to the **TI**.

The analysis of the **TS** bond lengths and the RMSD values shows that the geometry of the **TS** found with PM6 is in best agreement with the HF reference geometry. It is also noted that the PM6 method has recently been reported to provide DFT grade geometries [20].

Molecular enyzme model size

The definition of a molecular model appropriate to use in the study of enzyme activity is subject to the following conditions. In the context of the proposed screening approach, the molecular model is required to include at least all sites which are potential targets for mutations. The upper boundary for the size of the

molecular model is controlled essentially by the computational effort required for the calculation. For industrial applications, it is usually desirable to obtain results within 24 hours of wall clock time. In addition, it is assumed that the catalytic effect of a mutation located more than 10 Å away from the active site is negligible.

The molecular model and the configuration of the MOPAC program are assessed by constructing three molecular models of different sizes, Fig. 4. All three models (**a**), (**b**) and (**c**) are based on the atomic coordinates of the crystal structure and are generated by selecting a specific set of residues (complete amino acid sequence given in Text S1).

To afford the computational cost, the molecular model is optimized using the MOZYME LMO method and subsequent single point energy calculations are carried out using PM6 without using MOZYME. This is required since it is possible that the MOZYME energy accumulates error during geometry optimization. This observation is further discussed below.

In (**a**), only the catalytic triad, the oxyanion hole and few other residues in the active site are included. In (**b**), all residues within 8 Å of S105 and in (**c**) all residues within 12 Å of S105 are included. In case the backbone chain of the selection of residues is interrupted by only one residue, this residue is included as well. Crystal waters are also included into the molecular model. All N-termini introduced by interrupting the backbone chain are set to carry zero charge, all C-termini are modeled as -CHO groups. The benzylacetamide substrate ($CH_3CH_2CONHCH_2C_6H_5$) is introduced by molecular modeling to be in overlay with the inhibitor molecule of the crystal structure. In doing so, perfect binding is assumed. The substrate is modeled to be covalently bonded to the active site S105 and with the carbonyl carbon in tetrahedral geometry.

The effect of the MOPAC configuration is studied by optimizing the structure and computing the heat of formation, $\Delta_f H$, of the **TI**. In Table 2, results for a set of 9 different MOPAC configurations for all three models are shown, the time requirements are further discussed below.

In (**a**), $\Delta_f H$ is essentially independent of the gradient convergence criterion. This can be explained by the fact that the number of local minima is limited (compared to (**b**) and (**c**)) and that a gradient convergence criterion of 5 kcal/(molÅ) is sufficiently strict to lead to an optimization of all local minima. It is also observed, that the computed $\Delta_f H$ does not significantly change when optimizing the structure using a NDDO cutoff of 12 or 15 Å.

In (**b**), significant differences in $\Delta_f H$ using gradient convergence criteria of 5.0, 1.0 or 0.5 kcal/(molÅ) are observed. It can be assumed that the strict gradient convergence criteria are required to sufficiently optimize the large number of local minima of the model, the implications of which are further discussed below. Interestingly, the optimization using a gradient convergence criterion of 0.5 kcal/(molÅ) and a NDDO cutoff of 12 Å leads to a geometry with lower $\Delta_f H$ (-4323.5 kcal/mol) compared to optimization with the NDDO cutoff set to 15 Å (-4318.2 kcal/(molÅ)). The observed reason for this is that although using identical starting geometries, different NDDO cutoff settings can result in different final hydrogen bonding networks which are eventually lower in energy. This observation is made with the residue S50, which is located on the surface of model (**b**), Fig. 5a. Initially, O^γ of S50 is roughly equally distant of the backbone carbonyl groups of P45 and Q46, Fig. 5b. After optimization, new hydrogen bonds are formed differently when optimizing with a NDDO cutoff of 12 or 15 Å, respectively, Figs. 5c, d.

Table 1. Energy difference between TS and ES of SE structures.

Energy Differences	ΔE [kcal/mol]
B3LYP/6-31G(d)//HF/3-21G	20.95
B3LYP/6-31G(d)//AM1	19.85
B3LYP/6-31G(d)//RM1	14.15
B3LYP/6-31G(d)//PM3[a]	6.62
B3LYP/6-31G(d)//PM6	18.20

Electronic energy difference between **TS** and **ES** in (**1**), not corrected for ZPE.
a: PM3 value only for proton abstraction.
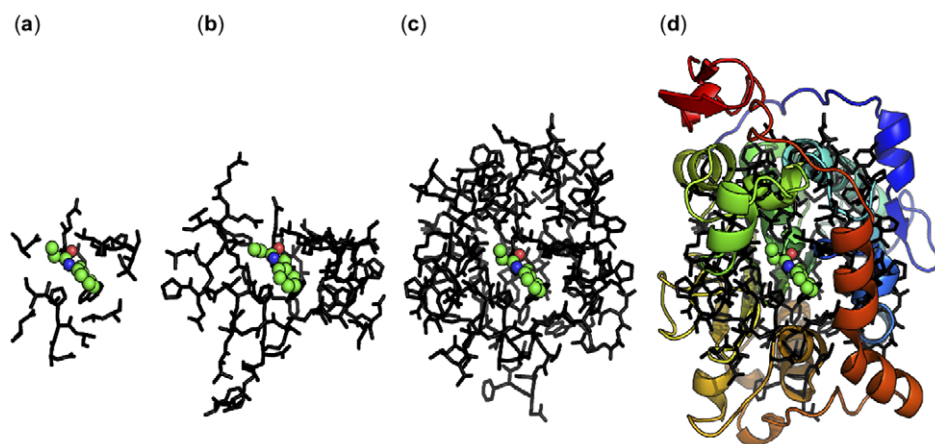doi:10.1371/journal.pone.0049849.t001

Figure 4. Enyzme molcular model size assessment. (a): 17 residues; **(b)**: 55 residues; **(c)**: 118 residues; **(d)**: Full enzyme (316 residues) in cartoon with **(c)** overlayed in sticks. Charge on models **(a)**, **(b)** and **(c)**: −1, −4, −6, respectively. Protonation states of ionizable residues (at hypothetical pH of 7.4) determined with PROPKA v3.1 [26], except E188 which is deprotonated.
doi:10.1371/journal.pone.0049849.g004

The rearrangement of surface residues has to be considered an inheritant artifact of the method, however it is interesting to note that different NDDO cutoffs can lead to different arrangements of the hydrogen bonding network.

It is further observed (Table 2) that the required time to optimize the system (**b**) using strict gradient convergence criterion and NDDO cutoff is within the time frame offered in industrial environments.

Model (**c**) consists of around half of all residues of the full enzyme leading to a large number of local minima on a flat potential energy surface. A strict gradient convergence of 0.5 kcal/(molÅ) combined with a high NDDO cutoff distance of 15 Å is required to completely optimize all parts of the model. In a model of this size, $\Delta_f H$ is considerably reduced both with gradient convergence and NDDO cutoff distance. Model (**c**) possibly provides the most detailed description of the active site, however,

the computational time required to optimize the structure makes it prohibitive to use in a screening approach.

The required computational wall clock time for optimization of models (**a**), (**b**) and (**c**) in dependence of gradient convergence criterion is summarized in Fig. 6.

It is observed that the required wall clock time for complete optimization of the molecular model increases non-linearly with model size. Only when using gradient convergence criterion of 1.0 kcal/(molÅ) is linear scaling of wall clock time with model size observed for NDDO cutoff of 15 Å. Using strict gradient convergence criterion of 0.5 kcal/(molÅ), linear scaling of wall clock time is approached only for NDDO cutoff distance of 9 Å.

From considering the observed time requirements, it is concluded that an intermediately sized model like (**b**) is adequately suited for the proposed screening approach.

Wild type reaction barrier estimation

In establishing an enzyme activity screening technique, it was tested if an approach similar to the one discussed above can be used to study activity in (**b**). Using the GEO_REF keyword [21], the MOPAC program offers an optimization routine where two structures on either side of a reaction barrier are provided to the program. The one higher in energy is used as a reference structure for the one lower in energy. An adjustable penalty potential (based on the geometrical difference between the two structures) is then applied in the optimization of the low energy structure, which will be forced to move towards the transition state on the potential energy surface (PES). After a few cycles of optimization (using the penalty potential), in principle a guess of the transition state is obtained which can be refined using a transition state search routine. However, despite extensive testing of different magnitudes of the penalty potential, it was frequently observed that the optimization is unsuccessful in generating a valid estimate of the transition state for the reaction under consideration. Instead, the structure under the penalty potential remains on one side of the barrier or completely passes the barrier. The exact location of the transition state in large systems by this method is thus not routinely feasible and the approach is not applicable in an industrial context where semi-quantitative estimates of the overall activity are requested within one day of CPU time. This limitation becomes even more apparent when a large library of mutants is to be studied.

Table 2. Comparison of MOPAC configurations and molecular model size.

GCC	PM6//C9	PM6//C12	PM6//C15
(a): 17 Residues			
5.0	−1548.2(0.3)	−1560.1(0.7)	−1561.3(0.5)
1.0	−1548.2(0.4)	−1560.1(0.7)	−1561.3(0.5)
0.5	−1548.2(0.4)	−1560.1(0.6)	−1561.3(0.7)
(b): 55 Residues			
5.0	−4281.4(2.2)	−4288.6(3.7)	−4286.0(3.1)
1.0	−4311.6(7.9)	−4323.5(12.6)	−4317.6(16.1)
0.5	−4311.6(7.3)	−4323.5(12.4)	−4318.2(18.0)
(c): 118 Residues			
5.0	−9315.8(18.9)	−9302.5(29.5)	−9296.7(25.0)
1.0	−9327.7(37.2)	−9310.3(54.4)	−9306.0(43.6)
0.5	−9327.6(37.5)	−9311.8(63.1)	−9316.4(85.8)

PM6 $\Delta_f H$ (in kcal/mol) of **TI** depending on GCC (in kcal/(molÅ)) and NDDO cutoff (in Å), wall clock time for geometry optimization in parenthesis (in h). "PM6//C9": PM6 $\Delta_f H$ computed on structure optimized with MOZYME and NDDO cutoff set to 9 Å. Average wall clock time (in h) for SPE calculations in (**a**), (**b**) and (**c**): 0.01, 1.4, 21.0.
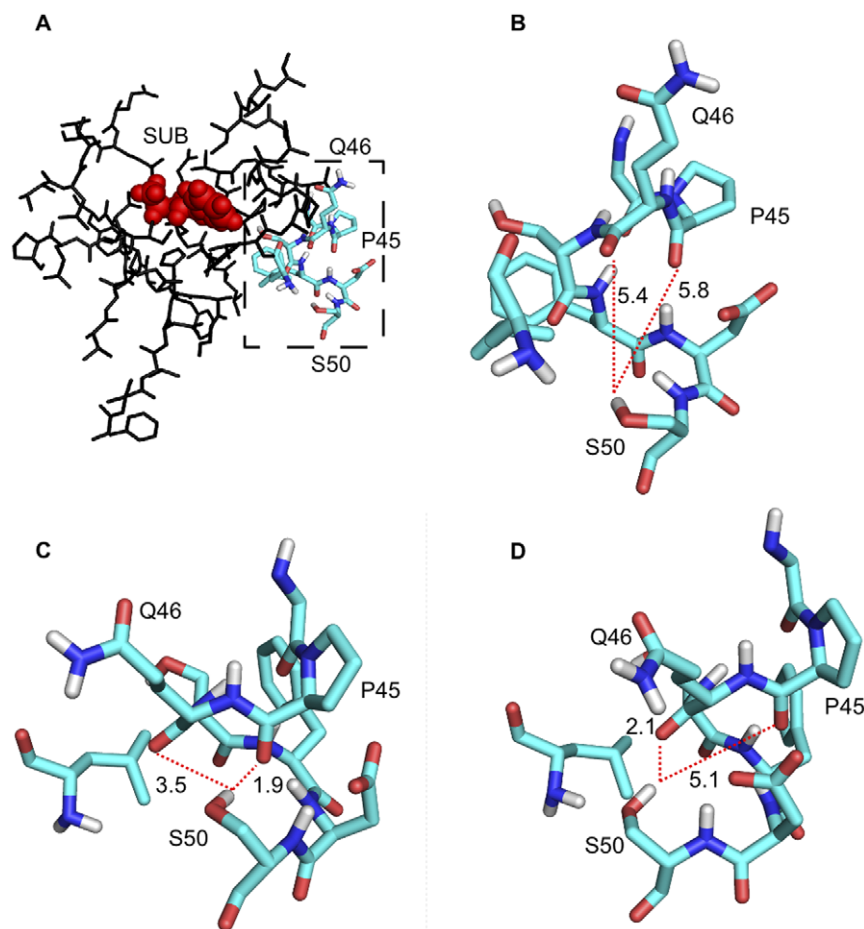doi:10.1371/journal.pone.0049849.t002

Figure 5. Hydrogen bonding network rearrangement. A. Model (**b**) overview showing the location of the residues undergoing different rearrangement in optimizations using NDDO cutoff of 12 or 15 Å, respectively. SUB: Substrate. B. Detail view of initial starting geometry. C./D. Hydrogen bonding network after optimization using NDDO cutoff of 12 Å or 15 Å, respectively.
doi:10.1371/journal.pone.0049849.g005

Based on these experiences and the results from above, it is therefore required to estimate a transition state, as described below. In the following, the notation M(C15, G0.5)'' means that a geometry optimization is carried out with the NDDO cutoff set to 15 Å and the gradient convergence criterion is set to 0.5 kcal/(molÅ) using the MOZYME LMO method. In this section all calculations are referring to the wild type (WT) structure.

In the procedure, first the molecular model of the **TI** is generated as described above. The **TI** model is then optimized with M(C15, G0.5). The optimized **TI** is then used as a template for the structure of the **ES** complex. To generate a model for the **ES** structure, the covalently bound substrate is replaced by the non-bonded, planar substrate and H^{γ} of S105 is transferred back onto O^{γ} of S105 using molecular modeling. The **ES** structure is
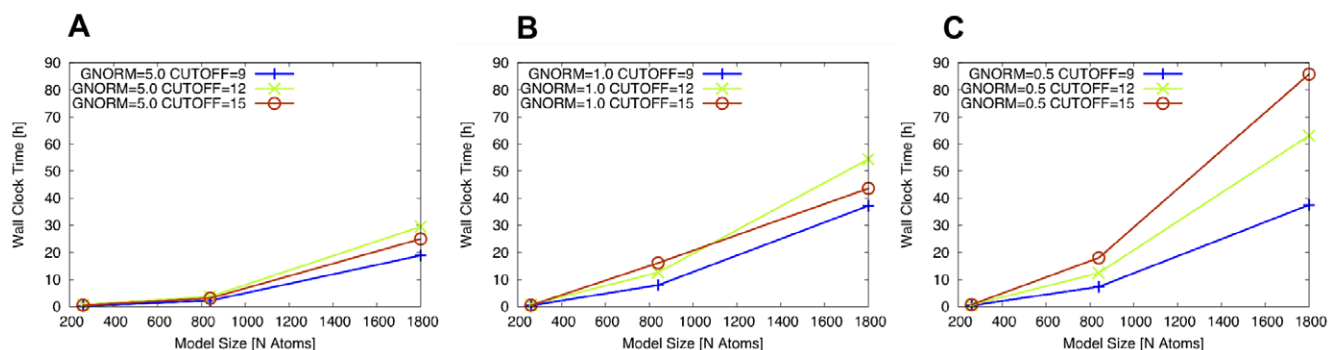

Figure 6. Time comparison for MOPAC configuration. Wall clock time requirements for optimization of tetrahedral intermediate using different GCC and model sizes.
doi:10.1371/journal.pone.0049849.g006

then optimized with M(C15, G0.5). These two optimized reaction end point structures are used in the linear interpolation scheme. To assess which distance between substrate C20 and O^γ of S105 is appropriate in the starting geometry of the **ES** complex, a number of different starting geometries were generated and optimized using M(C15, G0.5). The distance beteween C20 and O^γ in these starting geometries was varied in the range from 2.8 to 4.1 Å. The average distance of the optimized geometries is observed to be 3.55 Å and based on this, the distance between C20 and O^γ in the starting geometry of the **ES** complex was set to 3.5 Å. No significant differences in energy were observed for the optimized geometries of the different **ES** complexes.

The linear interpolation is carried out by dividing the geometrical distance between all atom pairs, $q_i^{TI} - q_i^{ES}$, where q_i is any of the cartesian coordinates of the atom i, by 10 and adding this difference incrementaly to q_i^{ES}. Every interpolation frame generated by this procedure is then optimized with M(C15, G0.5) where in each frame, the distance between O^γ and C20 of the substrate is kept fixed during the optimization. The separation O^γ/C20 is considered as defining the reaction coordinate and is fixed to a given value in a specific interpolation frame. The distance between C20 in the **ES** complex and C20 in the **TI** is observed to be 2.2 Å. The division of this distance by 10 interpolation steps leads to translation of C20 by 0.22 Å towards O^γ of S105 in each interpolation frame. To test for convergence with MOPAC configuration, every interpolation frame is also optimized with M(C12, G1.0) and M(C09, G5.0), where the same atom pair is kept fixed during the optimization. The structure corresponding to the highest point on the obtained energy profile estimate is considered as the approximation to the **TS**. This estimate is further analysed below. The estimated barriers for three MOPAC configurations are shown in Fig. 7.

The estimated barrier of 6.0 kcal/mol (using M(C15, G0.5)) is compared to a free energy of activation of 17.8 kcal/mol for the formation of tetrahedral intermediate in a high level QM/MM study [22] of trypsin and 15–20 kcal/mol in experimental studies [23]. The observed difference is possibly explained by the way the **ES** complex is modeled. In our presented approach, the molecular model of the **ES** complex is based on the optimized model of the
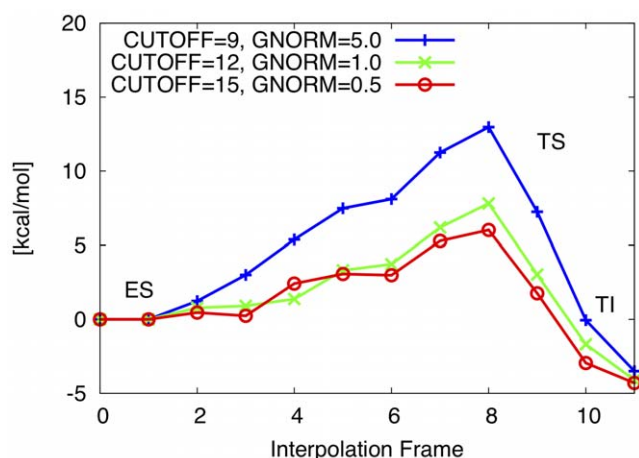

Figure 7. WT reaction barrier estimation. Convergence of estimated reaction barrier in WT. Estimated barriers are observed to converge to a lower boundary with strict GCC and higher NDDO cutoff. All constraints discarded in first and last interpolation frame. Estimated barriers are (in kcal/mol) PM6//M(C9, G5.0): 13.0, PM6//M(C12, G1.0): 7.8, PM6//M(C15, G0.5): 6.0.
doi:10.1371/journal.pone.0049849.g007

TI. By placing the non-covalently bonded substrate into the active site of the **TI**, a perturbation of this structure is introduced. However, the overall geometrical configuration of the active site is still very likely to the **TI** state (which itself is based on the crystal structure of the enzyme with covalently bound tetrahedral inhibitor) and therefore the optimization of the model can not completely leave the local minimum of the **TI** and arrive at the **ES** state with lower energy.

Given the very similar structure found for the enzyme-substrate complex for virtually all mutants, the effect of using a higher energy conformation on the barrier height will likely cancel. As a result it will have a relatively small effect on the relative barrier heights, which is the key parameter in this study. However, this is another approximation invoked to keep the method efficient.

It has to be noted that since the starting geometry for the M(C9, G5.0) and M(C12, G1.0) calculations is the optimized geometry from the M(C15, G0.5) calculation (of the stationary points), the optimized hydrogen bonding network is not expected to restructure. This is the reason why the **TI** obtained from optimizing with M(C12, G1.0) does not have the same relative energy as in Table 2, where the structure obtained from M(C12, G1.0) is lower in energy than the one obtained from M(C15, G0.5).

The estimated barriers using M(C12, G1.0) and M(C15, G0.5) are characterized by the same shape, while the estimated barrier using M(C9, G5.0) is significantly different. The apparent difference when going from less strict to strict gradient convergence is possibly explained by the fact that the PES of the system contains a huge amount of local minima. Using strict gradient convergence, it is ensured that also those parts of the gradient corresponding to shallow local minima are minimized. This in turn is apparently responsible for quite significant lowering of overall energy of the system.

From the above, it can be concluded that using a NDDO cutoff of at least 12 Å and a gradient convergence criterion of at least 1.0 kcal/(molÅ) is required for converged estimation of the reaction barrier.

Transition state verification

The optimized interpolation frame corresponding to the highest point on the energy profile (Frame 8 in Fig. 7 of the M(C15, G0.5) calculation) is subjected to partial Hessian vibrational analysis [24] (PHVA) using PM6 (without MOZYME, this function is provided by the FORCETS keyword in MOPAC). One imaginary frequency is found ($91.9i\text{cm}^{-1}$). The normal mode vibration is sketched in Fig. 8. An animation of the vibration is available in Text S1.

It has to be noted that the distance O^γ/C20 is constrained in the interpolation and results from the (arbitrary) division of the reaction coordinate into ten interpolation frames. Nevertheless, in the interpolation frame 8, the distances of S105 O^γ/C20 and O^γ/H^γ are 1.88 Å (fixed) and 1.27 Å (optimized), respectively, which is in very close agreement to the transition state distances found in model (**1**) using PM6 (Fig. 3). It can be concluded that the highest point on the reaction barrier estimate occurs at a geometry which is quite similar to the completely optimized **TS** structure of model (**1**).

Carrying out partial Hessian vibrational analysis using MO-ZYME LMOs returns only positive frequencies. Also it is observed, that all frequencies are positive after carrying out a partial transition state search for the atoms of the PHVA in the optimized interpolation frame 8.
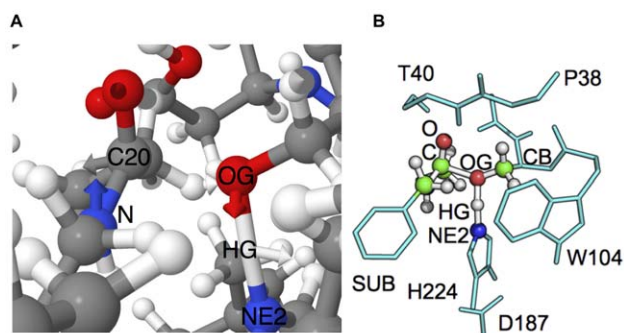
Figure 8. Illustration of PHVA in WT TS. A. Normal mode vibration on C20 of substrate towards S105 O^γ (OG), H^γ (HG) normal mode vector is towards H224 (NE2). B. Atoms included in PHVA shown as spheres (16 atoms in total). PHVA required 24.3 h of wall clock time.
doi:10.1371/journal.pone.0049849.g008

Comparison of PM6 and MOZYME energies

In the MOZYME method, in geometry optimization step τ_i, where $i > 0$, the LMOs from the step τ_{i-1} are used as the starting LMOs in the SCF procedure. The error originating from the truncation of the LMOs in step τ_{i-1} is therefore also present in the SCF cycle of the τ_i step. This can lead to different MOZYME and PM6 energies and differences in the estimated reaction barriers. In principle, this effect is avoided if the energy of the final geometry is evaluated using the 1SCF keyword to form a reorthogonalized set of LMOs, see Fig. 9.

As shown, the loss of orthogonality increases with the number of SCF cycles required in the geometry optimization. This is apparent in frames 0 and 11 of Fig. 9. The number of complete SCF cycles in these frames are 494 and 1896, respectively, compared to 25 (frame 1) and 437 (frame 10). Further comparisons between $\Delta_f H$ values obtained using different NDDO cutoff distances compared to PM6 are given in Text S1.
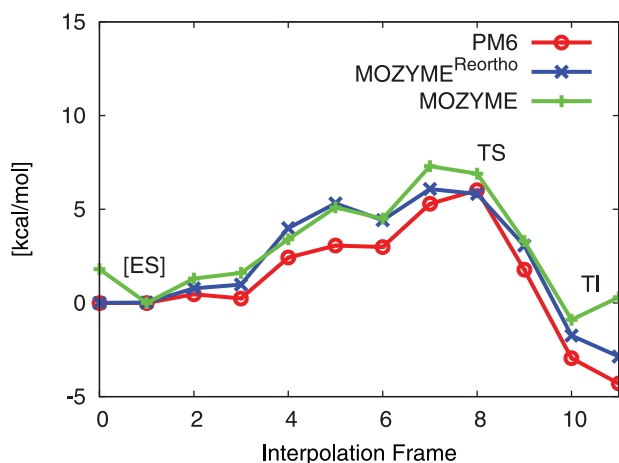

Figure 9. Comparison of PM6 and MOZYME Δ_fH. Comparison of $\Delta_f H$ using the final MOZYME LMOs, MOZYME LMOs after reorthogonalization and delocalized orbitals with PM6. Increases for $\Delta_f H$ in frames 0 and 11 of the MOZYME curve are due to loss of orthogonality of the LMOs. Calculations involving MOZMYE are done using M(C15, G0.5). Estimated barriers are (in kcal/mol) MOZYME: 7.3, MOZYME^{Reortho}: 6.1, PM6: 6.0.
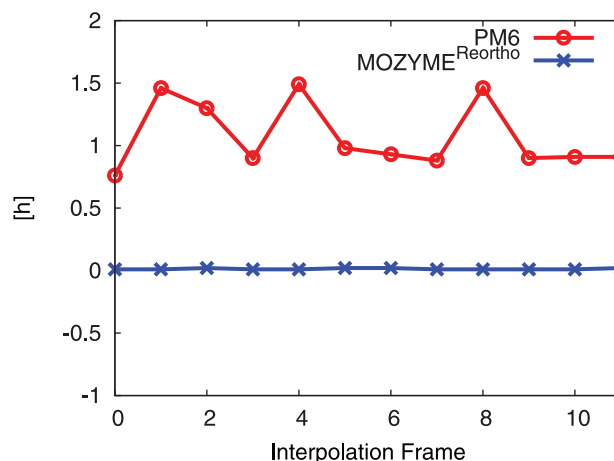doi:10.1371/journal.pone.0049849.g009

Figure 10. Comparison of CPU time required for PM6 and MOZYME SPE calculation. Average CPU time (in h): 1.07 (PM6), 0.01 (MOZYME).
doi:10.1371/journal.pone.0049849.g010

The required computational time to calculate single point energies using MOZYME is significantly different compared to using non-localized MOs, see Fig. 10.

Variant model preparation and single mutation screening

In the optimized model of the stationary points of the wild type, the molecular model of the variant v is generated by mutating the respective position in the backbone using the PyMOL [25] *Mutagenesis Wizard* function. The two molecular models ($\mathbf{ES}^v$ and $\mathbf{TI}^v$) are then used in a similar linear interpolation scheme as described for the wild type above. To illustrate the approach, the (single) mutations G39A, T103G and W104F are studied. Of the three discussed variants, G39A and W104F are located in the active site, C^α of T103G is located 8.7 Å away from O^γ of S105.

After introducing the mutation, the atoms of the new side chain are adjusted by molecular modeling to be in overlay with the wild type side chain and to fit into the available geometrical space. Each amino acid of the protein is then stored into a separate PDB file (called "fragment", (1) in Fig. 11). The water molecules and the substrate are stored as separate PDB files as well. By substituting the PDB fragment of the wild type at a given position by the fragment PDB file of a mutated side chain, the PDB structure file of a mutated enzyme can be assembled ((2) in Fig. 11).

In the optimization of the interpolation frames of the variants, it was observed that the introduction of a big side chain in the active site can lead to significant rearrangment of side chains on the surface of the molecular model. From this, the bonding topology between the wild type and the mutant can become significantly different and lead to reaction barrier estimates with unconclusive shapes. It was therefore required to fix the atoms of a number of side chains on the surface of the molecular model to remain in the position of the optimized wild type structure. In particular, the side chains of the residues S50, P133, Q156, L277 and P280 are fixed. Other than the constraints on the atoms of the reaction coordinate (which are removed in frames 0 and 11), these constraintes also remain effective in the optimization of the reaction end points. The reaction barrier estimations obtained after carrying out the linear interpolation and the constrained optimization of the variant structures are presented in Fig. 12.

The reaction energy profile of the G39A mutant shows a slight decrease in energy at interpolation frame 3. This is explained by
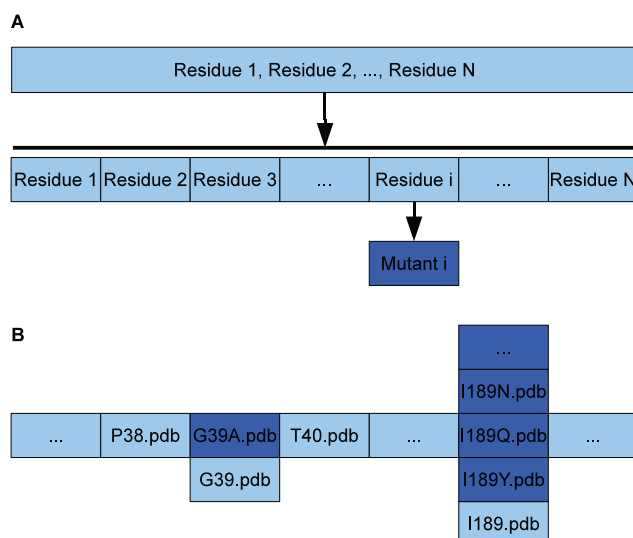
A

| Residue 1, Residue 2, ..., Residue N | | | | | | |

| Residue 1 | Residue 2 | Residue 3 | ... | Residue i | ... | Residue N |

Mutant i

B

...	P38.pdb	G39A.pdb	T40.pdb	...	I189Q.pdb	...
		G39.pdb			I189Y.pdb	
					I189.pdb	

(... I189N.pdb above I189Q.pdb)

Figure 11. Mutant file assembly. A. Preparation of fragment PDB files of WT and mutant. B. Sketch of variant structure file assembly from fragment PDB files using CalB. Light blue boxes indicate WT amino acids, dark blue boxes indicate variant side chains. The figure illustrates the hypothetical double mutation G39A-I189Q.
doi:10.1371/journal.pone.0049849.g011

the presence of a local minimum with lower energy than the initial **ES** state which becomes available to the system at the third step of the interpolation. However, this decrease in energy is not observed in the optimizations using M(C9, G5.0) or M(C15, G0.5). A similar effect is observed in the profile of the T103G mutant, both for the calculations using M(C9, G5.0) and M(C12, G1.0).

The estimated barrier of the G39A mutant is very close to the WT barrier and the lowest of all three mutants. Based on this, it would be concluded that the G39A mutant is the most likely candidate for showing increased overall activity. The complete approach outlining the various steps included in the presented screening technique is summarized in the overview Fig. 13.

Time requirements

It was observed that a significant amount of CPU time can be saved by basing the molecular model of the **ES**^{WT} and the variants on the optimized **TI** of the wild type. Since the molecular model of the wild type is based on the crystal structure, a major proportion of the structure is already optimized when the mutation is introduced. In Fig. 14, it is shown how the required wall clock time for the optimization of the wild type and three variants depends on the interpolation frame.

In the figure, a trend towards higher time requirements for the interpolation frames for the non-stationary points is observed. The average time per interpolation frame is highest in the G39A mutation. This appears reasonable considering the fact that a sterically demanding group is being introduced into a restricted environment, which requires considerable rearrangement of the surroundings. The time requirement in all three variants is greatly reduced by basing the molecular model of the variant on the optimized structure of the **TI** of the wild type. Also, the optimization of frame 1 of the wild type appears to require only very little CPU time (0.1 h). This is explained by its high similarity to frame 0, which is completely optimized already.

It is worth noting, that the interpolation frames can be optimized in parallel and thus the CPU time requirement for the evaluation of the energy profile is only determined by the optimization of that interpolation frame with the highest wall clock time.

Conclusions

A fast computational enzyme activity screening method is presented. The method is designed towards the efficient estimation of the barrier height of an enzymatic reaction of a large number of mutants. Based on the presented approach, the barrier height of a mutant can be computed within 24 hours on roughly 10 processors. In the approach, the PM6 method as implemented in the MOPAC2009 program is used. The approach is tested and applied to the study of the first step of the amide hydrolysis reaction as catalysed by *Candida Antarctica* lipase B (CalB). In particular we show that

1. PM6 reproduces the RHF/3-21G transition state (**TS**) structure (Fig. 3) and B3LYP/6-31G(d)//RHF/3-21G barrier height (Table 1) for a small model system.
2. PM6 combined with the MOZYME method can be used to geometry optimize a structural model containing all residues within 8 Å of the active site (Fig. 4b) in about 18 hours on a single processor (Table 2). A gradient convergence criterion of 0.5 kcal/(molÅ) and a NDDO cutoff distance of 15 Å are needed for reliable results.
3. The **TS** search algorithm implemented in MOPAC2009 was found too computationally demanding and not consistently reliable for our purposes. Instead we devised an adiabatic
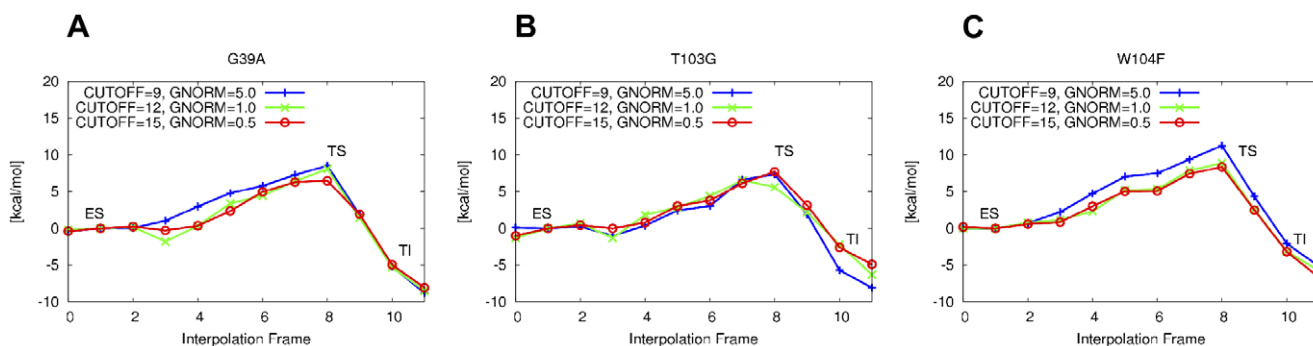

Figure 12. Single mutation screening between TS and ES in (b). Electronic energy difference not corrected for ZPE. The difference is defined by locating the highest point on the PES and subtracting the energy of the lowest point before it. Energy differences from PM6 SPE calculations of M(C15, G0.5) optimized geometries (in kcal/mol): A. G39A: 6.9. B. T103G: 8.7. C. W104F: 8.3. For comparison (see Fig. 7) WT: 6.0.
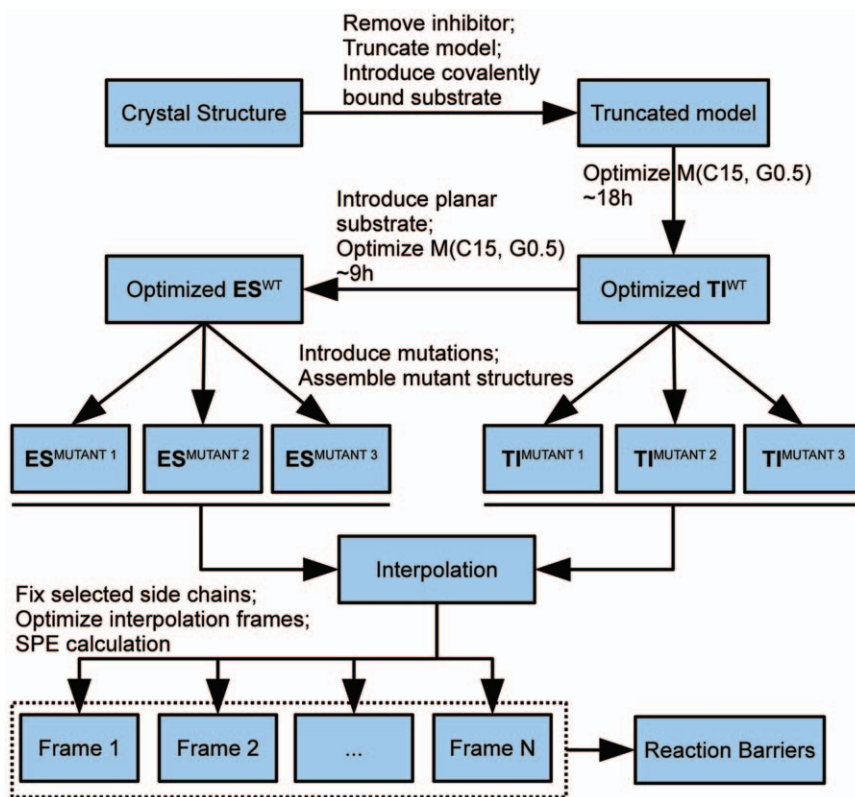doi:10.1371/journal.pone.0049849.g012

Figure 13. Calculation setup. Interpolation of WT model (not indicated) is between optimized **ES** and **TI** structure.
doi:10.1371/journal.pone.0049849.g013

mapping method for estimating the **TS** structure and barrier height (Fig. 7), where key bond lengths are kept constrained at a series of intermediate values while the rest of the protein structure is optimized using MOZYME. The optimized geometries are then used for conventional (i.e. not MOZYME) PM6 single point energies, because the energy difference

between conventional PM6 and MOZYME-PM6 is too large compared with the effect of mutations (Figs. 9, 12).

4. The average CPU time needed per point on the energy profile is 4–5 hours on a single processor (Fig. 14) and each point can be computed independently leading to trivial parallelization (Fig. 13).

5. Both the preparation of input files for the optimization of all interpolation frames on the reaction coordinate as well as the generation of energy profiles are automated to a large degree. In the current setup, manual effort is required only in the molecular modeling of the mutant side chain fragment PDB files, Fig. 11, and the molecular modeling of the substrate in the non-covalently bound reactant state. However, since a side chain fragment for a given mutant can be used in any number of combination mutants including this mutation, the required manual effort only scales with the number of distinguishable point mutations.
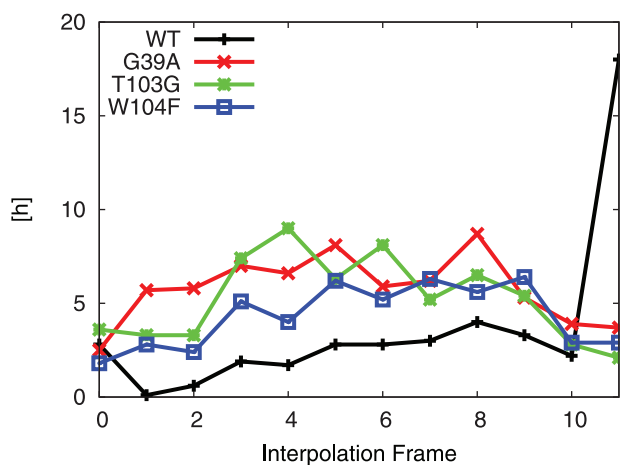
The method described here is *in principle* generally applicable to efficiently identify promising mutants for further study for any enzyme-catalyzed reaction for which the structure is known and which does not involve open-shell species (which can not currently be handled with MOZYME). When applying the method to a new system it is of course important to re-check the validity of using the PM6 method by, for example, comparison to *ab initio* results for small model systems, as was done here. In addition, the usual caveats associated with *all* computational studies of enzymatic reactivity apply: identifying a reaction coordinate that uniquely defines the mechanism can be difficult and is ultimately a matter of trial and error. Mechanisms that involve large structural rearrangements of the enzyme and/or large changes in solvation

Figure 14. Wall clock time for geometry optimization. Optimization of **TI**^{WT} requires more time than mutant **ES** and **TI** structures. Average wall clock time per interpolation frame (in h): 4.1, 5.8, 5.2, 4.3 for WT, G39A, T103G and W104F, respectively. All optimizations done using M(C15, G0.5).
doi:10.1371/journal.pone.0049849.g014

energy are difficult to model accurately, and the predicted effects of mutations may be less reliable.

As an initial application, the barrier heights of nearly 400 single to four-fold combination mutants in CalB have been estimated and, for 22 mutants, compared to experimentally measured activities with promising results (a preprint of this as yet unpublished study is available at http://arxiv.org/abs/1209.4469).

Supporting Information

Text S1 Difference MOZYME/PM6 energies; Amino acid sequences in models (**a**), (**b**) and (**c**); Transition state verification animation and URL of repository for modeling scripts.
(PDF)

Author Contributions

Conceived and designed the experiments: MRH LDV WB AS JHJ. Performed the experiments: MRH. Analyzed the data: MRH. Wrote the paper: MRH.

References

1. Claeyssens F, Harvey J, Manby F, Mata R, Mulholland A, et al. (2006) High-Accuracy Computation of Reaction Barriers in Enzymes. Angewandte Chemie 118: 7010–7013.
2. Parks J, Hu H, Rudolph J, Yang W (2009) Mechanism of cdc25b phosphatase with the small molecule substrate p-nitrophenyl phosphate from qm/mm-mfep calculations. The Journal of Physical Chemistry B 113: 5217–5224.
3. Hermann J, Pradon J, Harvey J, Mulholland A (2009) High level qm/mm modeling of the formation of the tetrahedral intermediate in the acylation of wild type and k73a mutant tem-1 class a β- lactamase. The Journal of Physical Chemistry A 113: 11984–11994.
4. Noodleman L, Lovell T, Han W, Li J, Himo F (2004) Quantum chemical studies of intermediates and reaction pathways in selected enzymes and catalytic synthetic systems. Chemical reviews 104: 459–508.
5. Syrén P, Hult K (2011) Amidases have a hydrogen bond that facilitates nitrogen inversion, but esterases have not. ChemCatChem 3: 853–860.
6. Tian L, Friesner R (2009) Qm/mm simulation on p450 bm3 enzyme catalysis mechanism. Journal of chemical theory and computation 5: 1421–1431.
7. Altarsha M, Benighaus T, Kumar D, Thiel W (2010) Coupling and uncoupling mechanisms in the methoxythreonine mutant of cytochrome p450cam: a quantum mechanical/molecular mechanical study. Journal of Biological Inorganic Chemistry 15: 361–372.
8. Friesner R, Guallar V (2005) Ab initio quantum chemical and mixed quantum mechanics/molecular mechanics (qm/mm) methods for studying enzymatic catalysis. Annu Rev Phys Chem 56: 389–427.
9. Klamt A, Schüürmann G (1993) Cosmo: a new approach to dielectric screening in solvents with explicit expressions for the screening energy and its gradient. J Chem Soc, Perkin Trans 2:799–805.
10. Stewart J (2007) Optimization of parameters for semiempirical methods v: Modification of nddo approximations and application to 70 elements. Journal of Molecular Modeling 13: 1173–1213.
11. Stewart J (1990) Mopac: a semiempirical molecular orbital program. Journal of Computer-Aided Molecular Design 4: 1–103.
12. Stewart J (1996) Application of localized molecular orbitals to the solution of semiempirical selfconsistent field equations. International Journal of Quantum Chemistry 58: 133–146.
13. Dewar M, Zoebisch E, Healy E, Stewart J (1985) Development and use of quantum mechanical molecular models. 76. am1: a new general purpose quantum mechanical molecular model. Journal of the American Chemical Society 107: 3902–3909.
14. Stewart J (1989) Optimization of parameters for semiempirical methods i. method. Journal of Computational Chemistry 10: 209–220.
15. Rocha G, Freire R, Simas A, Stewart J (2006) RM1: A reparameterization of AM1 for H, C, N, O, P, S, F, Cl, Br, and I. Journal of Computational Chemistry 27: 1101–1111.
16. Schmidt M, Baldridge K, Boatz J, Elbert S, Gordon M, et al. (1993) General atomic and molecular electronic structure system. Journal of Computational Chemistry 14: 1347–1363.
17. Dewar M, Thiel W (1977) Ground states of molecules. 38. the mndo method. approximations and parameters. Journal of the American Chemical Society 99: 4899–4907.
18. Pople J, Beverdige D (1970) Approximate Molecular Orbital Theory. New York: McGraw-Hill.
19. Uppenberg J, Oehrner N, Norin M, Hult K, Kleywegt G, et al. (1995) Crystallographic and molecular-modeling studies of lipase b from candida antarctica reveal a stereospecificity pocket for secondary alcohols. Biochemistry 34: 16838–16851.
20. Schenker S, Schneider C, Tsogoeva S, Clark T (2011) Assessment of popular dft and semiempirical molecular orbital techniques for calculating relative transition state energies and kinetic product distributions in enantioselective organocatalytic reactions. Journal of Chemical Theory and Computation.
21. Stewart J (2009) Application of the pm6 method to modeling proteins. Journal of molecular modeling 15: 765–805.
22. Ishida T, Kato S (2003) Theoretical perspectives on the reaction mechanism of serine proteases: the reaction free energy profiles of the acylation process. Journal of the American Chemical Society 125: 12035–12048.
23. Fersht A (1999) Structure and mechanism in protein science: a guide to enzyme catalysis and protein folding. New York: WH Freeman.
24. Li H, Jensen J (2002) Partial Hessian vibrational analysis: the localization of the molecular vibrational energy and entropy. Theoretical Chemistry Accounts: Theory, Computation, and Modeling (Theoretica Chimica Acta) 107: 211–219.
25. The PyMOL Molecular Graphics System, Schrödinger LLC (2010).
26. Søndergaard C, Olsson M, Rostkowski M, Jensen J (2011) Improved treatment of ligands and coupling effects in empirical calculation and rationalization of p k a values. Journal of Chemical Theory and Computation 7: 2284–2295.