ACS Catalysis

pubs.acs.org/acscatalysis

Research Article

Combinatorial Approach for Exploring Conformational Space and Activation Barriers in Computer-Aided Enzyme Design

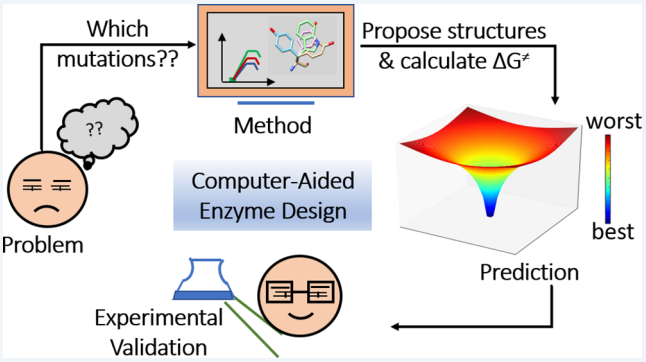
Dibyendu Mondal, Vesselin Kolev, and Arieh Warshel*

Cite This: *ACS Catal.* 2020, 10, 6002−6012

Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information

ABSTRACT: Computer-aided enzyme design is a field of great potential importance for biotechnological applications, medical advances, and a fundamental understanding of enzyme action. However, reaching a predictive ability in this direction is extremely challenging. It requires both the ability to predict quantitatively the activation barriers in cases where the structure and sequence are known and the ability to predict the effect of different mutations. In this work, we propose a protocol for predicting reasonable starting structures of mutants of proteins with known structures and for calculating the activation barriers of the generated mutants. Our approach also allows us to use the predicted structures of the generated mutant to predict structures and activation barriers for subsequent set of mutations. This protocol is used to examine the reliability of the *in silico* directed evolution of Kemp eliminase and haloalkane dehalogenase. We also used the results of single and double mutations as a base for predicting the effect of transition-state stabilization by multiple concurrent mutations. This strategy seems to be useful in creating an activity funnel that provides a qualitative ranking of the catalytic power of different mutants.

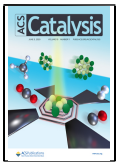


KEYWORDS: *computer-aided enzyme design, Kemp eliminase, dehalogenase, empirical valence bond, distinct rotamer generation, directed evolution*

1. INTRODUCTION

Designing a new protein with targeted functionality has wide implications in chemistry and biology. Progress has been made in recent years in designing proteins by various methods, including starting from scratch (*de novo* design) or improving the function of a protein scaffold by mutating rationally or randomly.[1,2] While such methods are promising, they encounter major challenges. For instance, *de novo* design of a protein presents an exceptionally difficult problem, since only a few out of infinite possibilities in sequence space can lead to a stable and functional protein. Finding arbitrarily that right sequence could be like searching for a needle in a haystack.[3] Of course, finding a reasonable sequence still requires the ability to calculate the catalytic power of the generated protein. Arguably, success is more likely to be achieved by template-based designs (rational or directed evolution). In fact, various design strategies have been used to improve the activity of already existing proteins.[4,5] A related recent study explored the ability to design enzyme active sites by remodeling a reasonable catalytic site, first reducing its activity by introducing one or more mutations and then increasing it by adding another mutation.[6] Directed evolution is another template-based method, and its success is much less dependent on a prior knowledge about the system.[7] The implementation of such a method usually involves a start from a low-functionality protein and then iteratively applying random mutagenesis and screening processes, to generate proteins with improved activities.[8] The greatest challenge with implementing a directed evolution is the vastness of the sequence space that needs to be screened, while performing the random mutations. This problem is overcome by biological evolutions due to the long time available for such processes. An informed screening can reduce the number of possible random mutations and enhance the predictability of the process. Thus, it is tempting to rely on computational tools to guide the directed evolution or to provide alternative directions.[2,9−14] While some of the available computational tools target the structural stability and sequence−activity relation,[2,9,10,12] information about the reaction energetics (transition state stability) is in principle a more reasonable direction *to guide directed evolution*. Even though some QM/MM studies have been proven to provide a partial rationale to directed evolution of proteins, they are restricted to cases with a relatively small number of mutation

Received: March 13, 2020
Revised: April 17, 2020
Published: April 27, 2020



https://dx.doi.org/10.1021/acscatal.0c01206
ACS Catal. 2020, 10, 6002−6012

possibilities, due to the high computational cost.[15] Furthermore, all energy calculation methods mentioned are limited by the need to employ an adequate sampling method and a protocol to avoid encountering a nonoptimal starting structure (wrong conformation of mutated residues). In case of processing nonoptimal starting structures, even very long simulations are not sufficient to avoid ending up at a kinetically trapped state.

In previous studies we relied on a semiempirical QM/MM approach—the empirical valence bond (EVB) method, which is arguably the most effective way of simulating chemical reactions in the condensed phase.[16] This method has been widely implemented previously to explore complex reactions and has successfully reproduced experimental findings without the need of extensive computational resources. However, the performance of this method in reproducing the observed results of directed evolution has not yet been established. For example, our recent study of Kemp eliminase[17] was able to reproduce the effects of sequence change by directed evolution but did not allow us to move in a correct way between the structures of different sequences. That is, using the structure of a given sequence we were able to obtain reasonable results, but trying to move from a structure that corresponds to one sequence to one that corresponds to another has not reproduced the proper change in structure and catalytic effect. Thus, it is crucial to advance a formulation that could be useful for predicting the effect of multiple mutations.

Predicting the effect of multiple mutations on a chemical step by computer simulations also depends on the starting structures provided. Considering initially a single conformer for each mutated residue can lead to the generation of problematic structures, which are kinetically trapped in some local minima (close to the actual ground-state minima). To avoid that kind of problem, it is essential to develop a procedure that would generate multiple distinct rotamers for each mutation. Such rotamers should then be used as starting structures for EVB simulations. Although such an approach is formally correct, it might lead to an unrealistically high number of simulations, due to the large amount of mutations that may be needed to be examined. Hence, we need some type of enhanced sampling that will allow us to estimate the effect of multiple mutations without simulating all the possible combinations. One possibility is to use sets of single and double mutations to predict the effect of multiple mutations.

In this work we implement the concept of using a combinatorial approach to explore the rotamer conformational space, combined with calculations of activation free energies. This approach is examined here in two test cases, namely the reactions of Kemp eliminase and haloalkane dehalogenase (DhlA). Our results show that the method can be useful for predicting mutational effects on the change of the rate of the chemical steps of enzymatic reactions. It is also demonstrated that the method can be implemented as a useful tool in computational directed evolution.

However, we wish to point out that the focus of this work is not about getting perfect results in modeling the effects of sequence changes but rather about examining the error range of a procedure that considers a very extensive configurational sampling in studying mutational effects.

2. METHODS

Our simulation protocol consists of two main steps: (1) a distinct rotamer generation and (2) the use of EVB calculations in a pragmatic way to reduce the computational cost of the corresponding simulations.

2.1. Rotamer Generation. To explore efficiently the free energy landscape, it is important to generate multiple rotamers as starting structures for the specific residues of interest. Employing only one starting structure (one set of rotamers) may require an extremely long relaxation process to bring the simulated conformation of the given mutant close to the most stable configuration. Even if the high computational cost of the relaxation is not an obstacle, the process may converge away from the actual ground state, due to the ruggedness of the landscape. This implies a possible start from a high-energy ground state that will end up obtaining reasonable protein activity for the wrong reasons. Thus, it is necessary to examine each newly proposed mutation on the basis of several starting conformers. One possible option of the sampling protocol is to choose three rotamers for each mutation and thus generate 3^N starting structures for a case of N mutations. This should be done while it is kept in mind that the rotamers selected for the simulations should be as distinct as possible.

Our first step in that direction is to convert every residue Q_i selected for a mutation, into a coarse-grained (CG) representation, by replacing its side chain with an effective atom (named X).[18] The X atom is located at the geometric center of the side chain's heavy atoms of the residue Q_i (Figure 1A). The additional atom D (dummy atom) is introduced along the C_α–C_β bond (this bond exists in the all-atom form of the side-chain). The main-chain atoms are kept unaltered in the CG representation. Having those steps completed, the next one is the generation of the rotamers of the amino acid j to which the current residue Q_i is to be mutated. For the first rotamer of Q_j, the geometric center of the side chain's heavy atoms of Q_j is placed as close as possible to the X atom of the residue Q_i. In other words, for the first rotamer of Q_j, the X atom of the residue Q_j is placed close to the X atom of the residue Q_i. For the remaining two rotamers, the X atoms of the new rotamers are placed in a way to maximize the distance among the X atoms of all three rotamers. Finally, the explicit forms of the side chains are generated and their coordinates are optimized, by removing any potential intra- or intermolecular clashes. In the case of performing mutations at N positions of a sequence of a protein Q, the aforementioned steps are followed until the requested 3^N starting structures are generated.

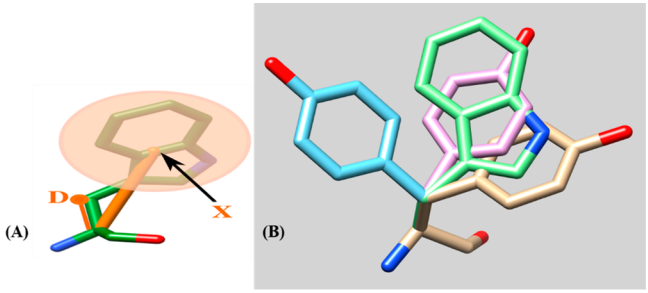


Figure 1. Application of the rotamer generation protocol for mutating a tryptophan residue Q_i to a tyrosine residue Q_j of a protein Q: (A) replacement of the side chain of tryptophan by an atom X and a dummy atom D, positioned along the C_α–C_β bond; (B) three distinctively different rotamers of the residue Q_j (pink, the rotamer closest to the residue Q_i (green); blue, the farthest rotamer). Refer to the text for more details.

https://dx.doi.org/10.1021/acscatal.0c01206
ACS Catal. 2020, 10, 6002−6012

The implemented search of the most distinguishable side-chain conformers is done with respect to the specific local environment inside the cavity that accommodates the given side chain. That environment is defined by the short- and long-range nonbonded potentials. However, we note that the adequacy of the CG simulations used in our procedure depends on the adequacy of the trimmed all-atom side chain. Thus, having adequate all-atom side chain conformation is very important when one generates mutations, because the number of mutations is very small with respect to the total number of residues. In this case, the proper assessment of the specific local environment is of primary importance and therefore a specific method for generating mutations is required. One may use standard rotamer libraries when the number of mutations is proportional to the number of residues in the simulated protein. In such a case, an external rotamer library, derived as an ensemble averaged over hundreds of protein molecules, might fit to the average environment. However, when the number of mutations is very small, only a specific approach, such as that implemented here, can bring adequate results. Our rotamer generation protocol is also adequate for large side chains. That is, if one can estimate the influence of the multipoles in the Coulomb expansion (up to octapoles), it is quite easy to see that the larger the side chain, the more precise the trimming approach, especially if restrained electrostatic potential (RESP) model based charges are employed.

2.2. Calculations of Activation Free Energies for a Protein with *N* Mutations. The EVB method is applicable for studying reaction kinetics in condensed phases and proteins. Here we use this method to calculate the activation free energies of a reaction (catalyzed by the protein Q) and to screen suitable structures for conducting new mutations of the protein Q. In order to calculate the activation free energy of the reaction ($\Delta G^{\ddagger}_{Q}$), we performed 3^N EVB simulations (based on 3^N starting structures generated by our rotamer generation protocol for mutating N residues of the protein Q). The activation free energy of the reaction, $\Delta G^{\ddagger}_{Q}$, is then given by

$$\Delta G^{\ddagger}_{Q} = \sum_{i=1}^{3^N} \Delta G^{\ddagger}_{Q_i} \frac{e^{-(\Delta G^{\ddagger}_{Q_i} - \Delta G^{\ddagger}_{min})/RT}}{\sum_{j=1}^{3^N} e^{-(\Delta G^{\ddagger}_{Q_j} - \Delta G^{\ddagger}_{min})/RT}} \quad (1)$$

Here $\Delta G^{\ddagger}_{Q_i}$ is the calculated activation free energy of the reaction, involving the ith starting structure. $\Delta G^{\ddagger}_{min}$ is the minimum of the calculated activation free energy, estimated after examining each of the 3^N generated starting structures. R and T are the gas constant and the simulation temperature, respectively. Ideally, the minimum of the calculated activation energies should be the activation energy of the reaction (i.e., $\Delta G^{\ddagger}_{min} \approx \Delta G^{\ddagger}_{Q}$), but it is also possible that other structures can be sufficiently preorganized in a way that they could contribute to the transition-state stabilization. Thus, $\Delta G^{\ddagger}_{Q}$ of eq 1 is calculated as an average over all starting structures that are used in the EVB calculations. It is worth noting that, even if the contributions of all the starting structures are included in eq 1, the structures that yield activation energies close to $\Delta G^{\ddagger}_{min}$ are those that contribute mostly to the $\Delta G^{\ddagger}_{Q}$. Note that the structure which corresponds to $\Delta G^{\ddagger}_{min}$ is used as the structure of the protein for introducing new mutations.

The above protocol of the EVB simulation has been applied here for the enzyme DhlA, known to catalyze the conversion of toxic haloalkanes to alcohols.[19] This enzyme supports the conversion of 1,2-dichloroethane into chloroethanol, in a process that consists of a series of steps, including an S_N2 reaction (Figure 2).[20] This S_N2 reaction is not the rate-

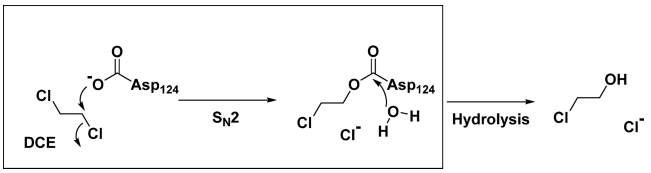


Figure 2. Illustration of the reaction in haloalkane dehalogenase. The square box marks one of the most important chemical steps, the S_N2 reaction, and DCE stands for dichloroethane.

determining step for the wild-type protein of DhlA,[21−23] whereas for many mutants of DhlA the S_N2 reaction was reported to be the rate-determining step.[6,23−25] Fortunately, the rate constants of the S_N2 reaction are experimentally known in cases where the S_N2 reaction is not rate determining. Thus, to validate the correctness and applicability of our EVB approach, we attempted to simulate the S_N2 chemical step and compared the calculated barrier of the reaction to the corresponding experimental results. Also, we paid special attention to the understanding of the cases where our previous calculations failed to reproduce the experimental finding.[6]

The DhlA system is clearly too simplistic to fully validate our approach, but it should be interesting enough to provide some important insights. Most of the mutations studied in DhlA are single mutations or at most double mutations. Our method can successfully reproduce most of the observed effects (see Results and Discussion), and the results encouraged us to go further and implement it to investigate a case that is more complicated.

The Kemp eliminase system (the proposed scheme of the reaction is given in Figure 3) has been a subject of extensive

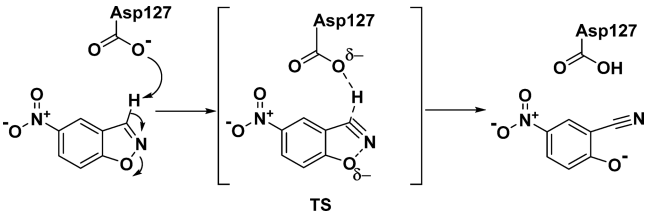


Figure 3. A scheme illustrating the base-catalyzed Kemp elimination of 5-nitrobenzisoxale. TS stands for transition state.

computational protein design as well as directed evolution studies.[26,27] Reference 27 demonstrates that 17 rounds of mutagenesis and screening on the *in silico* designed protein HG3 generates a protein, HG3.17, which decreases the activation energy of the reaction by ∼3.3 kcal/mol (in comparison to HG3).

The rate of the reactions for 5 mutants (generated using 17 rounds of mutagenesis and screening process) is reported in ref 27 (see Table 1), and the structures of the first and last systems are available in the Protein Data Bank.[27,28] Thus, we need to predict reasonable structures of the intermediate mutants to reveal how the enzyme has evolved from HG3 to HG3.17. One way to study the effects of different mutations in different stages of the mutagenesis is to mutate the residues using available software (starting from a known structure) and then study the reaction using the EVB.[17] On the other hand, we can start from the least active protein HG3, implement our method

https://dx.doi.org/10.1021/acscatal.0c01206
ACS Catal. 2020, 10, 6002−6012

Table 1. Point Mutations with Respect to HG3 in Different HG3 Variants and the Corresponding Experimental Activation Free Energies[a]

HG3 Variants	Point Mutations	$\Delta G^{\neq}_{exp}$ (kcal/mol)[b]
HG3	-	16.9
HG3.3b	V6I, K50H, M84C, S89R, Q90D, A125N	16.0
HG3.7	V6I, Q37K, K50Q, M84C, S89R, Q90H, A125N	14.1
HG3.14	V6I, Q37K, K50Q, G82A, M84C, Q90H, T105I, A125T, T142N, T208M, T279S, D300N	13.9
HG3.17	V6I, Q37K, N47E, K50Q, G82A, M84C, S89N, Q90F, T105I, A125T, T142N, T208M, F267M, W275A, R276F, T279S, D300N	13.6

[a]The typeface represents whether a specific position of the sequence has been mutated one (black), two (blue), or three (brown) times. [b]All experimental $\Delta G^{\ddagger}$ values are calculated from the k_{cat} values reported in ref 27, assuming T = 300 K.

to predict the most probable structure (which corresponds to $\Delta G^{\ddagger}_{min}$) in one step, and then use it in the next round of prediction to move forward (similar to natural evolution). The implemented protocol is schematically presented in Figure 4. More details regarding its design and the related EVB parameters are available in the Supporting Information.

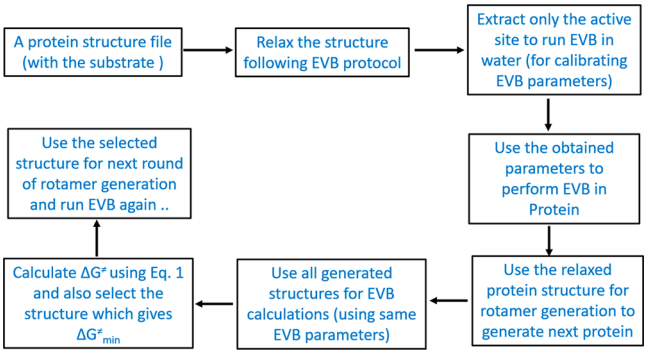


Figure 4. Schematic presentation of the stages and internal subroutines included in the implemented simulation protocol.

All Python and Bash scripts and some input files that were used in this project can be found in the git repository (https://github.com/dibyendu92/In-silico-Enymze-deisgn).

3. RESULTS AND DISCUSSION

In this section, we examine the results obtained for dehalogenase and Kemp eliminase.

3.1. Haloalkane Dehalogenase. Our protocol is tested first on the DhlA system. In an earlier study from our group,[6] we showed both computationally and experimentally that, while W175Y reduces the catalytic efficiency, another mutation, E56N, restores the catalysis.[6] In that study, we were unable to reproduce the effect of the double mutation W125F/V226Q. Therefore, we paid a special attention here to this mutation. The results that are presented in Table 2 show that, in most cases, the predictions given by our protocol are in good agreement with the corresponding experimental findings.

The results obtained for double mutations are showing the maximum deviation from the experimental results. In our previous work, we suggested one possible reason for the mismatch observed in the case of W125F/V226Q mutation.[6] Our current calculated result for the W125F/V226Q mutation is in better agreement with the observed result, in comparison

Table 2. Experimental and Calculated Activation Free Energies for the S_N2 Step in the Reaction of Haloalkane Dehalogenase[a]

mutation	$\Delta G^{\ddagger}_{cal}$ (kcal/mol)	$\Delta G^{\ddagger}_{exp}$[b] (kcal/mol)	$\lvert\Delta G^{\ddagger}_{exp} - \Delta G^{\ddagger}_{cal}\rvert$
wild type	14.7	15.3[c]	0.6
V226A	15.9	16.0[c]	0.1
W125F	17.5	17.6	0.1
W175F	18.4	18.3	0.1
F172Y	16.1	17.3	1.2
F175W	16.3	16.8	0.5
F164A	21.9	19.4	1.5
E56Q	15.3	15.9	0.6
W175Y	21.6	18.3	3.6
W175Y/E56N	18.4	15.6[c]	2.8
W125F/V226Q	17.5	n.a.	

[a]The last column represents the absolute difference between the experimental and calculated activation free energies. n.a. = not active. [b]The activation barriers are for the S_N2 step. All experimental activation free energies are taken from the compilation in ref 6 (see also refs 13, 14, 29, and 30 in ref 6). The activation free energies were calculated for a temperature of 300 K. [c]Reactions where the S_N2 step is not the rate-determining step.

to the results in ref 6 (significantly higher calculated barrier in the current work). However, we still cannot reproduce the inactivity that might be due to partial unfolding.[6] We also note that the calculated results for W175Y and W175Y/E56N are overestimated by ∼3 kcal/mol. One possible explanation for observing this deviation could be the selection of the model adopted for generating rotamers in the case of Y175. That is, our model generates side-chain samples by rotating the atoms about the C_α–C_β axis, deliberately ignoring the possible variation of the position of the aromatic ring. Since the wrong orientations of the aromatic ring in the starting structures of the rotamers could not be entirely fixed by reorientation during the simulation, they might cause the observed overestimations. Hence, an improved version of the protocol for sampling may be needed. At any rate because we found our overall results (for 8 out of 11 mutants) encouraging, we decided to explore the more complicated case of Kemp eliminase.

3.2. Kemp Eliminase. In the case of dehalogenase, most of the calculations used only a single mutation for checking the efficiency of our protocol and deciding whether eq 1 is applicable to calculate the activation energy ($\Delta G^{\ddagger}_{Q}$ for a protein Q). In the case of Kemp eliminase, our goal here is to check whether we can transform the protein from a less active to a more active form by an *in silico* approach, when we know in advance what mutations should be introduced during each step. Note that the overall decrease in activation free energies caused by introducing 17 mutations is only 3.3 kcal/mol (Table 1)[27] and an accurate reproduction of the effect of mutations is very challenging (because of that an accumulated error in any early step can influence the calculated results significantly). Below we discuss all the considerations we made to simulate the computational evolution from HG3 to HG3.17.

3.2.1. HG3. Three residues (K50, S89, and Q90) were selected and mutated back to the *same* residues (by applying our rotamer generation protocol) to generate 3^3 starting structures and use all of them in EVB calculations, to estimate the activation free energy of the reaction in HG3. These positions in the sequence space were chosen because the

https://dx.doi.org/10.1021/acscatal.0c01206
ACS Catal. 2020, 10, 6002−6012

mutated residues are situated very close to the active site and were used during the mutagenesis step (in experimental directed evolution) to optimize the protein. It is expected that wrongly predicted rotamer orientations of residues near the active site should severely affect the outcome of the simulations. Thus, the *self-mutation-based* calculations for three positions near the active site region are a reasonable choice to test our protocol. We have used eq 1 to calculate the activation free energy of HG3 (see Table 3), from the 3^3 EVB

Table 3. Experimental and Calculated Activation Free Energies of the Reaction in Kemp Eliminase[a]

HG3 variant	$\Delta G^{\ddagger}_{cal}$ (kcal/mol)	$\Delta G^{\ddagger}_{exp}$ (kcal/mol)	$\lvert\Delta G^{\ddagger}_{exp} - \Delta G^{\ddagger}_{cal}\rvert$
HG3	16.9	16.9	0.0
HG3.3b	14.6	16.0	1.4
HG3.7	14.4	14.1	0.3
HG3.14	14.2	13.9	0.3
HG3.17	14.1	13.6	0.5

[a]The last column represents the absolute difference between the experimental and calculated activation free energies.

calculations performed with the 3^3 starting structures. The structure corresponding to the minimum calculated activation free energy was selected to generate the structure of HG3.3b.

3.2.2. HG3.3b. In the case of HG3.3b, six residues should be mutated to convert HG3 to HG3.3b. V6I and M84C are the only 2 mutations that are present in all proteins from HG3.3b to HG3.17 (see Table 1). Since 6 residues are to be mutated, 3^6 (=729) starting structures and 729 EVB calculations have to be performed if we explore all 6 positions explicitly. On the other hand, if we limit the number of mutations to be explored, we can reduce the number of EVB calculations. As rationalized below, we mutated V6I and M84C using the Dunbrack backbone-dependent rotamer library (rotamers with the highest probability)[29] implemented in Chimera (version 1.10.2).[30] We then applied our rotamer generation protocol only for the remaining 4 mutations and generated 3^4 (81) starting structures (instead of 729 starting structures) and ran 81 EVB calculations. We specifically choose V6I and M84C to be mutated using the Dunbrack rotamer library because V6I and M84C are the only two mutations that are present in all proteins from HG3.3b to HG3.17. Thus, if any error occurs by not exploring (explicitly) these mutations, then that would affect all of the calculated results from HG3.3b to HG3.17 in a similar manner. Additionally, selecting residues which are common in all mutant variants would lead to a consistent comparison without doing a large number of EVB calculations. Like HG3, the catalytic effect is also predicted using eq 1, and the most probable structure (corresponding to $\Delta G^{\ddagger}_{min}$) is selected for use in the next round.

3.2.3. HG3.7. HG3.7 is different from HG3.3b by only 3 mutations (Table 1), and thus the movement between these 2 protein sequences is a reasonable test case. The structure corresponding to $\Delta G^{\ddagger}_{min}$ in the EVB calculation of reaction of HG3.3b was employed as a starting point for generating the 3^3 structures of HG3.7 (Figure 5). At this stage we also checked to what extent the order of introducing the mutations can influence our results. Starting from a variant of HG3.3b (without introducing M84C mutations), we followed 4 different paths to reach a variant of HG3.7.

Note that we are examining whether our choice of picking which mutation to perform first (in a set of three mutations)

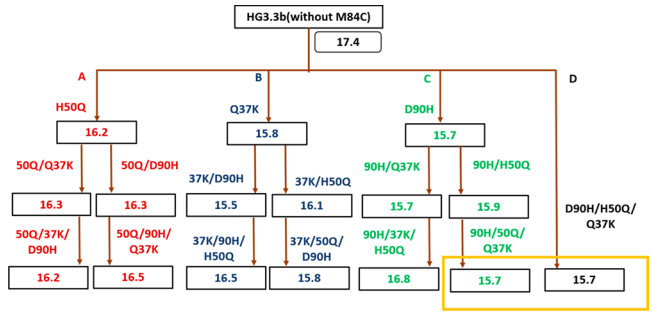


Figure 5. Illustrating the effect of the order of mutations in the process of generating starting structures on the calculation of activation free energies. Four different pathways are considered (A−D) to convert a variant of HG3.3b (without M84C) to a variant of HG3.7 (without M84C). The calculated activation free energies of the reaction in different mutated proteins are given in the boxes (in kcal/mol), and the mutations that lead to the corresponding proteins (starting from the root) are traced with arrows. The convergence of the calculations is stated in the yellow box (see main text). The calculated activation free energy of reaction in the variant of HG3.3b (without M84C) is 17.4 kcal/mol (in a black rounded rectangle near the root).

has any considerable effect on the calculated activation free energies.

Figure 5 helps to reveal that the order of mutation is not of critical importance. In pathways A−C, all mutations displayed are introduced one at a time, while the structure corresponding to minimum activation free energy is taken as the starting point for the next step of simulation. It is evident that the calculated values do not vary significantly. The maximum difference in the calculated activation free energies after all three mutations among routes A−C is ∼1 kcal/mol. Note that the error range for calculating activation free energies is still 1−2 kcal/mol. This is not our specific shortcoming, since the entire computational community would agree that 1−2 kcal/mol to be a legitimate error limit for calculating activation free energies. At any rate, this is just a very careful check of the stability of our calculations. The only thing that we want to point out is that the result coming from the most common protocol (i.e., pathway D) is not very different from those obtained in the other pathways where the mutations are attempted one at a time. The lowest activation free energy route in A−C (D90H → D90H/H50Q → D90H/H50Q/Q37K) converges decisively to the results for D, where all three mutations are introduced simultaneously. It is worth mentioning that the test for checking the ordering of mutations could be performed with any other set of mutations. The calculation starting with a variant of protein HG3.3b (without M84C) is just an example. As indicated in Figure 5, the calculated activation barrier for HG3.7 (without M84C) is 15.7 kcal/mol (pathway D). The activation free energy of HG3.7 (with M84C) is 14.4 kcal/mol (see Table 3). Thus, it can be suggested that the presence of CYS84 helps the other mutations to cooperatively enhance the activity of the protein. Furthermore, the results may explain why M84C is consistently carried over in all rounds of the experiment (possibly after round 1b; see Supplementary Table 4a of ref 27) without being mutated to anything else.

3.2.4. HG3.14. The move from HG3.7 to HG3.14 requires the introduction of eight point mutations. If we need to implement our method in its fullest form, then 3^8 (=6561) structures and EVB calculations are required. Thus, we need an

https://dx.doi.org/10.1021/acscatal.0c01206
ACS Catal. 2020, 10, 6002−6012

approximated strategy to tackle the problem by performing fewer computations. We can try to employ an approach similar to that used in the case of HG3.3b, where we reduced the number of computations by not exploring explicitly the conformational space of the V8I and M84C mutations. It is not always possible to find cases where some mutations are presented in all the mutant proteins. Thus, for cases where we cannot find common mutations, a more general approach is needed. An effective way of reducing the number of computations, without considerably compromising the reliability of the free energy calculation, is to predict a small number of reasonable starting structures that would mainly contribute to the calculation of activation free energy. In other words, we have to predict only the starting structures that correspond to $\Delta G^{\ddagger}_{min}$ in eq 1, and those which contribute in a major way to the activation free energy calculation using eq 1, without performing explicit EVB simulations for all 3^N of the starting structures. One option is to predict the reasonable starting structures (corresponds to $\Delta G^{\ddagger}_{min}$ in eq 1 and related structures) by considering the information from a limited number of mutational calculations and use those predicted reasonable structures to calculate the activation free energy of the reaction. The simplest option is to consider all possible single-mutation cases to predict reasonable starting structures with multiple mutations in them. Such calculations are adequate only when multimutational effects can be explained by additive effects of single mutations. However, in most cases the effect of one mutation might depend on the presence of other mutation(s). The functional effect of one mutation in the presence of other mutation(s) could be beneficial, neutral, or deleterious, and the combined effect might deviate from the individual additive effects. This is called epistasis.[31] In most evolutionary processes, the multimutational effects are nonadditive in nature. Thus, in our initial screening approach to explore the effect of multiple mutations, we should consider at least all possible double mutations. Theoretically, the double-mutation term can be described as a combination of single-mutation terms and terms related to the effect of performing a second mutation, in cases where the first mutation has already been introduced. Since the contribution due to single mutations is already included, it should be subtracted from the double-mutation terms to avoid an overcounting. Here we applied an *ad hoc* approach, expressing the effect of multiple mutations by using

$$\Delta\Delta G^{\ddagger}_{Q\to R} \approx \alpha \sum_{i=1}^{N} \Delta\Delta G^{\ddagger}_{Q\to R_i} + \beta \sum_{k>l}^{N} (\Delta\Delta G^{\ddagger}_{Q\to R_{kl}} - \Delta\Delta G^{\ddagger}_{Q\to R_k} - \Delta\Delta G^{\ddagger}_{Q\to R_l}) \quad (2)$$

Equation 2 is an approximate estimate of the relative barrier obtained by mutating N residues during the transition from protein Q to protein R. The index i runs over the residues that are mutated during the transition, whereas k and l run over all double-mutation combinations. α and β are parameters to scale the single- and double-mutation terms in eq 2, and their values vary as $0 \le \alpha \le 1$ and $0 < \beta < 1$. In the current implementation, $\alpha = 1.0$ and $\beta = 0.25$.

The single and double terms ($\Delta\Delta G^{\ddagger}_{Q\to R_x}$) in eq 2 can be defined as

$$\Delta\Delta G^{\ddagger}_{Q\to R_x} = \Delta G^{\ddagger}_{Q\to R_x} - \Delta G^{\ddagger}_{Q} \quad (3)$$

The terms $\Delta G^{\ddagger}_{Q\to R_x}$ denotes the activation free energy of the reaction in protein R_x, where x can stand for a single or double mutation of protein Q. $\Delta G^{\ddagger}_{Q}$ is the calculated activation free energy, obtained during the previous step (or taken for the state previously used to generate the single/double mutations). Both $\Delta G^{\ddagger}_{Q}$ and $\Delta G^{\ddagger}_{Q\to R_x}$ are calculated using eq 1 and correspond to averaged activation free energies. The term $\Delta G^{\ddagger}_{Q\to R_x}$ can be also represented as the activation free energy for each rotamer or rotamer combinations (for double mutations). This helps us to use eqs 2 and 3 to rank 3^N protein configurations (see below).

It is now important to clarify our approach by assuming that we use eq 2 to select ~N starting structures, which are used in the EVB calculations to calculate the activation free energy of the reaction using eq 1. To explain this example, let us take a case where ABCDEF are the mutations that we wish to perform on a protein. Since in our original protocol (see Methods) we consider 3 rotamers for each mutated residue, residue A can have rotamers A1 (rotamer 1), A2 (rotamer 2), and A3 (rotamer 3) and similarly residue B can have B1, B2, and B3 rotamers. Thus, A1_B1_C1_D2_E2_F2 symbolically defines a protein having rotamer 1 for residues A, B, and C and rotamer 2 for residues D, E, and F. Since we are considering 6 mutations, on the basis of our original protocol (see Methods) we should generate 3^6 starting structures. However, we can use eq 2 to rank these protein configurations, by assuming that $\Delta G^{\ddagger}_{Q\to R_x}$ in eq 3 corresponds to the activation free energy of the reaction for a specific configuration of R_x. In this case, $\Delta G^{\ddagger}_{Q\to R_x}$ in eq 3 is not calculated using eq 1, as it is done to estimate the effect of multimutations (see above). As a result, for single (3 starting structures) and double mutation (9 starting structures), the $\Delta G^{\ddagger}_{Q\to R_x}$ term would have 3 and 9 values, respectively. In this way we can calculate $\Delta\Delta G^{\ddagger}_{Q\to R_x}$ for each configuration (a total of 3^N $\Delta\Delta G^{\ddagger}_{Q\to R_x}$ values) and then rank the 3^6 structures accordingly to A1_B2_C1_D2_E3_F1, A2_B1_C3_D1_E1_F2, A2_B2_C2_D2_E2_F2, ... from the most likely to the least likely starting structures. The most probable structure should have a minimum $\Delta\Delta G^{\ddagger}_{Q\to R_x}$ value among all 3^6 structures. For example, the ranking might predict A2_B2_C2_D2_E2_F2 as the most probable, then A2_B3_C1_D2_E3_F1 as the next most probable, and so on. Then we can consider the ~6 (N = 6) most probable structures in our EVB calculations and use eq 1 to calculate the corresponding activation free energy of the reaction. We have used this approach in the cases of HG3.14 and HG3.17, and the calculated activation energy of the reaction is reported in Table 3.

These prediction calculations (use of eqs 2 and 3) are only used to rank the most probable structures, which are then used for the direct calculations (when we are really performing EVB calculations with structures having 8 mutations of HG3.7). In the current implementation, for the HG3.7 to HG3.14 conversion, the best 6−10 structures (based on ranking) are designated as starting structures for the EVB calculations (direct calculations). The calculated activation free energy for HG3.14 reported in Table 3 was obtained after using eq 1 with the EVB results of the direct calculations. The structure corresponding to the minimum activation free energy in the direct calculation can be used for the next round of predictions.

https://dx.doi.org/10.1021/acscatal.0c01206
ACS Catal. 2020, 10, 6002−6012

As explained above, eq 2 can also be used to explore the effect of multimutations using the calculations of single and double mutations. For example, Table 4 represents the

Table 4. Calculated $\Delta\Delta G^{\ddagger}_{HG3.7\rightarrow HG3.14}$ for All of the Double and Single Mutations of H3.7, Used to Estimate the Overall Barrier Change in the HG3.7 to H3.14 Conversion

HG3.7 → HG3.14	$\Delta\Delta G^{\ddagger}_{Q\rightarrow R_x}$ (kcal/mol)
R89S	0.14
T208M	0.40
N125T	−0.36
G82A	0.97
T279S	−0.68
T142N	−0.63
D300N	0.38
T105I	0.37
R89S/T142N	−0.64
N125T/D300N	−0.40
N125T/T279S	−0.22
T105I/T279S	−0.03
R89S/T208M	−0.17
T142I/D300N	−0.38
T105I/N125T	−0.17
R89S/T279S	0.19
G82A/D300N	−0.43
G82A/T279S	0.12
R89S/T105I	−0.11
N125T/T142N	−0.59
T105I/D300N	−0.33
G82A/R89S.	−0.2
R89S/N125T	−0.29
G82A/T105I	−0.17
G82A/N125T	−0.43
N125T/T208M	−0.24
T142N/T208M	−0.2
R89S/D300N	−0.27
T208M/T279S	0.0
G82A/T142N	−0.97
G82A/T208M	−0.29
T208M/D300N	−0.02
T279S/D300N	−0.07
T105I/T208M	−0.11
T105I/T142N	−0.12
T142N/T279S	−0.07
$\Delta G^{\ddagger}_{Q}$	14.4
$\Delta\Delta G^{\ddagger}_{HG3.7\rightarrow HG3.14}$ (α = 1.0; β = 0.25)	**−2.10**

Boltzmann averaged contributions of the single and double mutations to the calculated $\Delta\Delta G^{\ddagger}_{Q\rightarrow R}$ for the conversion of protein HG3.7 to HG3.14. Thus, we have used eq 3 to calculate each row in Table 4, where each row represents the relative change in activation free energy of the Kemp elimination reaction (due to a single/double mutation of the protein HG3.7). The first row in Table 4 represents the relative change in activation free energy due to the mutation from R (ARG) to S (SER) at the 89th position of the protein HG3.7. In this case, $\Delta\Delta G^{\ddagger}_{HG3.7\rightarrow HG3.7_{R89S}}$ was calculated by subtracting $\Delta G^{\ddagger}_{HG3.7}$ (14.4 kcal/mol) from $\Delta G^{\ddagger}_{HG3.7\rightarrow HG3.7_{R89S}}$, which was calculated using eq 1 (the summation in eq 1 was done over three starting structures). Please note the different way of calculating the $\Delta G^{\ddagger}_{Q\rightarrow R_x}$ term in eq 3 in the cases (a) when 3^N starting structures are ranked to find most probable structures (previous paragraph) and (b) when the effect of multiple mutation is predicted from a limited mutation based simulation (current paragraph).

While the predicted $\Delta\Delta G^{\ddagger}_{HG3.7\rightarrow HG3.14}$ in Table 4 seems to be an overestimation, the error is relatively small depending on the approximation we made, as the double-mutation-based simulations are used to account for the effect of eight mutations. Ideally, if the values of α and β can be predicted using some optimization protocols, then the predicted $\Delta\Delta G^{\ddagger}_{HG3.7\rightarrow HG3.14}$ may have a better correlation with the experimental results, which is −0.2 kcal/mol. In our current calculation, the values of α and β were optimized after several trial values. The values α = 1.0 and β = 0.25 were used in all cases (wherever eq 2 was used in this work). It is worth noting that Table 4 is a demonstration of how different terms in eq 2 contribute to the prediction calculation (a limited mutation based calculation to explain the effect of multiple mutations). By looking at the results of double mutations in Table 4, we can state that the mutations of the distant residues (from the active sites), for example D300N, T279S, T208M, T142N, and 105I, make relatively small contributions to the overall stabilization. The results of the present study are consistent with the finding of Chica and co-workers,[32] where they found that the progress in catalysis upon moving to HG3.17 does not requires all of the 17 observed mutations.

3.2.5. HG3.17. To transform the structure of H3.14 into HG3.17 requires six more mutations. The set of rotamers used to generate the most likely configurations of the mutations are selected by performing single- and double-mutation-based calculations and using eqs 2 and 3 as explained above. The most probable protein structures are then used in the "direct" calculations (EVB calculation) to obtain the activation free energy of the reaction. In this "direct" calculation only 6 instead of 3^6 structures were used. Thus, the use of the aforementioned protocol has significantly lowered the computational cost calculating the activation free energy of a reaction. The above approach of using $\sim N$ starting structures instead of 3^N structures is found to be useful for the cases of both HG3.14 and HG3.17 calculations.

Equations 2 and 3 are used to rank protein configurations as well as to explain the extent of catalysis due to the introduction of certain mutations in a protein. The accuracy of the latter depends on the accuracy of the calculation of $\Delta G^{\ddagger}_{Q}$ term in eq 3. Even a small error in the calculation of $\Delta G^{\ddagger}_{Q}$ can significantly compromise the result, because that error will propagate through all single- and double-mutation terms in eq 2. Thus eqs 2 and 3 are both required only to support the process of ranking the starting structures. Hopefully, the errors in the calculation of $\Delta G^{\ddagger}_{Q}$ are almost the same for all of the structures considered in a single ranking process. In such a case these errors should not alter the ordering of the most probable structures.

Overall, our designed protocol seems to be useful for calculation of activation free energies, even if the structure of the protein is unknown in advance. If we can add a predictive tool for proposing the next probable mutations, our protocol can be applied as a promising tool for *in silico* enzyme design. In this regard, eq 2 appears to offer substantial help in predicting the next possible set of mutations. To examine the adequacy of this assumption, we implemented our approach in

https://dx.doi.org/10.1021/acscatal.0c01206
ACS Catal. 2020, 10, 6002−6012

Table 5. Mutation Combinations Used to Predict the Next Possible Mutations of HG3.3b

type	small	nucleophilic	hydrophobic	acidic	basic	aromatic	
Q37	Q37A	Q37S	Q37I	Q37D	Q37K	Q37F	
H50	H50A	H50S		H50D	H50K	H50F	H50Q
D90	D90A	D90S	D90I		D90K or D90H	D90F	D90Q

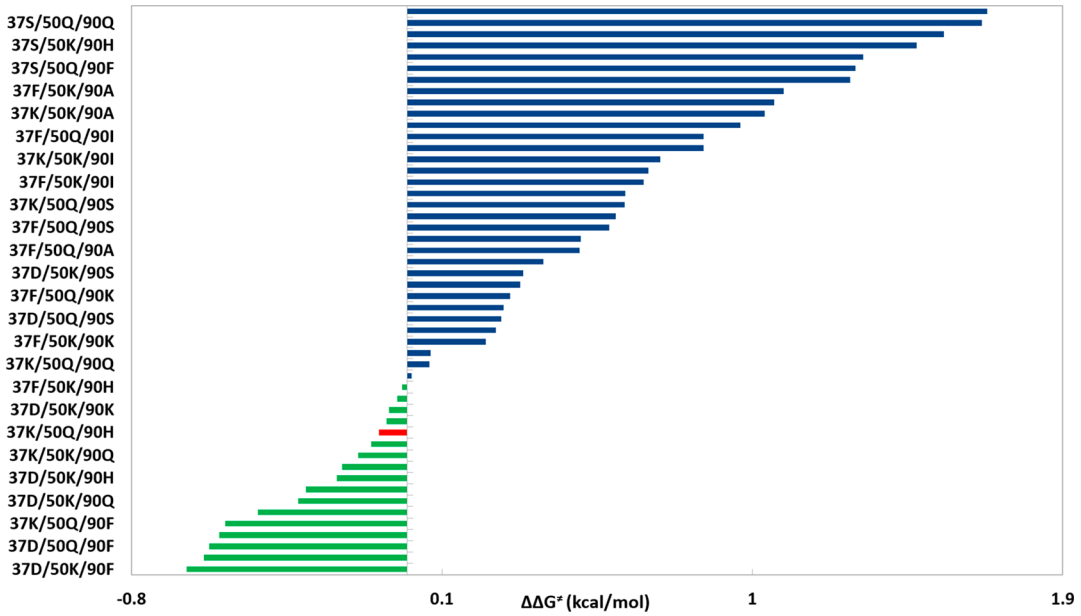


Figure 6. Ranking the predictions of three mutations of HG3.3b. The prediction for H50Q/Q37K/D90H is colored in red.

the mutations on HG3.3b, in an attempt to generate the next possible mutated protein, considering changes only at the positions of H50, Q37, and D90. For each of the selected positions various amino acid types were examined as substitutes (see Table 5).

If the assumption that eq 2 can be used as a predictive tool for selecting the next set of mutations works correctly, the result of the calculation should predict H50Q, Q37K, and D90H as the next probable mutations (the experimentally found sequence) out of all possible mutation options that are presented in Table 5. Our results ranked the combination H50Q/Q37K/D90H as 13th out of 252 combinations of tested triple mutations. The first 50 predictions are shown in Figure 6. Interestingly, the $\Delta\Delta G^{\ddagger}_{Q\rightarrow R}$ value predicted by eq 2 for the mutations H50Q/Q37K/D90H of HG3.3b is ∼−0.1 kcal/mol (see Figure 6 (red bar)), which is close to the calculated $\Delta\Delta G^{\ddagger}$ value in Table 3. This also shows that eq 2 may be used to estimate the effect of multiple mutations instead of doing so by performing all explicit calculations (EVB of 3^N structures), when the number of the mutations to be performed is small. This observation also proves the consistency in our calculations. Even if the ideal prediction is expected to be in favor of H50Q/Q37K/D90H as the most *catalytic* mutation, the relevant predictions are not that far from it. For example, our calculation proposed F to be the favorable mutation for D90 along with K/D and Q/K for the 37th and 50th positions of the sequence of the protein. The residue F is also found at the 90th position of the sequence of HG3.17, which shows that mutation at position 90 of the sequence by F should not destroy the catalytic gain. The decrease in activation free energy between HG3.7 (H in the 90th position of the sequence) and HG3.17 is ∼0.5 kcal/mol (Table 1), which shows that both histidine (H) and phenylalanine (F) are favorable at that position. Thus, for mutation at the 90th position in HG3.3b we can consider both possibilities of H and F. That makes our prediction reasonable.

An interesting insight can be obtained by placing different mutations, according to their relative activation free energies, on the surface of a hypothetical funnel, as shown in Figure 7. This type of diagram places the predicted mutations with a lower barrier at the bottom of the funnel. This gives hope that experimental design of the mutations at the lowest region of the funnel will allow a rational improvement of the given enzyme.

4. CONCLUDING REMARKS

The use of rotamers conformational space exploration and reaction energetics calculations have been shown to be useful for *in silico* directed evolution. However, the implementation of such a strategy is not obvious. The determination of a starting protein configuration (for any free energy based calculations), where the structure of the protein is unknown, is one of most important components, as it is also in *in silico* enzyme design. Here we developed such a strategy, where we generated three distinct rotamers for each mutated residue while considering the contributions to the average activation barriers from multiple reasonable structures. This approach of calculating the reaction barriers has been successfully validated for two model systems. We also introduced an equation that allows us to reduce the number of starting structures for our EVB calculations, whenever the number of possible mutations is very large. This approach approximates the catalytic/anticatalytic effect of N mutations, by considering single and double mutations, and is used to reduce the number of

https://dx.doi.org/10.1021/acscatal.0c01206
ACS Catal. 2020, 10, 6002−6012

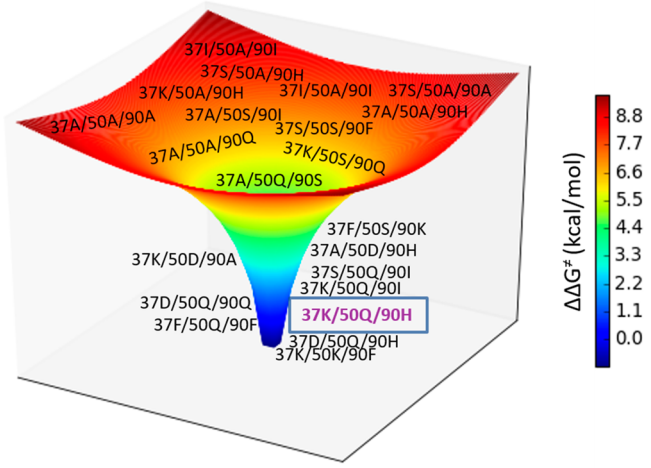


Figure 7. Comparison of the number of possible three-mutation combinations with the stabilization of the transition state. The color scaling indicates the extent of stabilization (red, least stable; dark blue, most stable) achieved by introducing mutations at three positions (Q37, H50, D90), using different types of amino acids. The flat top part of the 3D diagram corresponds to a large number of mutation combinations with a high $\Delta\Delta G^{\ddagger}$ value, whereas the narrow bottom indicates very low number of mutations leading to a low $\Delta\Delta G^{\ddagger}$ value. The experimentally found mutation combination (Q37K/H50Q/D90H) is highlighted (blue box).

probable starting structures for computing the activation free energies of N proteins instead of 3^N proteins.

At this point it might be useful to reiterate the issues of consistent computer-aided enzyme design. As we have argued repeatedly, consistent approaches must be validated by their ability to reproduce known results and subjected to scrutiny similar to that applied here. In this respect we wish to comment on a very recent interesting work of Hilvert and co-workers,[33] which has shown that Q50 can provide a significant catalysis in HG3.17 (most probably by forming an oxyanion hole), while other mutants at the 50th position might (M50, F50) or might not (K50, H50) help in catalysis. Although this finding seems to contradict the oversimplified views of the nature of oxyanion stabilization, it can be rationalized by the following facts. First, as has been emphasized in many of our works, the expected effect of mutations should not be comprehended by just looking at the structure (assuming it to reflect all the changes in interaction); it is essential to determine the effect of a mutation by careful free energy calculations that consider the overall reorganization and water penetration before and after the mutation. Second and more specifically, with regard to the catalytic contributions of oxyanions, it is useful to note that the first quantitative analysis of the energetics of oxyanions in serine proteases (e.g., refs 34 and 35) already indicated that the catalytic contribution of the oxyanion is not due to the interaction between the dipoles of the environment and the transition-state charges (which are similar in the protein and in water) but to the limited reorganization of the dipoles in the protein site. Major confusion in this respect has been clarified recently in ref 36, and it is also useful to clarify that the oxyanion contribution cannot be evaluated by just estimating the interaction energies using quantum mechanical calculations (that may give a major overestimation) but by considering the effect of the environmental relaxation in the protein and in the water reference system (this issue has been illustrated in ref 37). The nontrivial evaluation of oxyanion contributions has also been demonstrated in studies of ketosteroid isomerase (KSI),[38,39] where we established that there is a major catalytic contribution to the oxyanion hole, while demonstrating the problems with the analysis of Herschlag and co-workers,[40] who tried to argue that the oxyanion hole *does not* help to catalysis. At any rate, we agree that the study of ref 33 should provide an excellent benchmark for methods that are aimed at quantitative enzyme design.

Overall, our attempt to explore and extend *in silico* directed evolution has some encouraging aspects and may be improved with the help of a more robust functional form of eq 2. That kind of advancement should involve more accurate prediction of the individual activation barriers as well. Nevertheless, the current version of our approach can serve as a powerful tool in those experimental studies that require predicting which of the residues should be mutated next.

Our plans for future work include an extended exploration of the rotamer conformational space, by implementing a Monte Carlo based rotamer generator and optimizer based on a rotamer library, as a supplement to our simplified folding free energy calculation method. In addition, a further evaluation of the weights in eq 2, supported by adequate optimization methods, such as machine learning, may aid in increasing the accuracy of the results.

■ ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acscatal.0c01206.

Details of the computer simulations, a succinct description of the EVB method, and the EVB parameters that were used in the simulations (PDF)

■ AUTHOR INFORMATION

Corresponding Author

Arieh Warshel − *Department of Chemistry, University of Southern California, Los Angeles, California 90089, United States;* orcid.org/0000-0001-7971-5401; Email: warshel@usc.edu

Authors

Dibyendu Mondal − *Department of Chemistry, University of Southern California, Los Angeles, California 90089, United States;* orcid.org/0000-0002-5047-6985

Vesselin Kolev − *Department of Chemistry, University of Southern California, Los Angeles, California 90089, United States*

Complete contact information is available at: https://pubs.acs.org/10.1021/acscatal.0c01206

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

This research study was supported by the National Institutes of Health (R35 GM122472) and the National Science Foundation (Grant MCB 1707167). The authors wish to express their gratitude to the University of Southern California High Performance Computing and Communication Center and the Extreme Science and Engineering Discovery Environment's (XSEDE) Comet facility at the San Diego Super-

computing Center, for the provided computational resources. D.M. thanks Dr. Garima Jindal and Dr. Zhen Tao Chu for their guidance and helpful discussion.

REFERENCES

(1) Marcos, E.; Chidyausiku, T. M.; McShan, A. C.; Evangelidis, T.; Nerli, S.; Carter, L.; Nivon, L. G.; Davis, A.; Oberdorfer, G.; Tripsianes, K.; Sgourakis, N. G.; Baker, D. De novo design of a non-local beta-sheet protein with high stability and accuracy. *Nat. Struct. Mol. Biol.* **2018**, *25* (11), 1028−1034.
(2) Wu, Z.; Kan, S. B. J.; Lewis, R. D.; Wittmann, B. J.; Arnold, F. H. Machine learning-assisted directed protein evolution with combinatorial libraries. *Proc. Natl. Acad. Sci. U. S. A.* **2019**, *116* (18), 8852−8858.
(3) Huang, P. S.; Boyken, S. E.; Baker, D. The coming of age of de novo protein design. *Nature* **2016**, *537* (7620), 320−327.
(4) Kiss, G.; Celebi-Olcum, N.; Moretti, R.; Baker, D.; Houk, K. N. Computational Enzyme Design. *Angew. Chem., Int. Ed.* **2013**, *52* (22), 5700−5725.
(5) Khersonsky, O.; Rothlisberger, D.; Dym, O.; Albeck, S.; Jackson, C. J.; Baker, D.; Tawfik, D. S. Evolutionary Optimization of Computationally Designed Enzymes: Kemp Eliminases of the KE07 Series. *J. Mol. Biol.* **2010**, *396* (4), 1025−1042.
(6) Jindal, G.; Slanska, K.; Kolev, V.; Damborsky, J.; Prokop, Z.; Warshel, A. Exploring the challenges of computational enzyme design by rebuilding the active site of a dehalogenase. *Proc. Natl. Acad. Sci. U. S. A.* **2019**, *116* (2), 389−394.
(7) Chica, R. A.; Doucet, N.; Pelletier, J. N. Semi-rational approaches to engineering enzyme activity: combining the benefits of directed evolution and rational design. *Curr. Opin. Biotechnol.* **2005**, *16* (4), 378−384.
(8) Arnold, F. H. Directed Evolution: Bringing New Chemistry to Life. *Angew. Chem., Int. Ed.* **2018**, *57* (16), 4143−4148.
(9) Pavelka, A.; Chovancova, E.; Damborsky, J. HotSpot Wizard: a web server for identification of hot spots in protein engineering. *Nucleic Acids Res.* **2009**, *37*, W376−W383.
(10) Dalkiran, A.; Rifaioglu, A. S.; Martin, M. J.; Cetin-Atalay, R.; Atalay, V.; Dogan, T. ECPred: a tool for the prediction of the enzymatic functions of protein sequences based on the EC nomenclature. *BMC Bioinf.* **2018**, *19* (1), 1−13.
(11) Chen, C. Y.; Georgiev, I.; Anderson, A. C.; Donald, B. R. Computational structure-based redesign of enzyme activity. *Proc. Natl. Acad. Sci. U. S. A.* **2009**, *106* (10), 3764−3769.
(12) Fox, R. J.; Davis, S. C.; Mundorff, E. C.; Newman, L. M.; Gavrilovic, V.; Ma, S. K.; Chung, L. M.; Ching, C.; Tam, S.; Muley, S.; Grate, J.; Gruber, J.; Whitman, J. C.; Sheldon, R. A.; Huisman, G. W. Improving catalytic function by ProSAR-driven enzyme evolution. *Nat. Biotechnol.* **2007**, *25* (3), 338−344.
(13) Funke, S. A.; Otte, N.; Eggert, T.; Bocola, M.; Jaeger, K. E.; Thiel, W. Combination of computational prescreening and experimental library construction can accelerate enzyme optimization by directed evolution. *Protein Eng., Des. Sel.* **2005**, *18* (11), 509−514.
(14) van der Kamp, M. W.; Mulholland, A. J. Combined Quantum Mechanics/Molecular Mechanics (QM/MM) Methods in Computational Enzymology. *Biochemistry* **2013**, *52* (16), 2708−2728.
(15) Moliner, V.; Himo, F. Editorial: Challenges in Computational Enzymology. *Front. Chem.* **2019**, *7*, 1.
(16) Kamerlin, S. C. L.; Warshel, A. The empirical valence bond model: theory and applications. *Wiley Interdiscip. Rev.: Comput. Mol. Sci.* **2011**, *1* (1), 30−45.
(17) Jindal, G.; Ramachandran, B.; Bora, R. P.; Warshel, A. Exploring the development of ground-state destabilization and transition-state stabilization in two directed evolution paths of kemp eliminases. *ACS Catal.* **2017**, *7* (5), 3301−3305.
(18) Vicatos, S.; Rychkova, A.; Mukherjee, S.; Warshel, A. An effective Coarse-grained model for biological simulations: Recent refinements and validations. *Proteins: Struct., Funct., Genet.* **2014**, *82* (7), 1168−1185.
(19) Janssen, D. B. Evolving haloalkane dehalogenases. *Curr. Opin. Chem. Biol.* **2004**, *8* (2), 150−159.
(20) Verschueren, K. H. G.; Seljee, F.; Rozeboom, H. J.; Kalk, K. H.; Dijkstra, B. W. Crystallographic Analysis of the Catalytic Mechanism of Haloalkane Dehalogenase. *Nature* **1993**, *363* (6431), 693−698.
(21) Schanstra, J. P.; Ridder, A.; Kingma, J.; Janssen, D. B. Influence of mutations of Val226 on the catalytic rate of haloalkane dehalogenase. *Protein Eng., Des. Sel.* **1997**, *10* (1), 53−61.
(22) Schanstra, J. P.; Janssen, D. B. Kinetics of halide release of haloalkane dehalogenase: Evidence for a slow conformational change. *Biochemistry* **1996**, *35* (18), 5624−5632.
(23) Schanstra, J. P.; Ridder, I. S.; Heimeriks, G. J.; Rink, R.; Poelarends, G. J.; Kalk, K. H.; Dijkstra, B. W.; Janssen, D. B. Kinetic characterization and X-ray structure of a mutant of haloalkane dehalogenase with higher catalytic activity and modified substrate range. *Biochemistry* **1996**, *35* (40), 13186−13195.
(24) Kennes, C.; Pries, F.; Krooshof, G. H.; Bokma, E.; Kingma, J.; Janssen, D. B. Replacement of Tryptophan Residues in Haloalkane Dehalogenase Reduces Halide Binding and Catalytic Activity. *Eur. J. Biochem.* **1995**, *228* (2), 403−407.
(25) Krooshof, G. H.; Ridder, I. S.; Tepper, A. W. J. W.; Vos, G. J.; Rozeboom, H. J.; Kalk, K. H.; Dijkstra, B. W.; Janssen, D. B. Kinetic analysis and X-ray structure of haloalkane dehalogenase with a modified halide-binding site. *Biochemistry* **1998**, *37* (43), 15013−15023.
(26) Röthlisberger, D.; Khersonsky, O.; Wollacott, A. M.; Jiang, L.; DeChancie, J.; Betker, J.; Gallaher, J. L.; Althoff, E. A.; Zanghellini, A.; Dym, O. Kemp elimination catalysts by computational enzyme design. *Nature* **2008**, *453* (7192), 190−195.
(27) Blomberg, R.; Kries, H.; Pinkas, D. M.; Mittl, P. R.; Grütter, M. G.; Privett, H. K.; Mayo, S. L.; Hilvert, D. Precision is essential for efficient catalysis in an evolved Kemp eliminase. *Nature* **2013**, *503* (7476), 418−421.
(28) Privett, H. K.; Kiss, G.; Lee, T. M.; Blomberg, R.; Chica, R. A.; Thomas, L. M.; Hilvert, D.; Houk, K. N.; Mayo, S. L. Iterative approach to computational enzyme design. *Proc. Natl. Acad. Sci. U. S. A.* **2012**, *109* (10), 3790−3795.
(29) Dunbrack, R. L., Jr Rotamer libraries in the 21st century. *Curr. Opin. Struct. Biol.* **2002**, *12* (4), 431−440.
(30) Pettersen, E. F.; Goddard, T. D.; Huang, C. C.; Couch, G. S.; Greenblatt, D. M.; Meng, E. C.; Ferrin, T. E. UCSF Chimera—a visualization system for exploratory research and analysis. *J. Comput. Chem.* **2004**, *25* (13), 1605−1612.
(31) Starr, T. N.; Thornton, J. W. Epistasis in protein evolution. *Protein Sci.* **2016**, *25* (7), 1204−1218.
(32) Broom, A.; Rakotoharisoa, R. V.; Thompson, M. C.; Zarifi, N.; Nguyen, E.; Mukhametzhanov, N.; Liu, L.; Fraser, J. S.; Chica, R. A., Evolution of an enzyme conformational ensemble guides design of an efficient biocatalyst. *bioRxiv* **2020**, 2020.03.19.999235 (accessed April 16, 2020).
(33) Kries, H.; Bloch, J.; Bunzel, H. A.; Pinkas, D. M.; Hilvert, D. Contribution of oxyanion stabilization to Kemp eliminase efficiency. *ACS Catal.* **2020**, *10*, 4460.
(34) Hwang, J. K.; Warshel, A. Semiquantitative Calculations of Catalytic Free-Energies in Genetically Modified Enzymes. *Biochemistry* **1987**, *26* (10), 2669−2673.
(35) Hwang, J. K.; King, G.; Creighton, S.; Warshel, A. Simulation of Free-Energy Relationships and Dynamics of Sn2 Reactions in Aqueous-Solution. *J. Am. Chem. Soc.* **1988**, *110* (16), 5297−5311.
(36) Jindal, G.; Warshel, A. Misunderstanding the preorganization concept can lead to confusions about the origin of enzyme catalysis. *Proteins: Struct., Funct., Genet.* **2017**, *85* (12), 2157−2161.
(37) Kamerlin, S. C. L.; Chu, Z. T.; Warshel, A. On Catalytic Preorganization in Oxyanion Holes: Highlighting the Problems with the Gas-Phase Modeling of Oxyanion Holes and Illustrating the Need for Complete Enzyme Models. *J. Org. Chem.* **2010**, *75* (19), 6391−6401.
(38) Warshel, A.; Sharma, P. K.; Chu, Z. T.; Aqvist, J. Electrostatic contributions to binding of transition state analogues can be very

https://dx.doi.org/10.1021/acscatal.0c01206
ACS Catal. 2020, 10, 6002−6012

different from the corresponding contributions to catalysis: Phenolates binding to the oxyanion hole of ketosteroid isomerase. *Biochemistry* **2007**, *46* (6), 1466−1476.
(39) Kamerlin, S. C. L.; Sharma, P. K.; Chu, Z. T.; Warshel, A. Ketosteroid isomerase provides further support for the idea that enzymes work by electrostatic preorganization. *Proc. Natl. Acad. Sci. U. S. A.* **2010**, *107* (9), 4075−4080.
(40) Kraut, D. A.; Sigala, P. A.; Pybus, B.; Liu, C. W.; Ringe, D.; Petsko, G. A.; Herschlag, D. Testing electrostatic complementarity in enzyme catalysis: Hydrogen bonding in the ketosteroid isomerase oxyanion hole. *PLoS Biol.* **2006**, *4* (4), e99.

https://dx.doi.org/10.1021/acscatal.0c01206
ACS Catal. 2020, 10, 6002−6012