J Comput Aided Mol Des (2012) 26:1079–1095
DOI 10.1007/s10822-012-9601-y

Computational design of a Diels–Alderase from a thermophilic esterase: the importance of dynamics

Mats Linder · Adam Johannes Johansson · Tjelvar S. G. Olsson · John Liebeschuetz · Tore Brinck

Received: 23 April 2012 / Accepted: 3 September 2012 / Published online: 16 September 2012

Abstract A novel computational Diels–Alderase design, based on a relatively rare form of carboxylesterase from *Geobacillus stearothermophilus*, is presented and theoretically evaluated. The structure was found by mining the PDB for a suitable oxyanion hole-containing structure, followed by a combinatorial approach to find suitable substrates and rational mutations. Four lead designs were selected and thoroughly modeled to obtain realistic estimates of substrate binding and prearrangement. Molecular dynamics simulations and DFT calculations were used to optimize and estimate binding affinity and activation energies. A large quantum chemical model was used to capture the salient interactions in the crucial transition state (TS). Our quantitative estimation of kinetic parameters was validated against four experimentally characterized Diels–Alderases with good results. The final designs in this work are predicted to have rate enhancements of $\approx 10^3$–10^6 and high predicted proficiencies. This work emphasizes the importance of considering protein dynamics in the design approach, and provides a quantitative estimate of the how the TS stabilization observed in most de novo and redesigned enzymes is decreased compared to a minimal, 'ideal' model. The presented design is highly interesting for further optimization and applications since it is based on a thermophilic enzyme (T_{opt} = 70 °C).

Keywords Diels–Alder · Computational enzyme design · DFT · Molecular dynamics

Introduction

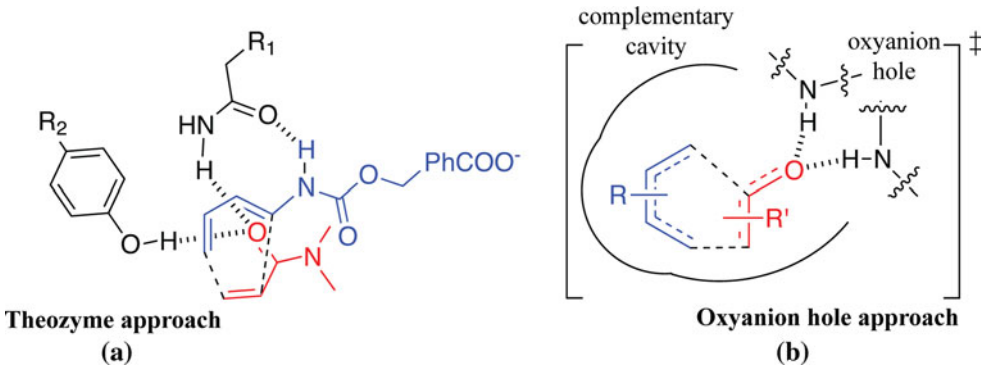
Designing an enzyme in a de novo sense, or to redesign an existing one to perform a mechanistically different task, is a daunting yet attractive problem in modern chemistry [1]. Enzymes' unmatched proficiency and selectivity are requisites for life itself, and the ability to control their function with the same delicate precision as nature would represent a paradigm shift in industrial catalysis. Still, we do not yet possess any rigorous tool to derive either function from structure, or structure from function. Add the fact that enzymes operate over timescales ranging from femtosecond to millisecond, and the challenges associated with predictive design become clear.

Nevertheless, many successful attempts at enzyme design have been reported in recent years. Design strategies include rational mutagenesis [2–6] and more recently computational approaches [1, 7, 8]. One particularly intriguing design problem that has captured the attention of several groups is the Diels–Alder reaction [9], both because of its paramount importance in synthetic chemistry and its scarcity in nature [10]. Of the known reactions catalyzed by putative Diels–Alderases in nature [10–12], the only intermolecular one has been shown not to follow a strict Diels–Alder mechanism [13, 14]. But the mechanism has been implemented in artificial biocatalysis. During the 1990s Hilvert, Janda, Lerner and others designed catalytic antibodies for the Diels–Alder reaction [15–19] by

Electronic supplementary material The online version of this article (doi:10.1007/s10822-012-9601-y) contains supplementary material, which is available to authorized users.

M. Linder · A. J. Johansson · T. Brinck (✉)
Applied Physical Chemistry, KTH Royal Institute of Technology, Teknikringen 30, 100 44 Stockholm, Sweden
e-mail: tore@physchem.kth.se

T. S. G. Olsson · J. Liebeschuetz
Cambridge Crystallographic Data Centre, 12 Union Road, Cambridge CB2 1EZ, UK
e-mail: john@ccdc.cam.ac.uk



Scheme 1 **a** The theozyme design leading to the experimentally validated Diels–Alderase 20_00_00 [29]. **b** Schematic diagram of the catalytic action of an oxyanion hole in a generic enzyme design

constructing haptens to resemble the putative transition state (TS) [20, 21].

Modern computational resources, along with increasing structural knowledge, has allowed for an increasing portion of the design process to take place in silico [6, 22]. The group of David Baker has developed a series of tools (Rosetta) that can be used to impose de novo functionalities on a given tertiary structure [23, 24]. Using Rosetta in concert with the 'inside-out' design strategy of Houk and coworkers, where optimal catalytic functionalities are predicted by designing 'theozymes' [25, 26], the group has successfully reported a series of de novo enzymes [27–29]. In the last study, they designed enzymes catalyze the intermolecular Diels–Alder reaction [22]. This was very recently followed up by an intriguing example of crowd-sourcing, when FoldIt [30] players were employed to help improve the DA_20_10 enzyme by a factor $\sim$20 [31].

This de novo Diels–Alderase is a major step toward efficient, predictive enzyme design. However, the inside-out design (illustrated in Scheme 1a) and subsequent optimization of a protein scaffold to incorporate the theozymes are static operations, and several studies have raised awareness of the protocol's limitations [32–36]. For example, molecular dynamics (MD) studies of the Retro–Aldolase design [27] reveal that the designed active site conformation is poorly represented in the resulting ensemble, accounting for the lower than predicted catalytic efficiency [32]. When studying the DA_00_00 design from Siegel et al. by MD methods, we observed that the designed enzyme-substrate interactions were poorly sampled in the trajectories, and hence that the substrates bound non-specifically to the active site [37].

These examples illustrate an increasing consensus that the dynamic behavior of enzyme-substrate complexes is important for improved success of computational design. Indeed, studies have begun to emerge that in various ways incorporate dynamics and-or flexibility in their design protocols with improved results [38, 39]. Apart from new advanced design tools [40], MD has come to play a vital part in both prediction and validation of new enzymes. As mentioned, recent investigations have shown that reasons for both active and inactive designs can be understood by MD evaluation [32, 34, 39, 41].

On top of these new directions in computer-aided design, there is a growing consensus that a combined use of several tools is instrumental for successful designs, where (semi-) rational design is one [22, 42]. In this spirit, we report a novel design originating from a combinatorial approach for finding suitable substrate motifs for a given enzyme candidate *combined* with rational mutations to better accommodate the substrates and introduce novel functionalities.

Our aim is to develop a design protocol without having to centre around two a priori selected substrates, which would limit the chemical search space [37]. We have concentrated our exploration of structures suitable for Diels–Alderase design to enzymes containing a so-called 'oxyanion hole', amongst which we have reported a redesigned lipase [43] and a hydroxynitrile lyase [37]. The idea is that like most organic Diels–Alder catalysts reported in the literature [44, 45], oxyanion holes can stablilize a build-up of negative charge on the dienophile in the TS of a normal electron demand Diels–Alder reaction, as illustrated in Scheme 1b. In this report, we expand our docking-based mutant evaluation [37] to involve MD-based screening of mutant candidates, and report the detailed computational design of one of the top candidates: a thermophilic enzyme with very high predicted catalytic power and diastereoselectivity.

Theory and methods

The workflow in the design protocol presented here can be said to consist of three stages, details of which have been described in a previous report [37]. A flowchart representation of the protocol is provided in Fig. S13. Stage A consists of PDB mining and virtual screening of a substrate

library in selected enzyme structures, followed by rational in silico mutagenesis and additional modifications of the substrates. Stage B involves MD simulations to validate and refine, if necessary, the preliminary design from stage A. For the resulting lead design(s), longer simulations are then undertaken and binding energies are calculated. In stage C, quantum chemical (QC) calculations are performed on a model active site to find the Diels–Alder TS and compute the catalytic activation energy.

In our previous work, the lead enzyme design was determined exclusively from a series of docking runs in stage A, while MD was used in stage B to refine and validate the selection of substrates [37]. In this work, we wanted to explore the potential of MD as a tool for refining the enzyme design as well, leading to a more extensive dynamic comparison of a number of mutants [38]. One advantage over a more automated screening protocol is that one is able to extract useful information even from the systems that prove unproductive [39].

Reaction thermodynamics

For an intermolecular enzymatic Diels–Alder reaction, assuming a random-ordered mechanism, the relative catalytic rate can be expressed as in Eq. 1 [19, 46–48]

$$\frac{v_{cat}}{v_{uncat}} = \frac{k_{cat}/k_{uncat}[\mathrm{E}^0]}{K_{M1}K_{M2} + K_{M1}[\mathrm{S}_2] + K_{M2}[\mathrm{S}_1] + [\mathrm{S}_1][\mathrm{S}_2]}. \quad (1)$$

The Michaelis constants K_{M1} and K_{M2} are generally measured for each substrate under pseudo-first-order conditions. The substrate concentrations are represented by $[S_1]$ and $[S_2]$, respectively. $[E^0]$ is the total enzyme concentration and k_{cat} and k_{uncat} are the rate constants obtained from the Eyring equation [49],

$$k = \kappa \frac{k_B T}{h} e^{-\Delta G^{\ddagger}/RT} \quad (2)$$

where $k_B T$ is the Boltzmann factor, h is the Planck constant and κ is the transmission coefficient, usually taken to be unity [50]. Hence, we evaluate a designed system by estimating the parameters k_{cat}, k_{uncat}, K_{M1} and K_{M2}. We note from Eq. 1 that a large relative rate is promoted by a high enzyme concentration and low substrate concentrations. There are obvious practical limitations, however, as enzyme concentrations must be held low e.g. to prevent aggregation, and since the absolute rate decreases with smaller substrate concentrations. Therefore, the Michaelis constants must be sufficiently low to achieve adequate catalytic acceleration in standard concentration ranges.

In relation to Eq. 1, one commonly describe enzyme catalytic power by their rate enhancement (RE), $\frac{k_{cat}}{k_{uncat}}$, and proficiency, $\frac{k_{cat}}{k_{uncat}}/K_{M1}K_{M2}$ [51]. The ratio $\frac{k_{cat}}{k_{uncat}}$ can be evaluated in terms of the differences in Gibbs free energy of activation, under the standard assumption that $\kappa_{cat}/\kappa_{uncat} \approx 1$.

$$\frac{k_{cat}}{k_{uncat}} = e^{-(\Delta G^{\ddagger}_{cat} - \Delta G^{\ddagger}_{uncat})/RT} = e^{-(\Delta\Delta G^{\ddagger}_{cat})/RT} \quad (3)$$

K_{M1} and K_{M2} can to first order be approximated with $K_S = 1/\exp(\Delta G_b/RT)$. The K_M constant expressed in the rate coefficients is (for a one-substrate enzyme):

$$K_M = \frac{[\mathrm{E}][\mathrm{S}]}{[\mathrm{ES}]} = \frac{k_{-1} + k_{cat}}{k_1} = K_S + \frac{k_{cat}}{k_1}, \quad (4)$$

and the error is thus in the order of k_{cat}/k_1. Because activation barriers for Diels–Alder reactions are large while binding rates are close to diffusion controlled, this assumption is well within the inherent computational errors of the methods employed.

The free energy of activation for the bond forming step ($\Delta G^{\ddagger}_{cat}$) has to be calculated by some QC method. Hence the reaction coordinate is divided into one classical regime, simulated by MD, and one QC regime, as illustrated in Fig. S14. To join these regimes is a nontrivial problem, which most importantly requires a common reference state. To this end, we employ 'near-attack conformers' (NACs) [52, 53], which can be represented by both MD (as a statistical distribution) and QC (as a representative structure). A NAC can be viewed as a point on the potential energy surface (PES) where the atoms involved in the bond forming process are aligned and within van der Waals distance. Toro-Labbe and coworkers have shown for a number of reactions that most electronic changes take place in a narrow transition region on the PES [54–56], which supports dividing it into two regimes.

The NAC for Diels–Alder can be described as a suprafacial overlap between diene and dienophile. The free energy of activation ($\Delta G^{\ddagger}_{cat}$) is then given by Eq. 5 (cf. Fig. S14).

$$\Delta G^{\ddagger}_{cat} = \Delta G_{\mathrm{NAC}} + \Delta G^{\ddagger}_{\mathrm{NAC}} \quad (5)$$

ΔG_{NAC} and $\Delta G^{\ddagger}_{\mathrm{NAC}}$ are the free energy differences going from the bound state to NAC, and from NAC to TS, respectively. ΔG_{NAC} can be seen as a measure of the enzyme's ability of pre-organization, since in a well-designed active site, the substrates are expected to be bound in TS-like configuration and ΔG_{NAC} should be small. If the enzyme-substrate complex is adequately sampled, ΔG_{NAC} can be calculated from an ensemble average:

$$\Delta G_{\mathrm{NAC}} = -RT \ln P_{\mathrm{NAC}} = -RT \ln \frac{N_{\mathrm{NAC}}}{N_{\mathrm{tot}}} \geq 0, \quad (6)$$

Note that sufficient sampling is required for this approach to be meaningful.

Methods and software

PDB mining and docking protocol (stage A)

All pre-processing of the initial set of enzymes was performed with the Relibase+ program [57, 58]. Exploration and comparing of cavities were done using the CavBase tool [59, 60], and all molecular docking was performed with GOLD 5.0 [61–65]. The combinatorial library used for virtual screening is based on two TS scaffolds (see Fig. 2), on which a set of 40 substituents has as been attached. The resulting ligands contain up to two substituents each. The docking protocol and combinatorial library have previously been described in ref. [37].

Molecular dynamics (stage B)

MD simulations were undertaken using the AMBER 10 package [66] and the ff03 force field [67]. Initial ligand poses were taken from the results in Stage A. The protocol is outlined in the electronic supplementary material (ESM) and is the same as described elsewhere [37].

For the mutant evaluation, 1 ns simulations (referred to as 'bursts') were undertaken from a common starting guess (the mutated crystal structure). Systems selected from the burst simulations were subject to 10 ns simulations to test the conformational stability of those mutants. With each substrate pair and mutant, we then performed three sets of 4 ns unconstrained simulations; one for the substrate pair in the active site, and two with the enzyme holding only the diene and dienophile, respectively.

Apart from evaluating the consistency of the interactions when one or both substrates are bound, we used these three systems to calculate the two binding constants (K_S) and ΔG_{NAC}. Binding energies are calculated from MD simulations using a slightly modified version of linear interaction energy (LIE) method [68–71]. In this method the free energy of binding is estimated from the differences in electrostatic and non-polar (van der Waals) interaction energies between the protein-bound and solvated ligands, using a parameterized relationship.

$$\Delta G_b = \alpha[\langle V_{l-s}^{vdw}\rangle_{\mathrm{bound}} - \langle V_{l-s}^{vdw}\rangle_{\mathrm{free}}] + \beta[\langle V_{l-s}^{elec}\rangle_{\mathrm{bound}} - \langle V_{l-s}^{elec}\rangle_{\mathrm{free}}] + \gamma \quad (7)$$

The generally accepted value of α is 0.18. β comes from the linear response to the electrostatic free-energy perturbation term and was originally set to 0.5. Now it is usually taken to be 0.43 for neutral ligands (or 0.37 if one OH group is present, as is the case of some ligands in this work) [70]. The parameter γ has been shown to depend on the hydrophobicity of the binding pocket [71, 72], but we have observed that it also varies within a series of ligands. For relatively non-polar, uncharged ligands, more accurate binding energies can be obtained by estimating γ from the non-polar solvation energy contribution (ΔG_{SA}) to ligand binding as calculated in the MM–GBSA routine of AMBER [37]. In particular, the deviations across a series of ligands of variable size are significantly reduced. This idea builds on the work by Jorgensen and coworkers [73–75] although we have chosen to keep the parametrization of Åqvist. We are currently looking into the details of this sort of correction [76], but for this study, as in the previous one, we have used a simple correlation of $\gamma = \Delta G_{SA}^{\mathrm{GB}}$.[1]

Quantum chemical calculations (stage C)

Since TS stabilization is central to an enzyme's catalytic power [50], an important evaluation of the enzyme design is to evaluate the TS structure in and outside the active site. TS stabilization is weighed in the design process at an early stage, since the combinatorial library represents 'loose' TS structures, but a TS optimization is needed to verify that there exists a gross catalytic effect.

As described previously [37], we have chosen to build a cluster model for the active site for QC evaluation, rather than a multilayer method such as QM/MM. It has been shown that remarkable accuracy can be achieved from cluster descriptions of native enzymatic systems [77, 78], while a QM/MM protocol was recently shown to be inferior to a combined MD and QC approach in differentiating active and inactive designs [34]. The active site model is built based on the enzyme-substrate interactions observed from the MD simulations, so as to represent the protein environment as well as possible. With contemporary computational resources, active site models of up to 250 atoms should be feasible.

Density functional theory (DFT) is well-established as an affordable and relatively accurate workhorse in computational chemistry. We have used the M06-2X functional [79–82] for both optimization and energy calculations. The functional has been designed to better capture several interactions crucial to Diels–Alder chemistry [83]. In our experience, M06-2X has been much more accurate than the commonplace B3LYP functional in predicting activation and reaction energies for Diels–Alder reactions, especially with larger basis sets and added solvent models [84, 85].

Activation and reaction energies were calculated relative to a representative NAC and compared to the most favorable uncatalyzed channel. Optimizations of the cluster were performed using Jaguar 7.6 [86] with the

[1] In this and our previous study [37], we have benchmarked the choice of LIE parametrization against several published de novo Diels–Alderases with good results [29, 31]. Additional comments are enclosed in the ESM.

6-31+G(d) basis set [87, 88], and single-point energies were obtained with Gaussian09 [89] at the M062X/6-311++G(d,p) level. Due to the extensive computational cost of calculating the Hessian of a large cluster model, and uncertainties associated with frozen atoms, thermodynamic corrections were added from the uncatalyzed states to estimate free energies and rate constants, which means that all thermodynamic effects between the NAC and TS cancel when calculating k_{cat}/k_{uncat}. This approximation is reasonable because the main entropic penalty occur in forming the bimolecular reaction complex, and we have previously estimated the difference in frequency corrections between cluster and the uncatalyzed system to be $\leq$1 kcal/mol [90]. A PCM [91, 92] solvent correction at the 6-31+G(d) level was added to the total energies, using $\varepsilon = 4$ to simulate a protein environment.

The uncatalyzed reactions were computed using Gaussian09 with the above-mentioned levels of theory. All TSs and ensuing products were optimized having the generally favored *endo-cis* geometry [93]. Activation and reaction free energies were calculated with respect to the separated, most stable conformations of the reactants (*trans* diene and *s-trans* dienophile [94]). The activation energy with respect to reactant complex was used to estimate TS stabilization in the enzyme. Solvent effects were added at the M06-2X/6-311++G(d,p) level using the SMD-PCM model [95] and $\varepsilon = 78.39$ (water).

Results and discussion

PDB mining

Although in principle any enzyme with an oxyanion hole could be redesigned to catalyze a reaction with an 'oxyanionic' TS, our efforts of mining the PDB for potential Diels–Alderase candidates have shown us that only a fraction of these enzymes can be seriously considered. The most limiting factor is, in our experience, cavity size and shape. Oxyanion holes are most prominent in EC class 3 (hydrolases) and 4 (lyases), and many of these enzymes have small cavities in which the sandwich-like TS of a Diels–Alder reaction cannot be accomodated. Furthermore, the oxyanion hole itself is often buried and is difficult to access even for native substrates, which indicates that they become directly bound only in the TS and acyl-enzyme stages of the hydrolytic reaction. Thus, out of all oxyanion hole containing enzymes, the actual number of structures to consider for Diels–Alderase design is quite small.

In a recent compilation of oxyanion hole contaning enzymes, Simón and Goodman [96] listed 310 structures. We decided to explore this subset of the protein data bank (PDB) [97], to see how many enzymes of this subset had necessary prerequisites for a potential Diels–Alderase. This means that the oxyanion hole should be accessible for the dienophile and that the cavity has to have a shape that allows the two substrates to be stacked on top of each other when forming the TS. Filtering was done by means of structure homology and visual inspection using the CavBase [59, 60] tool in Relibase+ [57, 58]. When $\approx$30 structures remained, they were clustered based on the cavities' mutual resemblance (see the ESM for details). A set of 10 structures with diverse properties were selected for virtual screening of TS models. The hydroxynitrile lyase presented in ref. [37] was one of them, and the focus of this study, the carboxylesterase Est30 from *Geobacillus stearothermophilus* (PDB entry 1TQH) [98] is another. Although both structures have an α/β-hydrolase fold [99] and share the Ser–His–Asp catalytic triad, the structures differ in how the active site is accessed. The former has a buried, tunnel-like active site [100], while in Est30 it is located between a α/β-hydrolase domain and a putative lid domain (see Fig. 1) [98]. It is not surprising the PDB mining resulted in a α/β-hydrolase structure, considering the vast size of this superfamily and the broad promiscuity exhibited by many of its members [4, 5, 101]. To date, Est30 represents a rare class of hydrolases [102], for which there is only one additional structures with a sequence similarity $\geq$20 % found in the PDB. This topic is further discussed in the ESM.

Stage A: virtual screening and static design

An overview of the semi-rational design process is given in Fig. 2a. As previously described [37], a combinatorial library was screened by adding substituents to a TS scaffold at six positions in two steps. Analysis of the results propelled rational refinement and trial of a number of substrates and mutants. An early mutation was the standard [37, 43, 103–105] removal of the nucleophilic serine in the catalytic triad (S94A), to quench hydrolytic activity and provide space around the oxyanion hole. Note that we changed the TS geometry from *exo* to *endo* after about half the refinement steps. This choice was induced by a preliminary MD simulation that falsified the proposed pose with the amide oxygen *cis* to the diene moiety (Fig. 2a), and in turn spurred numerous mutations to better accommodate the new motif, of e.g. F25, G130, V131, and V167. The M195Q mutation was attempted to potentially form a H-bond between the glutamine amide and the diene carbonyl in its new conformation. This event serves as an illustration of the benefits of iterating between two formally distinct steps in the design protocol.

Finally, the dienophile **1a** and dienes **2a** and **2b**, were chosen for the next stage. Figure 2b shows the histogram representation of the series of docking runs, illustrating

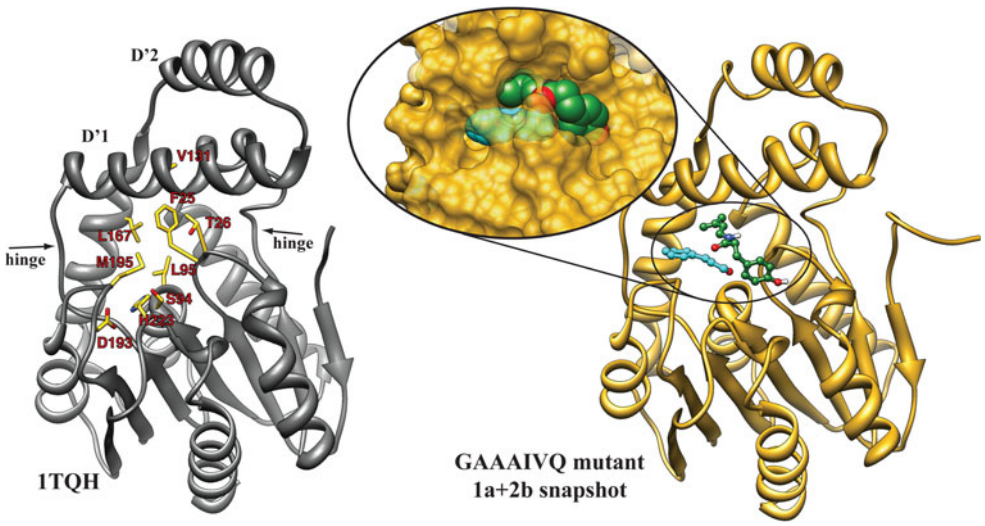
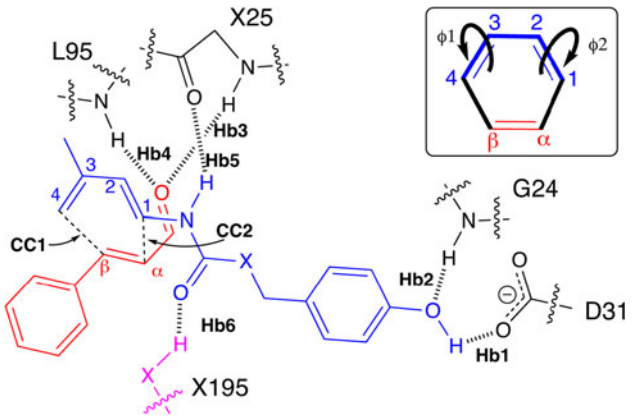


Fig. 1 Ribbon representation of carboxylesterase Est30 (PDB entry 1TQH), including some active site amino acids targeted in this study. The D′1 and D′2 helices belong to a putative lid, and the regions indicated by *arrows* are potential hinges (see ref. [98] and the ESM). Also displayed is a snapshot from an MD simulation of one of the lead mutants with iteratively designed substrates, to show how the diene and dienophile binds to the cavity. The *inset* shows a surface representation of the active site and a spacefill representation of the substrates

how scoring increased with refinement. Diene **2a** shares many properties with the diene used by Siegel et al. [29] in addition to several catalytic antibody designs [16–18, 106]. Our previous work also yielded dienes with a secondary amide in 1-position [37]. The reason for this apparent convergence is that there are few ways to catalyze the Diels–Alder reaction by direct hydrogen bonding to the diene end, and the amide group provides a feasible way to do so. The most notable novelty of this design is instead that all three designated catalytic elements are protein backbone atoms, in stark contrast to all previous Diels–Alderase designs. The H-bonding network observed in the final rounds of stage A is illustrated in Scheme 2, showing the diene–NH-group interacting with a backbone oxygen.

The diene variant **2b** was included because a potentially strained –O=C–O–CH_2 torsion was observed for **2a** in the



Scheme 2 Overview of all monitored geometrical variables between the enzyme and substrates **1** and **2**. The *inset* shows how the torison angles $\phi 1$ and $\phi 2$ are defined. The X group on the diene denotes either O or CH_2. Residue X195 is either Gln or Thr, if present, otherwise there is no hydrogen bond between the diene and this residue. For this reason, Hb6 is not included in the summation in Table 1

docking results. Indeed, a Mogul geometry check [107] shows that the dihedral angle of ≈60° for this motif is not present in the Cambridge Crystallographic Database (CSD) [108–110], while the corresponding **2b** torsion is well-represented. The same conclusion was reached from DFT calculations of the rotational profile of –O=C–X–CH_2 (Fig. S15).

Stage B: dynamic refinement and validation

Dynamic mutant screening

It is virtually impossible to judge whether a rational mutation gives the anticipated effect from docking results only. We therefore created an in silico ensemble of mutants, listed in Table 1, to evaluate and refine our design. Each mutant was evaluated by performing a 1 ns 'burst' simulation, once with the substrate composition **1a**+**2a** and once with **1a**+**2b**. The following mutations were attempted: F25 → G, A; S94 → A; G130 → A, V; V131 → L, I; L167 → V; M195 → Q, T. It soon became clear that T26 tends to obstruct the putatively catalytic H-bond between the diene's amide group and F/A/G25 carbonyl oxygen, by donating a H-bond from its sidechain. We therefore included this site as well and attempted T26 → A/V. All in all, we tested 20 variants including the wild-type (WT). Each variant is referred to as a seven-letter combination of the acronyms of the targeted sites and their particular mutation. The residues appear in numerical order (25, 26, 94, 130, 131, 167, 195). In this notation, the WT would be named FTSGVLM.

As descriptors for a comparative analysis of the variants, we chose to calculate a series of fractions, $p_x = N_x/N_{tot}$, where, N_x is the number of frames where x is within a predefined threshold and N_{tot} is the total number of frames

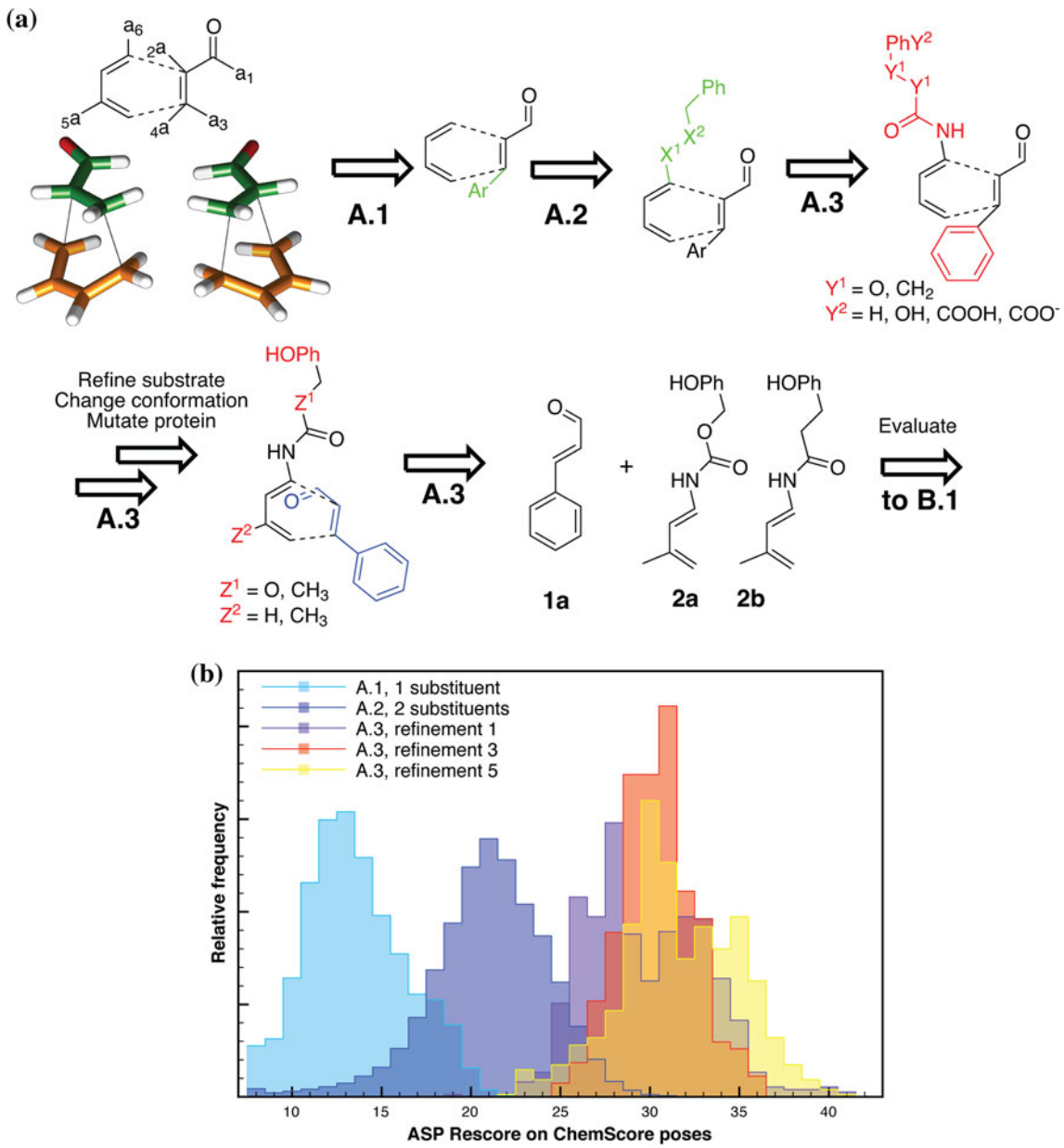


Fig. 2 Illustration of iterative substrate/enzyme design. **a** Model TS scaffolds and allowed attachment points (a1–a6), followed by evolution of substrate during stage A. Functionalities introduced in the first two phases are *colored green*. Subsequent substituents are *colored red*, while the conformational change of the dienophile is highlighted in *blue*. **b** Normalized *histograms* of ASP scores on docking poses of a series of refinement runs

(5,000 for a burst simulation). We decided to monitor a number of key hydrogen bonds, both the ones assumed to be catalytic in the design and those important for substrate binding and specificity. These descriptors are depicted in Scheme 2. The threshold for the p_{Hb} fractions was set to 2.10 Å. The sum of all fractions, $\sum p_{\mathrm{Hb}}$, was then taken as the overall H-bonding quality of the enzyme. The NAC fraction (p_{NAC}) is also treated as a key descriptor since it is a measure of how prone the substrates are at aligning themselves in a TS-like geometry. By visual inspection, we found that in some variants, the plane of the dienes's C_1-C_2-C_3-C_4 scaffold aligned perpendicular to the plane of the dienophile, which is of course undesired (see Fig. 3a and discussion below). The two incipient C–C distances are still short in these cases, however, leading to poor exclusion. We therefore added a constraint to our NAC description, forcing the diene and dienophile π-systems to be sufficiently coplanar. The total NAC constraint leading to the results reported in Table 1 can thus be summarized as in Eq. 8.

Table 1 Initial mutant evaluation

Mutant[a]	**1a + 2a**		**1a + 2b**		Both systems[d]	
	$\sum p_{\mathrm{Hb}}i$[b]	p_{NAC}[c]	$\sum p_{\mathrm{Hb}}i$[b]	p_{NAC}[c]	$\Delta\sum\sum p_{\mathrm{Hb}i}$	$\Delta\sum p_{\mathrm{NAC}}$
FTSGVLM[e]	1.37	0.00	1.24	0.00	0.00	0.00
ATAVVVM	1.46	0.05	1.31	0.12	0.16	0.16
GTAGVVM	1.64	0.05	1.19	0.18	0.22	0.22
GTAAVVM	1.11	0.02	1.35	0.25	−0.14	0.27
GTAAVVQ	1.38	0.23	1.32	0.26	0.10	0.49
GTAAVVT	1.79	0.17	1.40	0.31	0.58	0.48
GTAALVM	1.08	0.00	1.36	0.20	−0.17	0.20
GTAAIVQ	1.28	0.22	1.29	0.28	0.07	0.49
GTAVVVM	1.38	0.17	1.39	0.27	0.16	0.45
GTAVVVQ	1.53	0.25	1.15	0.35	0.07	0.59
GTAVVVT	1.45	0.13	1.29	0.30	0.14	0.43
GAAAVVQ	1.62	0.12	1.63	0.17	0.65	0.29
GAAAVVT	1.17	0.05	1.71	0.22	0.27	0.27
GAAVVVQ	1.59	0.06	1.55	0.23	0.54	0.28
GAAVVVT	1.59	0.03	1.56	0.16	0.54	0.19
GAAAIVM	1.67	0.12	1.51	0.33	*0.57*	*0.45*
GAAAIVQ	1.55	0.24	1.57	0.26	*0.52*	*0.49*
GAAAIVT	1.54	0.13	1.62	0.02	0.55	0.15
GVAAIVM	1.76	0.00	1.62	0.33	*0.78*	*0.33*
GVAAIVQ	1.74	0.06	1.55	0.27	*0.69*	*0.33*

[a] Mutated residues are shown in boldface

[b] H-bonds considered are displayed in Scheme 2. Hb1 was not included since $p_{\mathrm{Hb1}} \approx 1$ for all variants

[c] Defined according to Eq. 8

[d] Summation over both systems followed by subtraction of WT score

[e] Wild-type (WT)

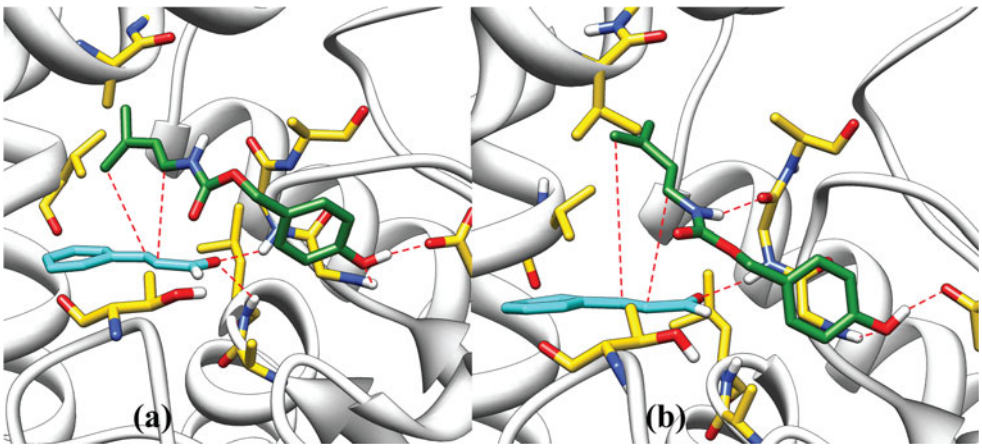


Fig. 3 Typical examples of unproductive poses from the burst simulations of **1a**+**2a**. **a** The diene is tilted almost perpendicular with respect to the dienophile, resulting in short C–C distances but no NAC. **b** The diene is bent away from the dienophile face and cannot form two short contacts

$$\mathrm{NAC}: \begin{cases} \mathrm{CC1} = d_{4-\beta} \leq 4.0\,\text{Å} \\ \mathrm{CC2} = d_{1-\alpha} \leq 4.0\,\text{Å} \\ \phi 1 = \phi_{\alpha-1-2-3} \geq 40^{\circ} \\ \phi 2 = \phi_{2-3-4-\beta} \geq 40^{\circ} \end{cases} \quad (8)$$

The ϕ variables represent dihedral angles, as shown in Scheme 2. It is still crude to judge a design solely based on an, albeit careful, attempt to quantify subtle differences. Therefore all simulations have also undergone visual inspection.

We note from Table 1 that the difference in $\sum p_{\mathrm{Hb}}i$ scoring between the variants is not so dramatic, meaning that the H-bond pattern is conserved in all variants. Most variants gave reasonable p_{NAC} scores during the limited simulation

time of the evaluation. Overall, the conformational integrity of each ligand was well-preserved. The differences become clearer if the scores from the WT are used as baselines, and subtracted from all scores. The resulting values, shown in the rightmost columns in Table 1, range from ≈ -0.2 to $+0.8$ for the H-bond descriptors, and the $\sum p_{NAC}$ descriptors are improved (from zero) to at most 0.5.

The main difference between the variants evidently lies in their ability to produce NACs. This ability is mostly affected by the subtle changes in cavity shape and volume introduced by the aliphatic mutations. We conclude from Table 1 that the **1a**+**2b** substrate pair gives higher p_{NAC} in general, presumably due to the more flexible $–CH_2–CH_2–$ motif in **2b** compared to $–O–CH_2–$ in **2a**. The reasons for the poor values of p_{NAC} recorded in a few cases are invariably that the diene motif bends away from the dienophile C–C plane in two ways, as illustrated in Fig. 3. Decreasing the space available directly above the dienophile somewhat mitigate this problem, as in the variants containing the G130A/V and V131I mutations.

The variants standing out as having the highest overall descriptor values are those containing the T26A/V, G130A, V131I and L167V mutations together with either M195 or Q195. Their scores are shown in italics in Table 1. The exception is GTAAVVT, but we chose not to proceed with this variant since the catalytic dne–NH–O–G25 H-bond is hindered by a consistently stronger T26–OH–O–G25 H-bond. Visual inspection of the trajectories of GTAAVVT and the similar GTAAIVQ revealed that although p_{NAC} (as defined in Eq. 6) is high for each substrate pair, there are tendencies of T-stacking (as in 3a) as well as 'shearing' between the planes of the substrates' π frameworks. Consequently, we selected the four variants GAAAIVQ, GAAAIVM, GVAAIVQ, GVAAIVM for detailed evaluation, even though we note that the results from several other variants are also promising.

Lead mutant evaluation

The four selected variants were then tested for stability with respect to fold and shape of the active site. This was done by performing 10 ns MD simulations and comparing the results to the WT. The results are discussed extensively in the ESM, and the most important conclusions are that all variants appear stable with respect to the wild-type and that the active site regions are inflexible and do not shift much during the simulations. Importantly, no major distortions of the residues surrounding the active site were found. The putative lid region (the D′1 and D′2 helices in Fig. 1) is indeed mobile with respect to the rest of the enzyme, suggesting that substrate binding could be coupled to an opening and closing motion of the enzyme.

Substrate evaluation

As described above, each variant was run with each of substrates **1a**, **2a** and **2b** in binary complexes. The augmented LIE binding energies estimated from these simulations are given in Table 2 (details on these calculations are provided in Table S1). Both dienes appear to be strong binders in all variants, which is reflected by the tight fit of the Ph–OH group between G24 and D31.

The dienophile does not bind specifically to the oxyanion hole when alone in the active site, and consequently the binding constants are quite poor. Since ΔG_b for the dienes is lower, it is however reasonable to assume that the dienophile binds cooperatively to an enzyme-diene complex. We therefore calculated the LIE binding energies of each substrate in ternary complexes (which are discussed below). The results, displayed in Table S2, show that **1a** benefits from the presence of the diene, with 1–1.5 kcal/mol lower binding energies. The diene binding is weakened by 1–2 kcal/mol, probably due to increased strain compared to the more relaxed conformations in the binary systems. Interestingly, the strongest diene-binding binary systems (GAAAIVM and GVAAIVM) are weakened the most.

Despite the improved binding of **1a** in the ternary systems, there is an imbalance in substrate affinity that seems symptomatic to attempts of Diels–Alderase design [29, 37]. Since the more weakly binding substrate will be limiting to specificity and relative rate (Eq. 1), we introduced a second

Table 2 Estimated binding energies from MD simulations (binary systems)

Variant	**1a**		**1b**		**2a**		**2b**	
	ΔG_b^a	K_M^b	ΔG_b^a	K_M^b	ΔG_b^a	K_M^b	ΔG_b^a	K_M^b
GAAAIVQ	−1.6	68	−4.7	0.38	−6.1	0.034	−5.7	0.059
GAAAIVM	−2.1	28	−5.5	0.087	−7.0	0.0068	−6.5	0.016
GVAAIVQ	−0.89	224	−4.9	0.25	−4.6	0.44	−4.9	0.24
GVAAIVM	−1.5	84	−4.7	0.34	−7.1	0.0065	−6.7	0.013
WT	−2.6	13	–	–	−4.8	0.30	−5.1	0.17

[a] Binding energies, reported in kcal/mol

[b] Binding constants in 10^{-3} M^{-1}

dienophile, *p*-nitro-cinnamaldehyde **1b**. We assumed it would reside more tightly in the oxyanion hole while also being more activated for Diels–Alder catalysis. As can be seen in Table 2, **1b** is a much stronger binder to all variants, and the simulations reveal that it binds more specifically to the oxyanion hole than **1a** in addition to forming a H-bond between the nitro group and the backbone –NH– of Lys122. (Additional statistics from simulations of the binary complexes are provided in the ESM.) The binding energy is essentially conserved in the ternary complexes (Table S2).

We note that the two M195 variants yield lower ΔG_b for the dienes than the Q195 variants. This finding contradicts one of the anticipated effects of introducing the glutamine, namely molecular recognition and stronger diene binding. When investigating the apparent stronger association to the M195 variants in the enzyme-diene trajectories, we found that Q195 does not form any H-bonds with the diene. However, in the M195 variants the dienes are more perturbed from their starting poses and even partially occupy the oxyanion hole. In the Q195 variants, they remain prearranged in a conformation similar to that of the designed ternary complexes. Hence for the dienes, there seems to be a trade-off between binding affinity and catalytic prearrangement.

In comparison with the lead mutants, the wild-type binds the substrates surprisingly well. For the dienes, this can be rationalized by the conserved G24, D31 functionality. The dienophile **1a** forms a H-bond to S94 instead of accessing the more buried oxyanion hole. From Table 1 it is however obvious that the wild-type is inferior to the mutants with regard to the ternary complex, which, as the previous paragraph, suggests a strong contrast between general affinity a the single ligand and the ternary reaction complex.

For the ternary complexes, we performed 4 ns simulations with the four lead mutants and substrate pairs **1a**+**2a**, **1a**+**2b** and **1b**+**2b**. We monitored the same descriptors as defined in Scheme 2, and their distributions from the last 2 ns are given in Fig. 4. As in the burst simulations, **Hb1** and **Hb2** are well maintained and have nearly identical distributions in all instances. The oxyanion hole interactions **Hb3** and **Hb4** are consistent distributions in all complexes, with **Hb4** being slightly weaker.

The diene H-bond donation to the enzyme (**Hb5**) is strong when **2a** is involved, while this interaction seems weaker in the **2b** cases. The **CC1** distribution in the **2a** complexes is however poor except in GAAAIVQ, which can be attributed to the rigidity of **2a** compared to **2b**. This resonates in the calculated NAC penalties, which are reported in Table 3 and follow the same trend as those in Table 1. GAAAIVQ is the only variant in which ΔG_{NAC} is <1 kcal/mol for **1a**+**2a**.

From Fig. 4, it is also clear that the diene-Q195 H-bond (**Hb6**) is nonconsistent. Two regions can be seen in **Hb6** distributions; one where a H-bond to the diene exists and one where larger distances indicate that the Q195 side chain is rotated away from the substrates and interacts with the solvent. Two snapshots of the GAAAIVQ–**1a**+**2b** trajectory are shown in Fig. 5, illustrating two typical conformations of Q195 corresponding to the two regions. By introducing Q195, we anticipated that NAC formation would be facilitated by aligning the diene with the dienophile. From Table 3, one cannot say with confidence that it has, although the Q195 variants yield smaller ΔG_{NAC} penalties compared to the corresponding M195 variants in three out of four cases.

Turning to the second varied residue, A/V26, we see from both Tables 2 and 3 that differences in the mutant's prowess for both substrate binding and NAC is small. They are not systematic with respect to varying substrate pairs, and we state that both mutations are promising for the present Diels–Alderase design.

The statistics from the **1b**+**2b** simulations reveal that the dienophile is more tightly bound to the oxyanion hole (**Hb3**, **Hb4**) than **1a**. Although yielding acceptable p_{NAC} values, they are lower than for the corresponding **1a**+**2b** ternary complexes. This is the result of a distorted alignment between incipient carbons due to the buriedness of **1b**. It can also be seen from Fig. 4 that the diene is distorted in the M195 variants; in GAAAIVM the Ph–OH 'anchor' leaves its subpocket for the solvent, while in GVAAIVM rotation of the C_2–C_3 bond is responsible for the poor p_{NAC}.

We finally tested the stability of the design by running a 10 ns simulation at 300 K as well as 4 ns simulations at elevated temperatures (details are found in the ESM). We selected the GAAAIVQ–**1a**+**2b** system based on the good overall results in binding and NAC formation outlined above. The longer simulation was used to test the statistical error of the p_{NAC} parameter as well as the sensitivity to changes in the definitions in Eq. 8. We found that the p_{NAC} and ΔG_{NAC} have errors of roughly 25 % of their total values, and that ΔG_{NAC} is increased by ≈ 1 kcal/mol when the NAC distance is decreased to 3.6 Å (see the ESM for details).

The heated simulations showed that the redesigned, thermophilic Est30 was able to maintain p_{NAC} of $\approx$3–5 % at temperatures as high as 365 K, with all key interactions conserved. A linear regression of ΔG_{NAC} from seven temperatures interestingly yielded ΔH_{NAC} and ΔS_{NAC} values of -7.7 kcal/mol and -28 cal/mol K (-8.4 kcal/mol at 300 K), respectively. In the 10 ns simulation, the average p_{NAC} value was $\approx$25 % overall and 18 % during the last nanosecond.

To conclude the stage B evaluation, we note that the recorded differences of 1–2 kcal/mol between the different variants are probably the highest available resolution at this computational level. Further quantitative distinction is not viable given the systematic errors associated with the applied methods. In all, the four mutants are promising

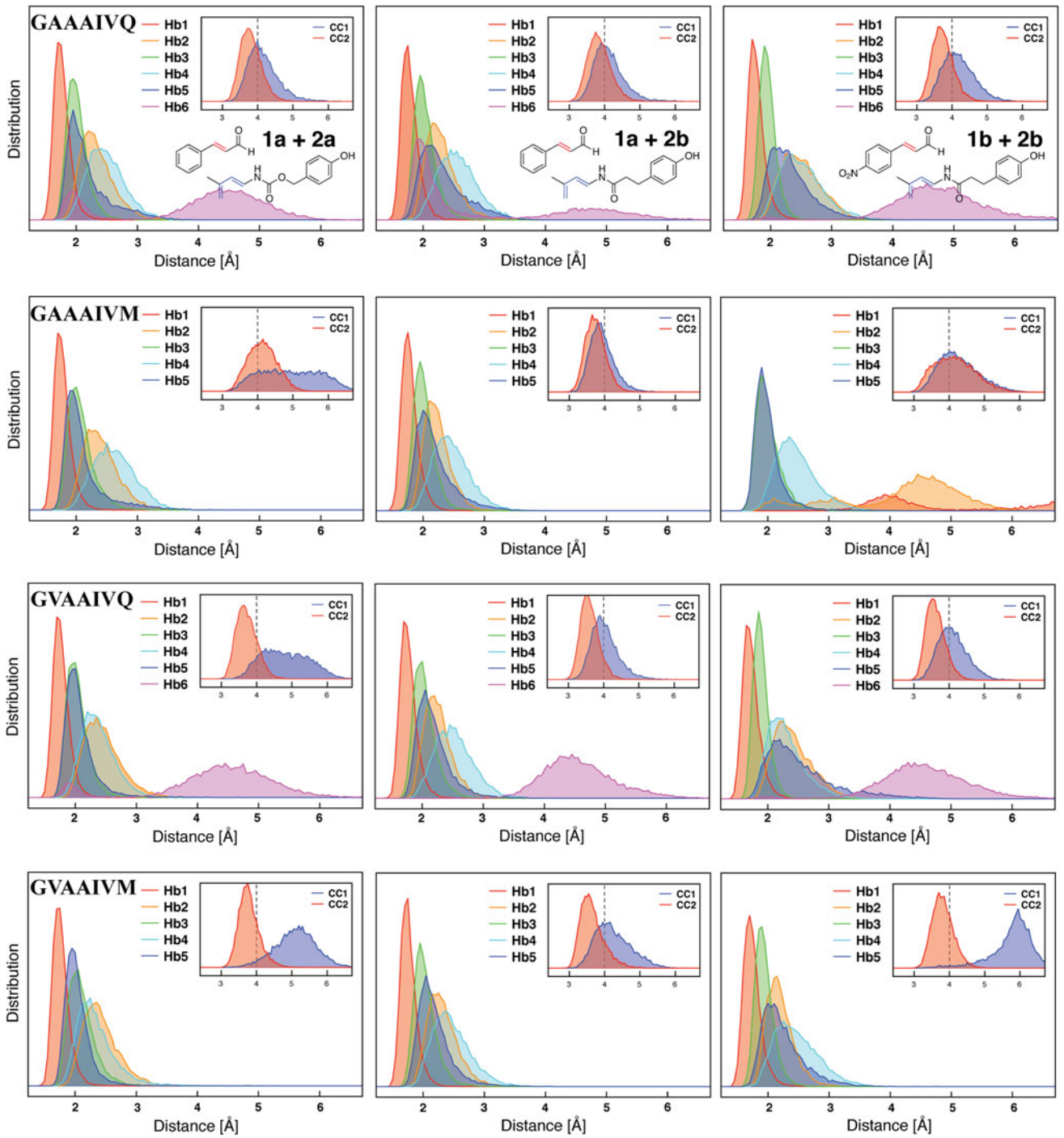


Fig. 4 Distance distributions from the last 2 ns (10,000 frames) of the four lead mutants (*rows*) with three substrate combinations (*columns*). The main *plots* show H-bond distributions and the *insets* show incipient C–C distributions. The *vertical line* at 4.0 Å represents the threshold for a NAC state. For visual reference of the distances (see Scheme 2; Fig. 5)

designs in terms of molecular recognition and specific substrate binding.

Stage C: quantum chemical evaluation

In the last stage, an active site model was designed from the crystal structure coordinates of the GAAAIVQ mutant. Although this does not take into account relaxation of the enzyme and active site, an approach based on several (weighed) representations of structure would increase computational time manyfold. Moreover, the relative rigidity of the cavities revealed in Fig. S3 justifies this single representation. Two models of the active site were constructed: a large version with 161 atoms and a minimal

Table 3 p_{NAC} and ΔG_{NAC} estimated from MD simulations (ternary systems)

Variant	**1a**+**2a**		**1a**+**2b**		**1b**+**2b**	
	p^{a}_{NAC}	ΔG^{b}_{NAC}	p^{a}_{NAC}	ΔG^{b}_{NAC}	p^{a}_{NAC}	ΔG^{b}_{NAC}
GAAAIVQ	0.24	0.85	0.26	0.80	0.11	1.29
GAAAIVM	0.11	1.29	0.29	0.73	0.20	0.95
GVAAIVQ	0.11	1.31	0.35	0.63	0.30	0.72
GVAAIVM	0.03	2.04	0.28	0.75	0.011	2.68

[a] Using definition in Eq. 8

[b] Calculated using Eq. 6 energies in kcal/mol

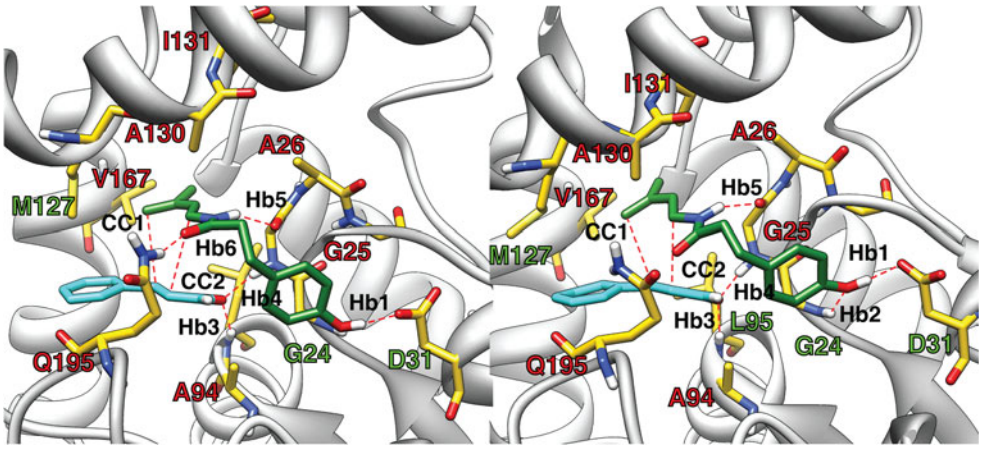


Fig. 5 Two late snapshots from the 4 ns simulation of **1a**+**2b** in GAAAIVQ. Distances reported in Fig. 4 are *labeled* in *black*, mutated residues in *red* and auxiliary, unaltered residues in *green*. Note the two typical modes of Q195; in the *left panel*, it interacts with the diene's carbonyl oxygen, while in the *right panel*, it is turned away and interacts with the solvent

version with 31 atoms. Including substrates, the large model consists of 211–215 atoms. The minimal model is comparable to a 'theozyme' [26, 29] in size and function, and can be regarded as an idealized verison of the binding site where the substrates are not so conformationally restricted. All residues were truncated to include important side chains and backbone peptides. The dangling carbons, as well as the hydrogen replacing the propagating chain, were frozen to maintain the active site structure. In addition, selected peptide atoms were frozen to compensate for the fact that atoms interacting with them were not included in the model. A detailed representation of the models with indicated frozen atoms is shown in Fig. S16.

The three substrate pairs considered above were modeled in both versions of the active site. In addition, **1a**+**2b** was modeled in a version with the wild-type M195 residue (representing the GAAAIVM and GVAAIVM variants) to deduce what effect the M195Q mutation can have on TS stabilization. We did not explicitly consider models with the T26V mutation, since the difference between Ala and Val was deemed to have little effect on the quantum chemical representation of the system.

Three points along the reaction coordinate were optimized; a reaction complex (representing the NAC), the TS and the product. Optimized structures with **1a**+**2b** are shown in Fig. 6. (Cartesian coordinates for all other optimized geometries are available in the ESM.) The TS and reaction free energies, including the NAC penalty from Table 3, are presented and compared to the corresponding uncatalyzed reactions in Table 4.

From Fig. 6 we see that Q195 forms a H-bond to the diene in all optimized structures. When comparing $\Delta G^{\ddagger}$ to that of the M195 model, Q195 seemes favorable for TS stabilization. One can argue that because Q195 spends so little time interacting with the diene in the MD simulations, a representative TS should not have this H-bond. However, as seen from the radial distribution function in Fig. S9, the diene carbonyl oxygen interacts with a number of solvent water on average. Because water essentially contributes in the same way as the Q195 NH_2 group, the diene–Q195 interaction can be viewed as a rough representation of a generic, if not completely resolved, enzyme-catalyzed TS.

The **1a**+**2b** substrate pair yields the lowest $\Delta G^{\ddagger}_{NAC}$ as well as the largest TS stabilization ($\Delta\Delta G^{\ddagger}_{cat,NAC}$ in Table 4). The estimated RE (k_{cat}/k_{uncat}) for the three substrate pairs, also provided in Table 4, are typical for recent de novo designs and promiscuous systems [8, 22, 27, 28], but higher than the Diels–Alderases reported by Siegel et al. [29] The modest TS stabilization is significative of H-bond catalyzed Diels–Alder [37, 43]. In contrast, ideal theozymes have shown a corresponding TS stabilization of $\approx$5–10 kcal/mol

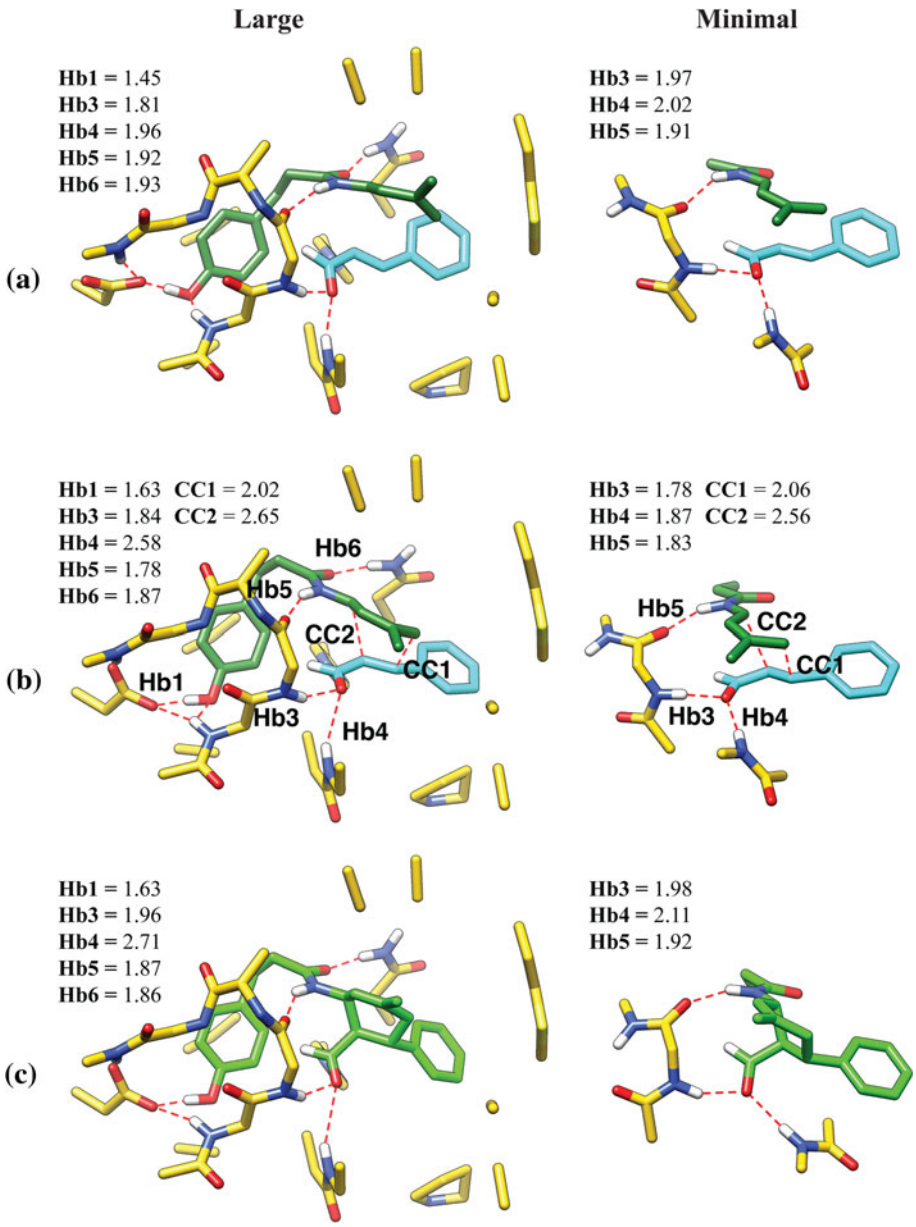


Fig. 6 Optimized geometries of **a** reaction complex (NAC), **b** TS and **c** product of **1a**+**2b** in the large and minimal active site models. All but key hydrogens are hidden for sake of clarity. The view is rotated approximately 180° about the vertical axis with respect to Fig. 5. **b** Displays bond labeling, which is identical with the statistical MD analysis. Each relevant distance is given in Angstroms for each *panel*. (**Hb2** was left out since the backbone hydrogen tends to bind to he aspartate residue in the models)

[29, 106], and this is indeed the kind of energy we obtain from the small active site models. It is reasonable to assume that the lower $\Delta G^{\ddagger}_{\mathrm{NAC}}$ values in the minimal model originate from strain relief relative to the large model (see Fig. 6). By superimposing the two clusters (Fig. S16), it is clear that the reactant conformations assumed in the minimal model is impossible in the real active site, and comparing $\Delta\Delta G^{\ddagger}_{cat,\mathrm{NAC}}$ of the two models is therefore a way of quantifying the cost of distorting these 'ideal' poses.

Methodology validation

The relative rates (Eq. 1) can be plotted as function of substrate concentrations (and constant $[\mathrm{E}]_0$) to make a composite ranking of the enzyme designs. It essentially includes all parameters that a computational design can optimize. We used four experimentally characterized de novo Diels–Alderases [29, 31] to benchmark our method, by performing MD simulations of ternary complexes and calculating K_{M1}, K_{M2} and ΔG_{NAC} as described. The 'theozyme' used by Siegel et al. [29] was used as QC model (see the ESM for additional details).

Correlations between computational predictions and experimental data, shown in Fig. 7a, are very good and serve as an adequate validation of our method in the absence of experimentally assessed mutants (see also Table S5). Specifically, the computations rank the Diels–Alderases correctly in the 'practical concentration range' marked by dotted lines in Fig. 7a. In this region, the relative rate is limited by k_{cat}/k_{uncat}, and since the same QC

Table 4 M062X/6-311++G(d,p) free energies of the active site models

Model	Substrates	$\Delta G^{\ddagger a}_{NAC}$	$\Delta G^{\ddagger a}$	ΔG^{b}_{rxn}	$\Delta\Delta G^{\ddagger c}_{cat}$	k_{cat}/k_{uncat}	$\Delta\Delta G^{\ddagger d}_{cat,NAC}$
Uncat.	**1a**+**2a**	19.3	26.0	−13.5			
	1a+**2b**	20.2	27.1	−13.3			
	1b+**2b**	19.6	24.7	−14.7			
Large-Q195 (GAAAIVQ)	**1a**+**2a**	18.4	19.2	−22.8	−6.8	1.0×10^5	−0.9
	1a+**2b**[e]	17.2	18.1	−20.6	−9.0	4.6×10^6	−3.0
	1b+**2b**	18.3	19.6	−20.6	−5.1	5.6×10^3	−1.3
Large-M195 (GAAAIVM)	**1a**+**2b**	18.8	19.5	−20.0	−7.6	3.6×10^5	−0.5
Small	**1a**+**2a**	13.4	–	−24.7			−5.9
	1a+**2b**	14.6	–	−27.0			−5.6
	1b+**2b**	11.5	–	−21.4			−8.1

Energies in kcal/mol. Optimized geometry and thermodynamic corrections are from the M062X/6-31+G(d) level. Solvent effects have been taken into account with $\varepsilon = 78.4$ for the uncatalyzed reactions and $\varepsilon = 4.0$ for the active site models

[a] For the enzyme-catalyzed reactions: calculated using Eq. 5 with values from Table 3

[b] Relative to separated reactants (*trans* diene and *s-trans* dienophile) in the uncatalyzed case, and relative the reaction complex in the active site models

[c] $\Delta\Delta G^{\ddagger}_{cat} = \Delta G^{\ddagger}_{cat} - \Delta G^{\ddagger}_{uncat}$

[d] $\Delta\Delta G^{\ddagger}_{cat,NAC} = \Delta G^{\ddagger}_{cat,NAC} - \Delta G^{\ddagger}_{uncat,NAC}$. This value gives an estimate of the explicit transition state stabilization in the active site

[e] Using data from the 10 ns simulation, where $p_{NAC} = 0.24$ during the last 5 ns

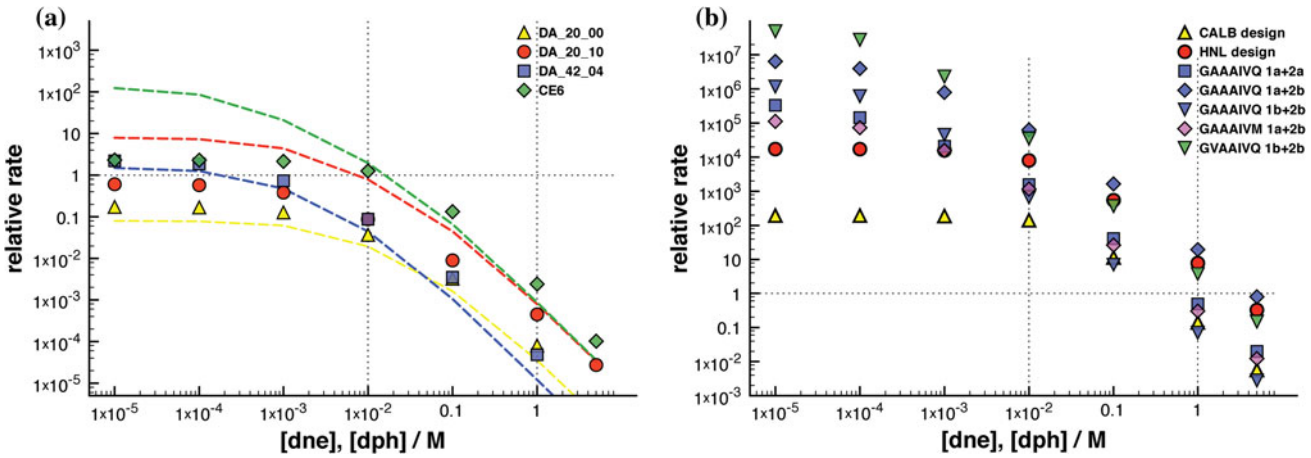


Fig. 7 v_{cat}/v_{uncat} (Eq. 1) *plotted* against projected substrate concentrations and a generic $[E]_0 = 10^{-5}$ M. The *horizontal dashed line* marks $v_{cat}/v_{uncat} = 1$, and the *vertical lines* show the practical substrate concentration range. Values at the *left*-hand side of the *graphs* are proportional to $\frac{k_{cat}}{k_{uncat}}/K_{M1}K_{M2}$. **a** *Curves* using experimental (*dashed*) and calculated values of k_{cat}/k_{uncat}, K_{M1} and K_{M2} for DA_20_00, DA_20_10, DA_42_04 [29] and CE6 [31] (b). Theoretical curves for selected systems reported in this article. An average value of ΔG_b from the binary and ternary complexes has been used for the corresponding K_M values in the *graph* (Tables 2 and S2, respectively). Two previous designs are included [37, 43] for comparison

model was used to compute $\Delta G^{\ddagger}_{NAC}$ for all enzyme variants, the ranking is governed solely by the ΔG_{NAC} values recorded from our simulations.

Summary and predicted activity

Turning back to our designs, we see from Fig. 7b that the GAAAIVQ/GVAAAIQ variants appear to give the best overall results, which can be attributed to balanced substrate binding, NAC propensity and TS stabilization. The substrate pair **1a**+**2b** yields the highest RE, partly due to a slower uncatalyzed reaction (Table 4) [111].

The Est30 designs are a leap forward in terms of enzyme-substrate complementarity, both compared to our previous designs and the Diels–Alderases from the Baker lab [29, 31]. From the estimated RE and proficiency values (Tables S4 and S5), the main differences between the

present designs and the Baker enzymes can be identidied as (1) a high propensity to form NACs and (2) slightly better substrate binding. The small $\Delta\Delta G^{\ddagger}_{cat,\mathrm{NAC}}$ values suggest that there is no dramatic increase TS stabilization from the NAC, and that a large part of the barrier reduction comes from pre-organization. A part of the difference in $\Delta\Delta G^{\ddagger}_{cat}$ can also be traced to a higher uncatalyzed barrier in the present case.

We note that at T_{opt} of the wild-type Est30 (70 °C), we estimate k_{cat} to be almost two orders of magnitudes larger than at 25 °C, although k_{cat}/k_{uncat} decreases to $\sim 10^2$. This estimate is based on a regression of the p_{NAC} at different temperatures (see the ESM for details), and rescaling of the uncatalyzed thermodynamic corrections. Hence, provided the designed mutants remain as thermophilic as the wild-type, they could be highly attractive as industrial biocatalysts. We furthermore propose that due to the specific binding conformations of the ternary complexes and the previously observed inability to bind the *exo* conformation, the enzyme-catalyzed reaction will be highly stereoselective.

We conclude the discussion with the sobering note that despite recent years' successful incorporation of new catalytic functions in enzymes, net RE values remain modest. Our repeated estimates of stabilizations of the Diels–Alder TS of $\leq$3 kcal/mol [37, 43] suggest possible limitations of H-bond catalysis of Diels–Alder. A comparison of the minimal and large cluster suggests that greater RE is possible, but it has proven difficult to 'have it all' in terms of binding, pre-organization and TS stabilization, regardless if one uses an 'inside-out' strategy (beginning with an idealized TS stabilization) or our combinatorial strategy (beginning with idealized binding of a TS dummy model). It has been widely discussed why computationally designed enzymes perform so poorly (rate accelerations $\leq 10^6$) [8], and which strategies should be used to overcome the problems. The results from our two cluster models, where the differences in $\Delta G^{\ddagger}_{\mathrm{NAC}}$ are $\sim$3–7 kcal/mol, show the sensitivity to structural distortion. A design that captured the 'ideality' of the minimal model would have a 2–4 orders of magnitude higher catalytic rate compared to the best instances in this study. It is however recognized that non-covalent enzyme catalysis has an upper limit [112], which muffles the potential of designing enzymes with native-like efficiencies as long as the target reaction mechanism remains simple. We have recently showed that much larger Diels–Alder TS stabilization is possible if one proceeds through an acid/base-catalyzed route; [90] this might be a fruitful model for future Diels–Alderases.

Our study emphasizes the difficult key challenge in computational enzyme design: to find the 'perfect match' of mutant and substrate(s). Optimize the TS stabilization and you run into problems caused by unexpected dynamic behavior; optimize for binding and you pay the toll in loss of TS stabilization. The reason for this is that regardless of which is prioritized, the 'ansatz' of the catalytic system is in a static form. A dynamic approach gives the possibility to single out more successful models from failing ones [34], but tools to correct flaws in the ansatz itself are, at best, blunt. In this work, we used a variable ansatz and focused on using dynamic simulations to single out the most promising systems, and it is conceivable that the design was too biased towards substrate binding, leading to the ternary complex becoming overly stabilized with respect to the TS. Nevertheless, we believe that this approach is a step toward a flexible and controllable design protocol, since one is never stuck with a rigid design template.

Conclusions and outlook

This is the second report where we use a combinatorial approach to substrate design while simultaneously manipulating the active site to maximise its ability to promote the Diels–Alder reaction. In this work, we employed more extensive MD screening and were able to take into account the dynamics of the system early in the design phase. Twenty protein variants were initially considered, of which four were selected for further study. We propose that the best performing mutant taking everything into consideration is the GAAAIVQ mutant. It has a high predicted proficiency and the specific interactions made to the TS model suggests a high degree of stereospecific catalysis. The main catalytic interactions come from backbone protein atoms, which have smaller positional uncertainty than side chain atoms and thus suggest a robust design. The ternary complex is furthermore shown to be stable with respect to increased temperatures for 4 ns. These are all attractive attributes in the context of industrial application [113].

We have demonstrated that the MD-based protocol for calculating binding and NAC energies correlates well with available experimental values, thus enabling us to discuss quantitatively the reasons behind less and more successful designs. Next to sufficient TS stabilization and substrate binding, a crucial parameter seems to be the propensity to pre-arrange substrates for the TS conformation, in this work represented by the NAC concept.

Finally, this study sheds light on the challenges for computational enzyme designers in balancing the incorporation of dynamics with designing for optimal TS stabilization. Our approach allowed us to find enzyme-substrate complexes with specific binding modes and high

complementarity, while adequately conserving the catalytic interactions found in the initial docking phase.

Acknowledgments This work has been supported by the Swedish Research Council (VR) and Cambridge Crystallographic Data Centre (CCDC). The authors thank Dr. Colin Groom of the CCDC for helpful discussions and comments.

References

1. Baker D (2010) Protein Sci 19:1817–1819
2. O'Brien PJ, Herschlag D (1999) Chem Biol 6:R91–R105
3. Kazlauskas RJ (2005) Curr Opin Chem Biol 9:195–201
4. Hult K, Berglund P (2007) Trends Biotechnol 25:231–238
5. Svedendahl Humble M, Berglund P (2011) Eur J Org Chem 2011:3391–3401
6. Gerlt JA, Babbitt PC (2009) Curr Opin Chem Biol 13:10–18
7. Pantazes RJ, Grisewood MJ, Maranas CD (2011) Curr Opin Struct Biol 21:467–472
8. Saven JG (2011) Curr Opin Chem Biol 15:452–457
9. Kelly WL (2008) Org Biomol Chem 6:4483–4493
10. Pohnert G (2001) ChemBioChem 2:873–875
11. Ose T, Watanabe K, Mie T, Honma M, Watanabe H, Yao M, Oikawa H, Tanaka I (2003) Nature 422:185–189
12. Kim HJ, Ruszczycky MW, Choi S-h, Liu Y-n, Liu H-w (2011) Nature 473:109–112
13. Guimaraes C, Udier-Blagovic M, Jorgensen W (2005) J Am Chem Soc 127:3577–3588
14. Serafimov JM, Gillingham D, Kuster S, Hilvert D (2008) J Am Chem Soc 130:7798–7799
15. Hilvert D, Hill KW, Nared KD, Auditor MTM (1989) J Am Chem Soc 111:9261–9262
16. Gouverneur VE, Houk KN, de Pascual-Teresa B, Beno B, Janda KD, Lerner RA (1993) Science 262:204–208
17. Blake JF, Lim D, Jorgensen WL (1994) J Org Chem 59:803–805
18. Yli-Kauhaluoma JT, Ashley JA, Lo C-H, Tucker L, Wolfe MM, Janda KD (1995) J Am Chem Soc 117:7041–7047
19. Kim SP, Leach AG, Houk KN (2002) J Org Chem 67:4250–4260
20. Hilvert D (2000) Annu Rev Biochem 69:751–793
21. Tremblay MR, Dickerson TJ, Janda KD (2001) Adv Synth Catal 343:577–585
22. Lutz S (2010) Curr Opin Biotechnol 21:734–743
23. Zanghellini A, Jiang L, Wollacott AM, Cheng G, Meiler J, Althoff EA, Röthlisberger D, Baker D (2006) Protein Sci 15:2785–2794
24. Kuhlman B, Baker D (2000) Proc Natl Acad Sci 97:10383–10388
25. Tantillo DJ, Jiangang C, Houk KN (1998) Curr Opin Chem Biol 2:743–750
26. Zhang X, DeChancie J, Gunaydin H, Chowdry AB, Clemente FR, Smith AJT, Handel TM, Houk KN (2008) J Org Chem 73:889–899
27. Jiang L, Althoff EA, Clemente FR, Doyle L, Röthlisberger D, Zanghellini A, Gallaher JL, Betker JL, Tanaka F, Barbas CF, Hilvert D, Houk KN, Stoddard BL, Baker D (2008) Science 319:1387–1391
28. Röthlisberger D, Khersonsky O, Wollacott AM, Jiang L, DeChancie J, Betker J, Gallaher JL, Althoff EA, Zanghellini A, Dym O, Albeck S, Houk KN, Tawfik DS, Baker D (2008) Nature 453:190–195
29. Siegel JB, Zanghellini A, Lovick HM, Kiss G, Lambert AR, St.Clair JL, Gallaher JL, Hilvert D, Gelb MH, Stoddard BL, Houk KN, Michael FE, Baker D (2010) Science 329:309–313
30. Cooper S, Khatib F, Treuille A, Barbero J, Lee J, Beenen M, Leaver-Fay A, Baker D, Popovic Z, Players F (2010) Nature 466:756–760
31. Eiben CB, Siegel JB, Bale JB, Cooper S, Khatib F, Shen BW, Players F, Stoddard BL, Popovic Z, Baker D (2012) Nat Biotechnol 30:190–192
32. Ruscio JZ, Kohn JE, Ball KA, Head-Gordon T (2009) J Am Chem Soc 131:14111–14115
33. Lassila JK, Baker D, Herschlag D (2010) Proc Natl Acad Sci 107:4937–4942
34. Kiss G, Röthlisberger D, Baker D, Houk KN (2010) Protein Sci 19:1760–1773
35. Frushicheva MP, Cao J, Chu ZT, Warshel A (2010) Proc Nat Acad Sci 107:16869–16874
36. Frushicheva MP, Cao J, Warshel A (2011) Biochemistry 50:3849–3858
37. Linder M, Johansson AJ, Olsson TSG, Liebeschuetz J, Brinck T (2011) J Chem Inf Model 51:1906–1917
38. Lassila JK (2010) Curr Opin Chem Biol 14:676 – 682
39. Privett HK, Kiss G, Lee TM, Blomberg R, Chica RA, Thomas LM, Hilvert D, Houk KN, Mayo SL (2012) Proc Natl Acad Sci 109:3790–3795
40. Huang P-S, Ban Y-EA, Richter F, Andre I, Vernon R, Schief WR, Baker D (2011) PLoS One 6:e24109
41. Morin A, Kaufmann KW, Fortenberry C, Harp JM, Mizoue LS, Meiler J (2011) Protein Eng Des Selec 24:503–516
42. Bommarius AS, Blum JK, Abrahamson MJ (2011) Curr Opin Chem Biol 15:194–200
43. Linder M, Hermansson A, Liebeschuetz J, Brinck T (2011) J Mol Model 17:833–849
44. Schreiner PR (2003) Chem Soc Rev 32:289–296
45. Cheong PH-Y, Legault CY, Um JM, Celebi-Ölcüm N, Houk KN (2011) Chem Rev 111:5042–5137
46. Michaelis L, Menten ML (1913) Biochem Z 49:333–369
47. Cleland WW (1970) Enzymes 2:1–65
48. Dalziel K (1975) Enzymes 11:1–60
49. Eyring H (1935) J Chem Phys 3:107–115
50. Garcia-Viloca M, Gao J, Karplus M, Truhlar DG (2004) Science 303:186–195
51. Wolfenden R, Snider MJ (2001) Acc Chem Res 34:938–945
52. Lightstone FC, Bruice TC (1996) J Am Chem Soc 118:2595–2605
53. Bruice TC, Lightstone FC (1999) Acc Chem Res 32:127–136
54. Toro-Labbe A, Gutierrerez-Oliva S, Murray J, Politzer P (2007) Mol Phys 105:2619–2625
55. Labet V, Morell C, Grand A, Toro-Labbe A (2008) J Phys Chem A 112:11487–11494
56. Toro-Labbe A, Gutierrez-Oliva S, Murray J, Politzer P (2009) J Mol Model 15:707–710
57. Hendlich M (1998) Acta Crystallogr D54:1178–1182
58. Hendlich M, Bergner A, Günther J, Klebe G (2003) J Mol Biol 326:607–620
59. Schmitt S, Hendlich M, Klebe G (2001) Angew Chem Int Ed 40:3141–3144
60. Schmitt S, Kuhn D, Klebe G (2002) J Mol Biol 323:387–406
61. GOLD 5.0 (2010) http://www.ccdc.cam.ac.uk/products/life_sciences/gold/
62. Jones G, Willet P, Glen R (1995) J Mol Biol 245:43–53
63. Jones G, Willett P, Glen RC, Leach AR, Taylor R (1997) J Mol Biol 267:727–748
64. Nissink JWM, Murray C, Hartshorn M, Verdonk ML, Cole JC, Taylor R (2002) Proteins 49:457–471
65. Verdonk ML, Cole JC, Hartshorn MJ, Murray CW, Taylor RD (2003) Proteins 52:609–623
66. Case D, Darden T, III TC, Simmerling C, Wang J, Duke R, Luo R, Crowley M, Walker RC, Zhang W, Merz K, Wang B, Hayik

S, Roitberg A, Seabra G, Kolossvary I, Wong KF, Paesani F, Vanicek J, Wu X, Brozell S, Steinbrecher T, Gohlke H, Yang L, Tan C, Mongan J, Hornak V, Cui G, Mathews D, Seetin M, Sagui C, Babin V, Kollman P (2008) AMBER 10, University of California, San Francisco, CA
67. Wang JM, Cieplak P, Kollman PA (2000) J Comput Chem 21:1049–1074
68. Åqvist J, Medina C, Samuelsson J (1994) Protein Eng 7:385–391
69. Åqvist J, Luzhkov VB, Brandsdal BO (2002) Acc Chem Res 35:358–365
70. Hansson T, Marelius J, Åqvist J (1998) J Comput Aided Mol Des 12:27–35
71. Brandsdal BO, Österberg F, Almlöf M, Feierberg I, Luzhkov VB, Åqvist J (2003) Free energy calculations and ligand binding. In: Daggett V (ed) Protein simulations, vol 66. Academic Press, London, pp 123–158
72. Almlöf M, Brandsdal BO, Åqvist J (2004) J Comput Chem 25:1242–1254
73. Carlson HA, Jorgensen WL (1995) J Phys Chem 99:10667–10673
74. Jones-Hertzog DK, Jorgensen WL (1997) J Med Chem 40:1539–1549
75. Lamb ML, Tirado-Rives J, Jorgensen WL (1999) Bioorg Med Chem 7:851–860
76. Linder M, Ranganathan A, Brinck T (2012) (Submitted)
77. Siegbahn PEM, Himo F (2009) J Biol Inorg Chem 14:643–651
78. Siegbahn PE, Himo F (2011) Wiley Interdiscip Rev Comp Mol Sci 1:323–336
79. Zhao Y, Schultz NE, Truhlar DG (2005) J Chem Phys 123:161103
80. Zhao Y, Schultz NE, Truhlar DG (2006) J Chem Theory Comput 2:364–382
81. Zhao Y, Truhlar DG (2006) J Chem Phys 125:194101
82. Zhao Y, Truhlar DG (2008) Theor Chem Acc 120:215–241
83. Pieniazek S, Clemente F, Houk K (2008) Angew Chem Int Ed 47:7746–7749
84. Linder M, Johansson AJ, Brinck T (2012) Org Lett 14:118–121
85. Linder M, Brinck T (2012) J Org Chem 77:6563–6573
86. Jaguar v. 7.6 (2009) http://www.schrodinger.com
87. Ditchfield R, Hehre WJ, Pople JA (1971) J Chem Phys 54:724–728
88. Hehre WJ, Radom L, Schleyer P, Pople JA (1986) Ab initio molecular orbital theory. Wiley, New York
89. Frisch MJ, Trucks GW, Schlegel HB, Scuseria GE, Robb MA, Cheeseman JR, Scalmani G, Barone V, Mennucci B, Petersson GA, Nakatsuji H, Caricato M, Li X, Hratchian HP, Izmaylov AF, Bloino J, Zheng G, Sonnenberg JL, Hada M, Ehara M, Toyota K, Fukuda R, Hasegawa J, Ishida M, Nakajima T, Honda Y, Kitao O, Nakai H, Vreven T, Montgomery JA, Jr.,; Peralta JE, Ogliaro F, Bearpark M, Heyd JJ, Brothers E, Kudin KN, Staroverov VN, Kobayashi R, Normand J, Raghavachari K, Rendell A, Burant JC, Iyengar SS, Tomasi J, Cossi M, Rega N, Millam JM, Klene M, Knox JE, Cross JB, Bakken V, Adamo C, Jaramillo J, Gomperts R, Stratmann RE, Yazyev O, Austin AJ, Cammi R, Pomelli C, Ochterski JW, Martin RL, Morokuma K, Zakrzewski VG, Voth GA, Salvador P, Dannenberg JJ, Dapprich S, Daniels AD, Farkas O, Foresman JB, Ortiz JV, Cioslowski J, Fox DJ (2009) Gaussian 09, revision A.02
90. Linder M, Johansson AJ, Manta B, Olsson P, Brinck T (2012) Chem Comm 48:5665–5667
91. Barone V, Cossi M (1998) J Phys Chem A 102:1995–2001
92. Cossi M, Rega N, Scalmani G, Barone V (2003) J Comput Chem 24:669–681
93. Kong S, Evanseck J (2000) J Am Chem Soc 122:10418–10427
94. Loncharich RJ, Schwartz TR, Houk KN (1987) J Am Chem Soc 109:14–23
95. Marenich AV, Cramer CJ, Truhlar DG (2009) J Phys Chem B 113:6378–6396
96. Simón L, Goodman JM (2010) J Org Chem 75:1831–1840
97. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE (2000) Nucleic Acids Res 28:235–242
98. Liu P, Wang Y-F, Ewis HE, Abdelal AT, Lu C-D, Harrison RW, Weber IT (2004) J. Mol. Biol. 342:551–561
99. Murzin AG, Brenner SE, Hubbard T, Chothia C (1995) J Mol Biol 247:536–540
100. Lauble H, Miehlich B, Förster S, Kobler C, Wajant H, Effenberger F (2002) Protein Sci 11:65–71
101. Jochens H, Hesseler M, Stiba K, Padhi SK, Kazlauskas RJ, Bornscheuer UT (2011) ChemBioChem 12:1508–1517
102. Montoro-García S, Martínez-Martínez I, Navarro-Fernández J, Takami H, García-Carmona F, Sánchez-Ferrer Á (2009) J Bacteriol 191:3076–3085
103. Branneby C, Carlqvist P, Magnusson A, Hult K, Brinck T, Berglund P (2003) J Am Chem Soc 125:874–875
104. Carlqvist P, Svedendahl M, Branneby C, Hult K, Brinck T, Berglund P (2005) ChemBioChem 6:331–336
105. Svedendahl M, Hult K, Berglund P (2005) J Am Chem Soc 127:17988–17989
106. Cannizzaro CE, Ashley JA, Janda KD, Houk KN (2003) J Am Chem Soc125:2489–2506
107. Bruno IJ, Cole JC, Edgington PR, Kessler M, Macrae CF, McCabe P, Pearson J, Taylor R (2002) Acta Cryst B 58:389–3397
108. Allen FH (2002) Acta Cryst B 58:380–388
109. Allen FH, Motherwell WDS (2002) Acta Cryst B 58:407–422
110. http://www.ccdc.cam.ac.uk/products/csd/
111. Cannon WR, Benkovic SJ (1998) J Biol Chem 273:26257–26260
112. Zhang X, Houk KN (2005) Acc Chem Res 38:379–385
113. Arnold FH (2001) Nature 409:253–257