Supporting Online Material for

De Novo Computational Design of Retro-Aldol Enzymes

Lin Jiang, Eric A. Althoff, Fernando R. Clemente, Lindsey Doyle,
Daniela Röthlisberger, Alexandre Zanghellini, Jasmine L. Gallaher, Jamie L. Betker,
Fujie Tanaka, Carlos F. Barbas III, Donald Hilvert, Kendall N. Houk, Barry L. Stoddard,
David Baker*

*To whom correspondence should be addressed. E-mail: dabaker@u.washington.edu

Published 7 March 2008, *Science* **319**, 1387 (2008)

DOI: 10.1126/science.1152692

This PDF file includes:

Other Supporting Online Material for this manuscript includes the following:
(available at www.sciencemag.org/cgi/content/full/319/5868/1387/DC1)

Design Model Coordinates in PDB Format

Supporting Material for
De novo computational design of retro-aldol enzymes

Lin Jiang*, Eric A. Althoff *, Fernando R. Clemente, Lindsey Doyle,

Daniela Röthlisberger, Alexandre Zanghellini, Jasmine L. Gallaher, Jamie L. Betker,

Fujie Tanaka, Carlos F. Barbas III,

Donald Hilvert, Kendall N. Houk, Barry Stoddard and David Baker[†]

* both authors contributed equally to this work

[†] To whom correspondence should be addressed. E-mail: dabaker@u.washington.edu

Table of contents

I. Kinetic parameters for designed retro-aldolases

Table S1. Rate enhancements for active designs

Design name	PDB code of scaffold protein	Catalytic lysine position	Number of amino acid changes	Active site motif	Additional mutations	Rate (min^{-1})* at 270 μM racemic substrate	Rate enhancement of lysine knockout mutation
Uncatalyzed rate						3.9×10^{-7}	
Jelly rolls							
Wild-type	**1m4w**					4.5×10^{-7}	
RA61	1m4w	176	8	IV		6.3×10^{-3}	1.6
RA60	1m4w	48	12	IV		3.4×10^{-3}	2
RA59	1m4w	176	9	IV		1.4×10^{-3}	n.d.
RA28	1f5j	180	6	III		1.3×10^{-4}	1
TIM barrels							
RA17	1thf	126	14	III		2.1×10^{-5}	4
RA58	1thf	169	10	IV		9.0×10^{-6}	1
RA31	1i4n	179	14	III		8.1×10^{-5}	1
RA32	1i4n	177	13	III		5.1×10^{-5}	110
RA33	1i4n	177	15	III		4.0×10^{-5}	120
Wild-type	**1lbf/1a53**					5.1×10^{-7}	
RA22	1lbf	159	13	III		$_b 1.3 \times 10^{-3}$ $_s 2.2 \times 10^{-4}$	2.5
					S210A	6.5×10^{-5}	n.d.
RA34	1lbf	159	13	IV		$_b 1.5 \times 10^{-3}$ $_s 1.9 \times 10^{-4}$	n.d.
					Y51L	3.2×10^{-4}	n.d.
RA35	1lbf	159	14	IV		5.1×10^{-4}	n.d.
RA36	1lbf	159	17	IV		1.6×10^{-4}	n.d.
RA47	1lbf	159	15	IV		7.2×10^{-4}	n.d.
RA39	1lbf	178	14	IV		7.7×10^{-5}	1
RA41	1lbf	178	17	IV		2.1×10^{-5}	n.d.
RA63	1igs	131	11	IV		6.2×10^{-4}	n.d.
RA6	1lbl	178	15	III		9.0×10^{-4}	1
					T159V	5.6×10^{-5}	n.d.
RA42	1lbl	178	15	IV		5.5×10^{-5}	n.d.
RA44	1lbl	178	17	IV		9.5×10^{-5}	n.d.
RA45	1lbl	180	13	IV		9.3×10^{-4}	2.4
					E233T	4.3×10^{-4}	n.d.
RA46	1lbl	180	14	IV		3.3×10^{-4}	3
RA48	1lbl	131	16	IV		7.0×10^{-4}	1
RA55	1lbl	159	12	IV		5.8×10^{-4}	n.d.
RA56	1lbl	159	16	IV		7.2×10^{-5}	n.d.
RA57	1lbl	159	13	III		1.5×10^{-3}	n.d.
RA49	1lbl	178	17	III		1.9×10^{-4}	n.d.
RA26	1a53	131	10	III		3.2×10^{-5}	1
RA40	1a53	178	17	IV		3.5×10^{-5}	n.d.
RA43	1a53	178	16	IV		3.8×10^{-5}	n.d.
RA53	1a53	159	16	IV		9.0×10^{-5}	n.d.
RA68	1a53	53	10	IV		5.4×10^{-4}	1

* Fluorescence measurements are converted to concentration of product which is divided by the enzyme concentration

Table S2. Enaminone formation rates with 2,4-pentanedione

Design Name	Additional Mutations	$T_{1/2}$ (h)
Jelly rolls		
RA61		1.5
RA60		1
RA59		3
RA28		8
TIM barrels		
RA17		10
RA58		2.5
RA31		3
RA32		6
RA33		6
RA22		10
	S210A	8
RA34		2
	Y51V	1.5
RA35		3
RA36		12
RA47		none
RA39		1
RA41		1
RA63		2
RA6		6
	T159V	12
RA42		none
RA44		6
RA45		2
	E233T	none
RA46		1
RA48		6
RA55		none
RA56		none
RA57		none
RA49		0.5
RA26		none
RA40		none
RA43		3
RA53		none
RA68		6
RA5*		4
RA54*		0.5

none = no 316nm absorbance increase due to enaminone formation was detected
* denotes designs which had no detectable aldolase activity

Other figures and tables appear in the section V.

II. Computational enzyme design methodology for multi-step reactions

Figure 1 in the main text provides an overview of the new computational enzyme design methodology for multi-step reactions. In the following sections, we describe the key steps in this protocol in detail.

Generation of idealized active site for each step in a reaction pathway

Given a schematic representation of a particular catalytic motif as in Figure 2C, the first step is to generate three-dimensional coordinates of the corresponding idealized active site. The chance of being able to perfectly recreate any single three dimensional arrangement of transition state and associated protein functional groups within a given scaffold set and subsequently achieve high affinity interactions with surrounding residues in the protein is small; therefore, it is desirable to generate as large and diverse a set of three dimensional realizations as possible for a given catalytic motif. We accomplish this by fine sampling along each of the degrees of freedom of the transition state-functional group complex. For a multi-step reaction such as the retro-aldol considered in this paper, it is necessary to ensure that the designed active site not only strongly stabilizes the highest energy transition state, but also interacts favorably with the reaction intermediates and other transition states along the reaction pathway (if it destabilizes any of these to the point that they become higher in energy than the original highest energy transition state, catalysis will be impaired). We thus must construct diversified structural ensembles representative of the key intermediates and transition states in the reaction.

The active site ensembles for catalytic motif I were created using quantum chemistry methods while the ensembles for catalytic motifs II-IV were created using a hybrid approach as described in the following paragraphs.

For catalytic site motif I, the geometry of the TS and the catalytic functional groups (two lysines and an aspartate) was fully optimized at the B3LYP/6-31G(d) level in the gas phase using Gaussian 03(*1*), as described previously for an aldol reaction(*2*). The 43 lowest energy carbon-carbon bond-breaking transition states, shown in Figure S1A, were used in the design calculations. For each of these 43 active sites, the other enantiomer of the TS was created by reflecting across the plane of the acetone moiety (enamine) and the resulting sites were also used in the design process.

Ensembles of transition state models were generated for motifs II, III, and IV by sampling around the lowest energy QM optimized structures. The C-C bond adjacent to the breaking C-C bond is the most constrained. As the retro-aldol process proceeds the breaking C-C σ bond is being transformed into a C=C π bond in the acetone moiety. The torsion around the forming C=C bond is constrained to be close to $\pm$ 90º (between the breaking C-C bond and the C-N bond of the imine, see Figure S1B). With the two enantiomers at the chiral center and the two possibilities for the χ_1 dihedral, four structures were generated from the lowest energy C-C bond-breaking TS found in the QM calculations (Figure S1C).

These four base states were diversified by sampling the degrees of freedom of the TS as indicated in Figure S1B to generate an ensemble of 1296 models. For each degree of

freedom independently, the deviation from the optimal structure producing a 0.5-1.0 kcal increase in energy after allowing the remaining degrees of freedom to relax was determined. Ensembles were generated by exhaustively enumerating conformations in which each degree of freedom took on the ideal value or the ideal value ± the deviation identified as above. This procedure is not optimal in that it neglects energetic couplings between simultaneously perturbed degrees of freedom.

Generation of composite active site description

The retro-aldolase reaction pathway shown in Figure 2B has multiple intermediates and transition states. A superposition of models based on the imine for all of these states obtained in QM calculations is shown in Figure S2A. It is evident that the diversity in this set of models can be largely represented by just two structures: the carbinolamine intermediate and the transition state for the carbon-carbon bond-breaking step. The carbinolamine mimics four distinct steps along the reaction pathway: the initial lysine attack and the dehydration step as well as the associated proton shuttling steps and also the reverse reactions required for enzyme recycling shown in Figure 2B. For computational efficiency, we chose to represent the ensemble of structures in Figure S2A by a superposition of the carbinolamine and the C-C bond-breaking step (Figure S2B). In the superposition, the naphthyl moiety is very close in space and for simplicity we use the naphthyl placement in the C-C bond-breaking TS as its geometry and orientation are more restricted. We simplify further by building the carbinolamine alcohol and the shift in the terminal methyl position onto the C-C bond-breaking transition state to create the composite TS model (Figure S2B).

For motif IV, a water molecule is placed to coordinate the two alcohol groups on the carbinolamine, and then situated on the composite TS based on the TS-carbinolamine superposition. Two additional hydrogen bonding sidechains (a combination of Ser, Tyr, and basic residues) are then incoporated to hydrogen bond to and position the water molecule; the two OH groups on the composite TS and the two hydrogen bonding sidechains together tetrahedrally coordinate the water.

For all motifs, the conformations of the imine-forming Lys residue (Figure S3A) and the other catalytic sidechains were diversified as described in (Figure S3B-F). For example, motif II begins with the addition of the methyl and the alcohol to the C-C bond-breaking TS. The composite TS is then diversified using the degrees of freedom outlined in the Figure S1B giving 1,296 conformations. These are combined with the 2,268 rotamers and 108 functional group placements of Lys shown in Figure S3A. For the Tyr residue responsible for proton abstraction, the dihedral angles $\chi 1$, $\chi 2$ and $\chi 3$ are sampled at 60° intervals, producing 6 x 6 x 6 = 216 discrete orientations for each of the 54 tyrosine sidechain rotamers (Figure S3C). Finally the π-stacking residues (Phe, Tyr, Trp) are placed in 1296 distinct positions for each of 189 rotamers (Figure S3B). In total, the diversification of the composite TS, the orientation of the TS relative to the key sidechains, and the internal geometry of the sidechains results in a total of 1,296 x 2,268 x 108 x 216 x 54 x 1296 x 189 = 9.1 x 10^{17} (Table S3) possible three dimensional realizations of the motif II catalytic motif.

Selection of scaffold set and pocket identification

The selection criteria for the scaffolds used in the designs were 1) high-resolution crystal structure available, 2) expression in *Escherichia coli* (*E. coli*) possible, 3) stable protein,

4) contain a preexisting pocket, and 5) span a variety of protein folds. The list of input protein scaffolds used in this study is provided in Table S4. For each scaffold, 3-dimensional grids representing the structural space available for possible transition state localization was defined by a 10 x 10 x 10 Å grid (0.25 Å resolution) centered on the bound ligand in the PDB, or centered on the catalytic residues if using an *apo* structure, were generated using an in-house pymol plugin; an example is shown in Figure S4. This first grid allows very rapid pruning of TS placements falling outside the desired pocket. A second grid representing backbone location was also generated for each structure to allow very rapid lookup based rejection of sidechain rotamer and transition state placements overlapping with the scaffold backbone. For each scaffold, positions within 6 Å of the first grid and with a Cα-Cβ vector pointing toward this grid were considered as possible sites for catalytic residue placement by Rosetta Match. This list of positions was further pruned in some cases for specific residues, for example to ensure a highly buried catalytic lysine.

Identification of sites in scaffolds where composite active site can be accurately created

The RosettaMatch method(*3*) employs geometric hashing to identify positions in a set of input protein scaffolds, which support the construction of a specified constellation of catalytic residues. For each composite active site description (the superimposed transition states and the associated catalytic residues), candidate catalytic sites were generated in a scaffold set (Table S4). At each position in the active site pockets on each scaffold (see previous section), each rotamer of each catalytic sidechain was placed and the ensuing position of the composite TS if there were no clashes with the scaffold backbone was recorded in the six-dimensional (6D) hash. To maximize the number of solutions, large

sets of sidechain rotamers were generated as detailed in Figure S3A-G. Following the filling out of the hash table, which scales linearly with the number of scaffold positions and number of sidechain rotamers, the hash was examined for TS positions compatible with all catalytic constraints; such positions are termed "matches".

Many of the possibilities for sidechain and composite active site placement were pruned, or trimmed from the hash. We used both general and reaction specific pruning criteria. The former criteria for elimination include sidechain or transition state clashes with the backbone grid, or key atoms of the transition state (for the retro-aldol: the imine C, Cαs, Cβ, β-OH, the C6 on the naphthyl, and their associated protons) not fitting inside the grid representing the potential binding site on a scaffold. Reaction specific criteria include restriction to TS orientations with the imine forming lysine at the bottom of the pocket and allowing only extended conformations of the imine forming lysine, which enable extensive packing with other residues to lower the pKa and keep the lysine fixed in place. This pruning reduces the search space and results in only rotamers and composite active site placements that are viable being added to the hash.

Using a π-stacking residue as a catalytic group

To benefit from the binding energy provided by a π-stacking interaction with the naphthyl moiety of the substrate, we developed an additional matching technique which treated the aromatic sidechains (Trp, Phe, Tyr) as catalytic residues in order to obtain active sites which also incorporated a key interaction for good binding affinity (Figure S3B). We searched for both parallel-displaced and T-shaped π-π interactions. This addition to the composite active site matching process greatly slowed down the search

when conducted in parallel with the other catalytic residues due to the considerable increase in site complexity. In order to overcome the slowdown due to this diversity, we often did an initial RosettaMatch run with only the catalytic residues shown in Figure 2C. This was followed by a second matching stage for the sole purpose of incorporating the aromatic π-stacking residue into the previously matched doublets or triplets of residues for motifs II and III, respectively.

Optimization of transition state rigid body orientation and the conformations of the catalytic sidechains

The site identification method described in the previous section produces initial active site configurations that are often suboptimal due to slightly strained catalytic residue geometry and/or small clashes between the composite transition state and the protein backbone. To eliminate these imperfections and to ensure a near ideal geometry for the subsequent full site design calculations, the rigid body degrees of freedom of the composite transition state and the chi angles of the catalytic sidechains were subjected to quasi-Newton optimization using a force field consisting solely of repulsive interactions, a constraint function representing the relevant catalytic constraints (see next section) and a sidechain torsional potential (*4*). Attractive interactions were omitted during this optimization as the amino acid composition of the site at this stage is still polyalanine or polyglycine. During the catalytic site optimization, the internal torsions of the composite transition state were also allowed to vary within specified tolerances (Figure S3) in order to maximize the possibility of the simultaneous satisfaction of all the catalytic constraints without large clashes of the composite TS with the scaffold backbone.

Constraint function

The constraint function consists of a flat bottom well with a width equal to the deviations listed in Figure S3 with a harmonic penalty for violations greater than these values. Freely rotatable torsional parameters were constrained with a periodic potential. As indicated in the figures, all degrees of freedom varied in the RosettaMatch search are optimized during the minimization along with a subset of the degrees of freedom kept fixed during RosettaMatch. The force constant for the distance, angle, and torsional deviations are scaled so the penalties are equal for deviations of 0.15 Å, 10°, and 10°, respectively. Constraints on each catalytic residue were weighted to reflect their importance to catalysis, in particular the constraints on the catalytic lysine was weighted 10x higher than those on the other residues.

Design of surrounding binding pocket.

We carried out full sequence optimization of the remaining (not included in the minimal catalytic site description) positions surrounding the docked composite TS model. We use the standard RosettaDesign Monte Carlo optimization using an extended version of the Dunbrack rotamer library (*4*) as carried out in the design of TOP7 (*5*), with protein-ligand interactions modeled as described previously(*6*). The backbone atoms were fixed in space, and a subset of residues were redesigned (varying the amino acid identity and sidechain conformation) and a second subset repacked (keeping the amino acid identity but varying the sidechain conformation). The first subset includes amino acid positions whose Cα atom are within 6.0 Å of the TS model and positions with Cα atom between 6.0 Å and 8.0 Å and with Cα-Cβ vector pointing toward the TS model. The second subset includes positions with Cα atom within 10.0 Å of the TS model and positions,

whose Cα atom is between 10.0 Å and 12.0 Å and with Cα -Cβ vector points to the TS model. Other positions are fixed in sidechain space.

Energy based refinement of completed designed active site

The design calculations employ a discrete rotamer representation of the sidechains, and after design the full energy function can be used in the refinement of the site. Hence, we next carry out simultaneous quasi-Newton optimization of the composite TS rigid body orientation, the sidechain torsion angles, and in some cases, the torsion angles of the composite TS and the backbone torsion angles in the site using the complete Rosetta energy function supplemented with the catalytic constraint function(*7*). This step does not change the designed sequence, but is essential to the subsequent assessment of the catalytic efficacy of the design.

Dock-perturbation to generate further minimized structures

To optimize interactions (H-bonding or packing) that are still missing at the end of the design process, we next carried out small random perturbations to the TS rigid-body degrees of freedom (0.2 Å for translational degrees of freedom; 10˚ for rotational degrees of freedom) to explore slightly different rigid body arrangements. We minimize the catalytic sidechains to ensure they are still within the geometric constraints. After catalytic sidechain minimization, the remaining pocket is redesigned with the pre-perturbed design as the starting point. Another round of energy-based refinement of the active site provides slight variations on the initial design which may have optimized properties. Usually several iterative runs of small dock perturbation, pocket design and refinement were carried to improve hydrogen bonding and packing interactions.

Ranking of designs

The resulting designs were filtered based on the following criteria:

1) Designs with a transition state-protein van der Waals attractive energy > -5.0 kcal/mol were removed.

2) Designs with fewer than 35 Cβ atoms within 10Å of the TS were eliminated (too exposed) as were designs with >85 Cβ atoms (too buried).

3) Designs in which the solvent accessible surface (SASA) of the TS was less than 10Å^2 were eliminated (no access to binding pocket).

The remaining designs were then ranked according to each of the following metrics independently:

1) Energy of binding of composite TS. Designs were ranked not only on the total binding energy, but also on each of the components separately (Lennard-Jones interactions, solvation, hydrogen bonding, and electrostatics)(*5, 6, 8*).

2) Catalytic geometry satisfaction as assessed with the harmonic constraint function described above.

3) Packing of the catalytic lysine sidechain assessed through its Lennard-Jones and solvation energy.

4) Atomic packing around transition state as assessed with a new algorithm for detecting packing imperfections and cavities in proteins (Sheffler, W. and Baker, D., submitted).

Designs which ranked in the top 40% according to each of these measures were further screened as follows:

1) The consistency of side chain conformations after repacking in the presence and absence of the TS model was determined, and designs in which there were large differences were discarded (we seek designs with a largely preformed active site).

2) The total folding free energy of the designed enzyme was compared to that of the original scaffold and designs predicted to be largely destabilized were eliminated.

Evaluation of top-ranking designs

The top 100 or so designs remained after these filtering and ranking steps. And two calculations are performed on those designs: the substrate rigid-body dock perturbation experiments (*6*) in order to check whether the substrate can fit into the pocket in the orientation appropriate for catalysis, the DDG calculation (*9*) which inspects the free energetic change upon the protein-transition state binding. And then designed satisfying the substrate docking and the DDG calculation were inspected visually to confirm routes for substrate entrance and product release, and the enzymatic activity of selected designs was experimentally characterized.

An overview of the active site search and design results is shown in Table S5. In a typical active site search using catalytic motif III, a total of 181,555 matches were found in the 71 different scaffolds matched against. Following optimization of the TS rigid body orientation and the identities and conformations of the surrounding residues, 343 of these had low TS binding energy and satisfied the catalytic constraints. Interestingly, a large fraction of the low energy matches were found in TIM barrel and jelly roll folds (Table S5).

III. Experimental Characterization

Sequences of active designs

A total of 72 designs with 10-20 amino acid identity changes in 10 different scaffolds were selected for experimental characterization. Soluble purified protein was obtained for

70 of 72 of the expressed designs. An overview of the rate enhancement of 70 designs is provided in Table S6. The sequences of enzymatically active designs, whose rate enhancement are provided in the Table S1, are shown in Table S7.

Experimental methods

Genes were synthesized by Codon Devices (Cambridge, MA) and inserted between *NdeI* and *XhoI* in pET29b+ (Novagen, Madison, WI) with a C-terminal Gly-Ser as a spacer between the protein and the *XhoI*-6xHis tag. Proteins were expressed in an auto-induction media(*10*). The cells were spun down and sonicated to expose soluble protein. The proteins were purified over 10 mL of Ni-NTA His•Bind Resin (Qiagen, Valencia, CA). After elution, the proteins were dialyzed three times against 100-fold larger volume into 25 mM HEPES, 100 mM NaCl, pH 7.5. The purity of the proteins was monitored by Coomassie staining which showed them to be >90% pure in all cases and generally significantly more pure than that especially for the designs fully characterized in the paper.

The protein concentrations were determined using the absorbance at 280 nm measured using an ND-1000 spectrophotometer (Thermo Fisher Scientific, Wilmington, DE). The molar extinction coefficient was calculated using the amino acid sequence- a tryptophan residue contributing 5500 cm^{-1} M^{-1} to the extinction coefficient, tyrosine contributing 1490 cm^{-1} M^{-1} and cysteine 125 cm^{-1} M^{-1}(*11*). Additionally, amino acid analysis was done on representative members of each scaffold, which confirmed the extinction coeffecient calculations for the protein concentration.

Site-directed mutagenesis was performed following the Kunkel protocol(*12, 13*), using oligonucleotides designed by a mutagenesis primer design algorithm (http://www.stratagene.com/tradeshows/feature.aspx?fpId=118) (*14*).

Retro-aldol reactions of 4-hydroxy-4-(6-methoxy-2-naphthyl)-2-butanone were performed in 2.7% acetonitrile, 25 mM HEPES, 100 mM NaCl, pH 7.5 at 27-29˚C. (180 μL buffer and protein with 5 μL small molecule in acetonitrile; normally, 6-16 μM protein and 270 μM substrate) and followed in a Spectramax M5^e (Molecular Devices, Sunnyvale, CA) in 96 Well Black Flat Bottom Polystyrene Non Binding Surface Microplates (Corning, Lowell, MA) using λ_{ex} of 330 nm (9 nm bandwidth) and λ_{em} of 452 nm (15 nm badwidth) with an additional filter at 435 nm for increased noise reduction. Quartz cuvette measurements were used to verify the plate measurements. The reactions were controlled for evaporation and were generally stable up to 4-5 hours with minimal evaporation-though most kinetic measurements were done within the first hour. Known concentrations of the product were used to quantitate the fluorescence (which is linear at low concentrations). HPLC was done to verify that the product of the reaction coeluted with the expected product. Using an Agilent 1200 Series HPLC monitoring at 254 nm, Agilent Zorbax ODS 5 μm 4.6 x 250 mm column, and eluting 0.5 mL/min 50% CH_3CN in H_2O, the product (6-methoxy-2-naphthaldehyde) has a retention time of 4.2 minutes while the starting material (4-hydroxy-4-(6-methoxy-2-naphthyl)-2-butanone) elutes at 8.6 minutes.

Product curves

Representative progress curves are shown in Figure S5. As noted in the main text, the progress curves associated with different designs can be quite different; many clearly do not follow the simple activated complex model used for Michaelis-Menten calculations.

RA60 and RA61 (Figure S5A and S5B) demonstrate nearly linear behavior for the length of the reaction. There is a slight lag or delay at the beginning for several minutes before they reach their maximal velocity. Both continue for multiple turnovers or until substrate is consumed.

RA45 and RA46 (Figure S5C and S5D) show a very distinct lag phase of many minutes where no product is formed- our hypothesis is that this represents a very slow imine formation rate. After this lag phase, the enzymes quickly attain their maximal velocity and continue on through multiple turnovers.

RA22 and RA34 (Figure S5E and S5F) display even more complex behavior. They begin with a similar short lag phase and take off similarly to RA60 and RA61, but after several minutes, slow down (almost like one would expect for burst kinetics); however, the enzymes slow down well before one equivalent of aldehyde product has formed. The difference between the two rates is ~7-fold lower for RA22 and ~8-fold lower for RA34. The rate after the slowdown is continued indefinitely; we report both the burst phase and the steady state rates in Table 2A. The amount of product formed during the burst phase is dependent on the substrate concentration as seen in the figures and is consistent

between independently purified preparations of designed enzyme. In experiments to examine the cause of the burst and steady state rate differences, pre-incubation of the enzymes (RA22 and RA34) with an excess of 6-methoxy-2-naphthaldehyde, the product, and subsequent addition of substrate resulted in only the steady state rate being observed. Pre-incubation with acetone had no similar effect. The interaction between the product and RA22 and RA34 results in a significant decrease in the fluorescent signal, which at least partially explains the observed non-stoichiometry of the burst phase. This also indicates that the true enzymatic rate is even faster than the burst phase, but is inhibited by binding of the product.

1a53 and 1m4w (Figure S4G), which are the original unmutated scaffold proteins used for many of the best designs, show no retro-aldolase activity. The product slope actually appears to be slightly downward and a much longer calculation is required to get a steady state rate which is calculated to be on the same order as the uncatalyzed rate.

Enaminone formation rates with 2,4-pentanedione

Enaminone formation was analyzed in 540 μM 2,4-pentanedione, 2.7% acetonitrile, 25 mM HEPES, 100 mM NaCl, pH 7.5. (180 μL buffer and protein with 5 μL small molecule in acetonitrile). Changes in absorbance at 316 nm were observed in 96 well plates either from a single reaction run (if fast enough) or by taking hourly point measurements. After the absorbance was saturatated, the total absorbance change was determined and the time at half ($T_{1/2}$) of the absorbance increase was interpolated and displayed in Table S1. Using the extinction coefficient of 15,000 M^{-1} cm^{-1} for the

enaminone(*15*), the active site lysine concentration appears to be less by 3-fold than the enzyme concentration we estimate based on tryptophan absorbance for RA60 and RA61; if the former is a more correct measure our reported k_{cat} values would be underestimates by 3-fold.

Uncatalyzed rate calculation

See Figure S6.

Crystallographic Methods.

RA22 S210A and RA61 M48K were both overexpressed from *E. coli* strain BL21 (RIL) as C-terminal and N-terminal his-tagged protein constructs, respectively, and purified via affinity chromatography against Talon metal-chelate resin (Clontech). The construct of RA22 S210A that was crystallized was expressed using the pET-29b vector (Novagen) and contained a non-cleavable C-terminal 6XHis-tag after the translated protein sequence of the designed enzyme. In contrast, the RA61 M48K construct that was crystallized was expressed using the pET-15b vector (Novagen), allowing removal of the N-terminal his-tag after purification (resulting in three exogenous amino acid residues, G-S-H, prior to the sequence of the designed enzyme).

Purified proteins were each concentrated to approximately 2 mg/mL by centrifugation against low molecular weight cut-off sieves (Centricon) and then screened for crystallization conditions using a commercial sparse matrix screen (Nextal Classic Suite; Qiagen). Crystals of RA22 S210A were grown by equilibration of the protein against a

reservoir containing 0.2 M ammonium sulfate, 0.1 M sodium acetate pH 4.6 and 27% PEG4000, followed by further equilibration after the initial crystallization over a well containing 0.2 M ammonium sulfate, 0.1 M sodium acetate pH4.6 and 30% PEG3350. The crystals were frozen in an artificial mother liquor matching the final crystallization reservoir, augmented with 20% glycerol v/v. Crystals of RA61 M48K were grown by equilibration of the protein against reservoirs containing 1.9 to 2.1 M ammonium sulfate, 0.1 M MES pH 6.5 and 5% v/v PEG 400. The crystals were frozen in an artificial mother liquor matching the final crystallization reservoir, augmented with 25% sucrose w/v.

X-ray data were collected on an R-AXIS IV/++ area detector (Rigaku) using X-rays corresponding to λ = 1.54 Å provided from the Cu-Ka line of a Micromax 007 rotating anode generator (Rigaku). A complete redundant data set was collected for each specimen, to 2.2 Å and 1.8 Å resolution, respectively. Data were processed and scaled using the CrystalClear software package (Rigaku/Molecular Structure Corporation). During scaling and subsequent molecular replacement and refinement, the space groups for the diffraction data sets were determined to be $P2_12_12$ and $P2_12_12_1$ for RA22 S210A and RA61 M48K, respectively. Molecular replacement was performed using the original PDB files (1LBF and 1M4W for RA22 and RA61, respectively) from the RCSB data base, corresponding to the parental proteins prior to computational redesign, with all sidechains and extended peptide regions that were subjected to redesign deleted. All stages of molecular replacement, model building, and refinement were performed using programs from the CCP4 computational suite(*16*) (PHASER, COOT and REFMAC). Data, model building and refinement statistics are shown in Table S8.

Verification of Enzyme Activity and Characterization

Amino Acid analysis was performed at the Texas A&M Protein Chemistry Laboratory on samples of RA22, RA22 K159M, RA34, RA60, RA61, and RA61 K176M, which had been purified over a Ni-NTA column and dialyzed to pH 7.5, 25mM HEPES, and 100mM NaCl. The concentrations observed corresponded quite well with the expected concentrations determined by absorbance at 280nm.

Mass Spectrometry was also performed on RA22, RA22 K159M, RA34, RA45, RA46, RA60, RA60 K48M, RA61, and RA61 K176M. In the cases of RA22, RA22 K159M, RA34, RA45, and RA46, masses were determined that corresponded to within 6 Daltons of the expected average mass. In the cases of RA60, RA60 K48M, RA61, and RA61 K176M, a mass was obtained that corresponded to within 6 Daltons of the calculated protein mass with the formyl-methionine remaining. These proteins are on the same scaffold and it is not surprising they behave similarly.

An imidazole gradient was run to elute the proteins from the Ni-NTA column (Figure S7). A gradient from 20 mM to 150 mM imidazole at 1 mL/min over 15 minutes eluted the proteins. The elution was monitored at 280nm. 1mL fractions were collected and assayed for retro-aldolase activity. Fractions with significant 280nm absorbances were run in 16.5% Tris-Tricine/ Peptide gels (BIO-RAD) and stained with GelCode Blue Stain Reagent (PIERCE). Separation is not great; however, the peak in catalytic activity for all designs tested (RA22, RA34, RA45, RA60 and RA61) corresponds to the peak the peak at 280nm and to the most intense protein bands on the stained gel (only results for RA61

shown for brevity, but it is representative of the remainder). The major bands on the gel correspond to the approximate molecular weight expected for the designed enzymes as well.

Size exclusion chromatography using an ÄKTA FPLC eluting at 0.8 mL/min and loading 1.5 mL of a Ni-NTA purified enzyme sample (10-20 μM) on a Superdex 200 column, which had been equilibrated and eluted with 150 mM NaCl, 25 mM HEPES, pH 7.5, allowed collection of 1 mL fractions. Elution was monitored at 280nm. All fractions were assayed for retro-aldolase activity and fractions with significant 280nm absorbances were run on PAGE gels. The results are displayed in Figure S8. Once again, the retroaldolase activity corresponds to the major protein fractions, which show single bands on a gel. RA34 is an interesting case in that it has a non-protein based impurity which elutes with a significant absorbance at 280nm. The impurity has no catalytic activity, however.

IV. Comparison to catalytic antibodies and related discussion

Our results demonstrate that novel enzyme catalysts for non-natural reactions can be created using computational enzyme design. Our success with the retro-aldol reaction is notable because of the complexity and large number of steps in the reaction—the enzyme design methodology used here is immediately applicable to other multi-step reactions. While the designs are less active than aldolase catalytic antibodies(*17*), they should be excellent starting points for generating improved catalysts using directed evolution due to their relatively small size and the robustness of the scaffolds which should allow for

increased expression, easier purification and library synthesis, *etc*. The more constrained designed active sites are also likely to have different substrate selectivities: for example, the nucleophilic lysine in the catalytic antibody combines rapidly with both the retro-aldol substrate and the diketone probe despite their very different structures, whereas many of our designs with considerable retro-aldolase activity interact very slowly or not at all with the diketone (Table S2).

The success in computational design of enzyme catalysts for the retro-aldol reaction is due to not only the enzyme design methodology described, but also to availability of sufficiently powerful computers (an average of 30,000 cpu hours per active site motif) and the availability of low cost and rapid gene synthesis capability which made possible the experimental testing of many designs for each of four different enzyme active site types in a wide range of protein scaffolds.

It is tempting to speculate that our admittedly primitive computationally designed enzymes and primordial enzymes that arose early in evolution resemble one another more than they resemble highly refined and sophisticated modern day enzymes. Whether or not this analogy is correct in detail, we look forward to developing increasingly powerful aldol catalysts by improving on the robust and stable designs described here both by incorporating additional backbone flexibility into the design process, particularly in loop regions, to increase the reactivity of the imine forming lysine and to lower K_M by making possible tighter substrate and transition state binding, and by using directed evolution for more subtle fine tuning further from the active site.

V. References

Full authorship of Gaussian

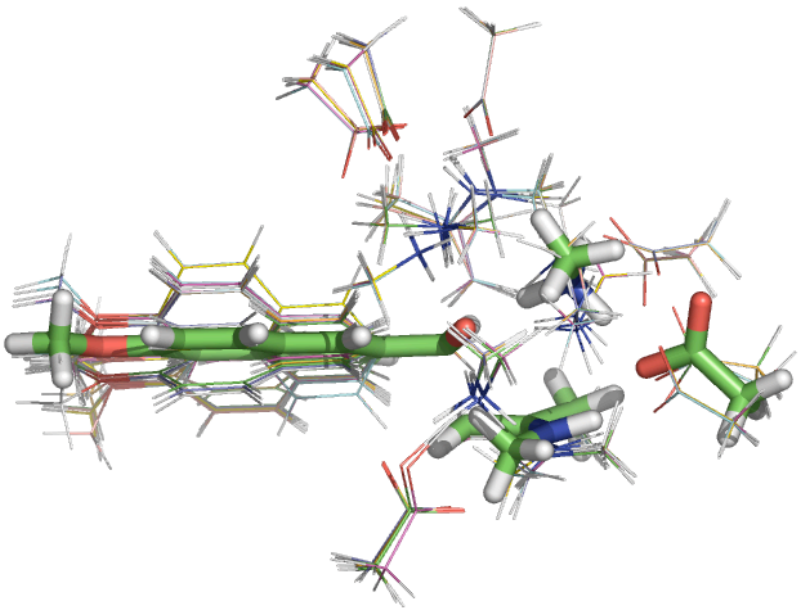
Gaussian 03, Revision C.02, M. J. Frisch, G. W. Trucks, H. B. Schlegel, G. E. Scuseria, M. A. Robb, J. R. Cheeseman, J. A. Montgomery, Jr., T. Vreven, K. N. Kudin, J. C. Burant, J. M. Millam, S. S. Iyengar, J. Tomasi, V. Barone, B. Mennucci, M. Cossi, G. Scalmani, N. Rega, G. A. Petersson, H. Nakatsuji, M. Hada, M. Ehara, K. Toyota, R. Fukuda, J. Hasegawa, M. Ishida, T. Nakajima, Y. Honda, O. Kitao, H. Nakai, M. Klene, X. Li, J. E. Knox, H. P. Hratchian, J. B. Cross, V. Bakken, C. Adamo, J. Jaramillo, R. Gomperts, R. E. Stratmann, O. Yazyev, A. J. Austin, R. Cammi, C. Pomelli, J. W. Ochterski, P. Y. Ayala, K. Morokuma, G. A. Voth, P. Salvador, J. J. Dannenberg, V. G. Zakrzewski, S. Dapprich, A. D. Daniels, M. C. Strain, O. Farkas, D. K. Malick, A. D. Rabuck, K. Raghavachari, J. B. Foresman, J. V. Ortiz, Q. Cui, A. G. Baboul, S. Clifford, J. Cioslowski, B. B. Stefanov, G. Liu, A. Liashenko, P. Piskorz, I. Komaromi, R. L. Martin, D. J. Fox, T. Keith, M. A. Al-Laham, C. Y. Peng, A. Nanayakkara, M. Challacombe, P. M. W. Gill, B. Johnson, W. Chen, M. W. Wong, C. Gonzalez, and J. A. Pople, *Gaussian, Inc.,* Wallingford CT, 2004.

1. M. J. Frisch *et al.*, Gaussian (2004).
2. F. R. Clemente, K. N. Houk, *J Am Chem Soc* **127**, 11294 (Aug 17, 2005).
3. A. Zanghellini *et al.*, *Protein Sci* **15**, 2785 (Dec, 2006).
4. R. L. Dunbrack, Jr., F. E. Cohen, *Protein Sci* **6**, 1661 (Aug, 1997).
5. B. Kuhlman *et al.*, *Science* **302**, 1364 (Nov 21, 2003).
6. J. Meiler, D. Baker, *Proteins* **65**, 538 (Nov 15, 2006).
7. W. H. Press, Teukolsky, Saul A., Vetterling, William T., Flannery, Brian P., *Numerical recipes in FORTRAN: The art of scientific computing*. (Cambridge University Press, 1992), pp.
8. G. Dantas, B. Kuhlman, D. Callender, M. Wong, D. Baker, *J Mol Biol* **332**, 449 (Sep 12, 2003).
9. T. Kortemme, D. Baker, *Proc Natl Acad Sci U S A* **99**, 14116 (Oct 29, 2002).
10. F. W. Studier, *Protein Expr Purif* **41**, 207 (May, 2005).
11. C. N. Pace, F. Vajdos, L. Fee, G. Grimsley, T. Gray, *Protein Sci* **4**, 2411 (Nov, 1995).
12. T. A. Kunkel, *Proc Natl Acad Sci U S A* **82**, 488 (Jan, 1985).
13. T. A. Kunkel, K. Bebenek, J. McClary, *Methods Enzymol* **204**, 125 (1991).
14. V. Z. A. Novoradovsky, M. Ghosh, H. Hogrefe, J.A. Sorge and T. Gaasterland, paper presented at the Technical Proceedings of the 2005 NSTI Nanotechnology Conference and Trade Show, Anaheim, 2005 2005.

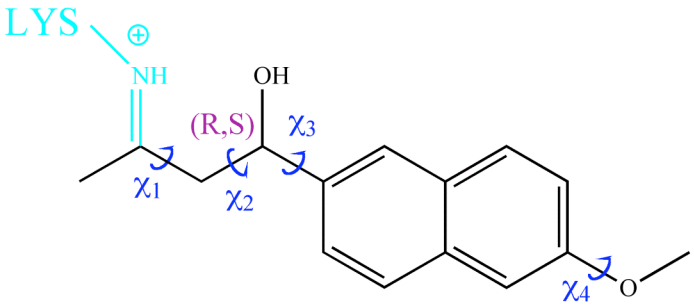
15. J. Wagner, R. A. Lerner, C. F. Barbas, 3rd, *Science* **270**, 1797 (1995).
16. *Acta Crystallogr D Biol Crystallogr* **50**, 760 (Sep 1, 1994).
17. G. Zhong *et al.*, *Angew Chem Int Ed Engl* **37**, 2481 (1998).
18. F. Tanaka, R. Fuller, C. F. Barbas, 3rd, *Biochemistry* **44**, 7583 (2005).
19. H. M. Berman *et al.*, *Acta Crystallogr D Biol Crystallogr* **58**, 899 (Jun, 2002).

VI. Figures and Tables

Figure S1. Transition state diversification



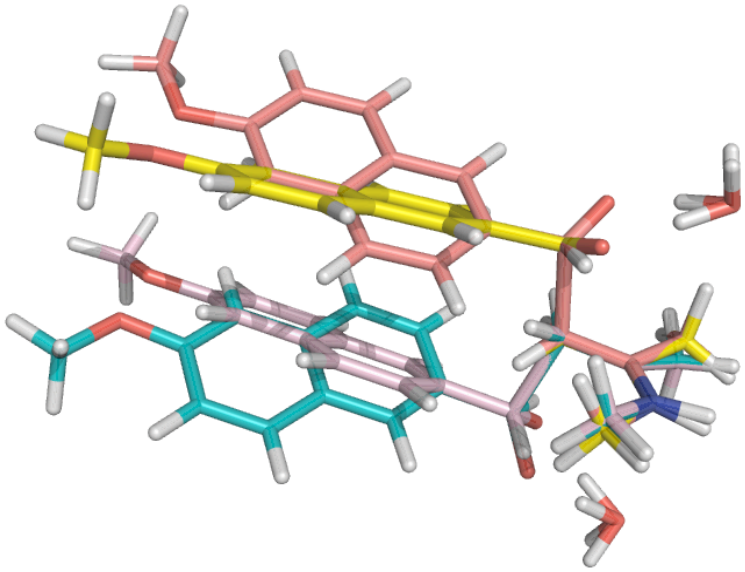
A. A graphical depiction of the active site ensemble for the carbon-carbon bond-breaking step in **motif I**. The lowest energy conformation appears as thicker sticks and the other conformations in the thinner lines. The xyz coordinates of the ensemble are provided in the online support. Only the terminal atoms of the catalytic sidechains are shown.



	optimal value[a]	deviation	# of conformations
χ_1	91°,-89°	+/-10°	6
χ_2	-177°	+/-10°	3
χ_3	109°,-71°	+/-10°	6
χ_4	0,180°	+/-20°	6
(R,S)-enantiomers			2
total			1,296

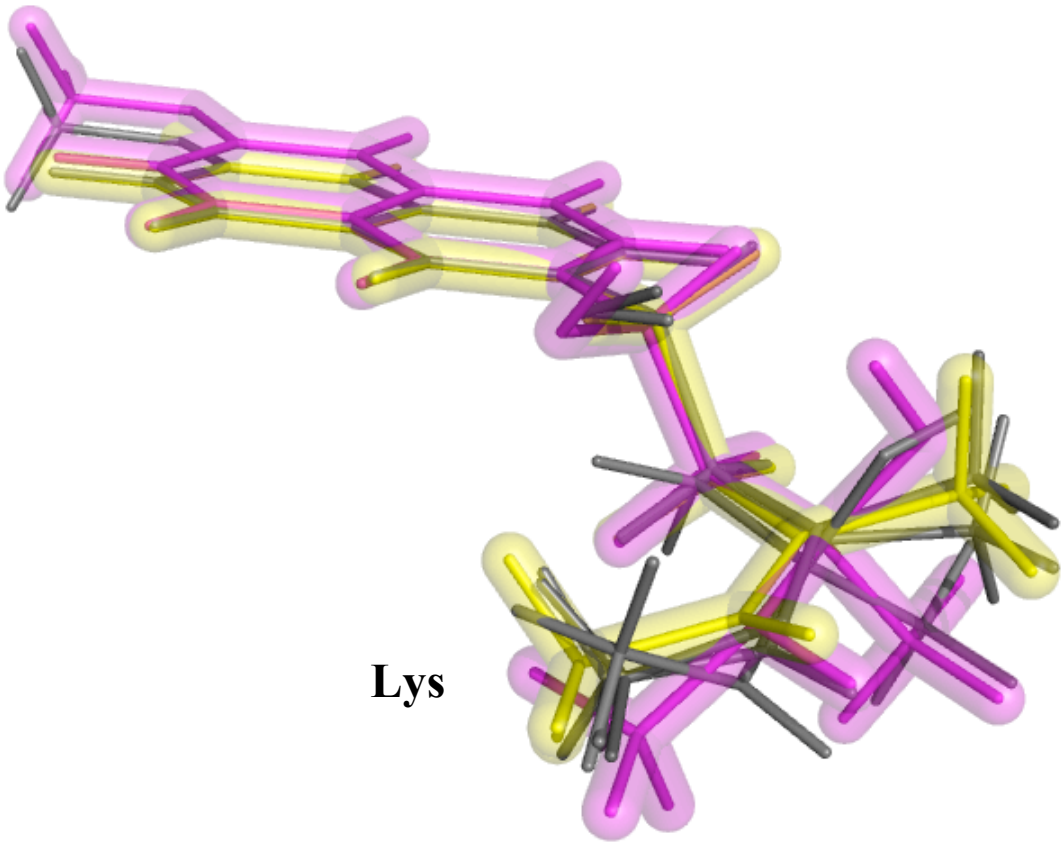
[a]. all torsion angles take their starting values from the QM lowest energy TS model of the R-enantiomer based on the QM caculation for C-C bond-breaking step

B. Diversification of the degrees of freedom of the C-C bond-breaking transition state for motifs II, III and IV.

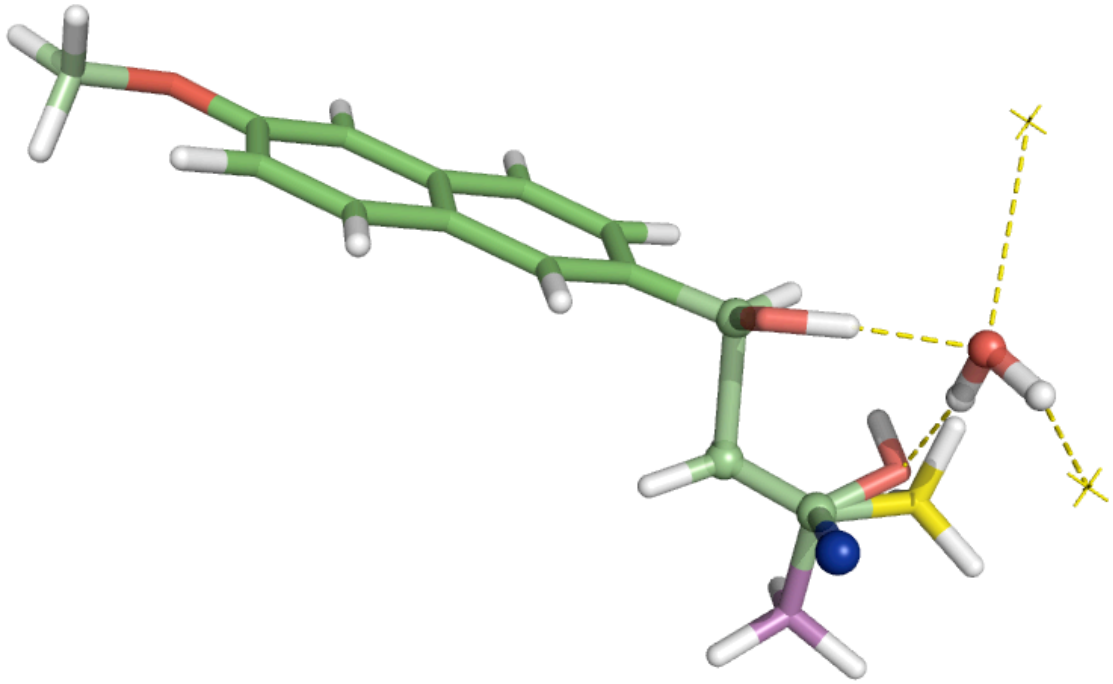


C. The four base transition states for the carbon-carbon bond-breaking step (R and S enantiomers; $\chi_1 = \pm 90°$) The Nζ (blue), Cε of the catalytic lysine, and the water in motif IV are shown.

Figure S2. Generation of the composite TS



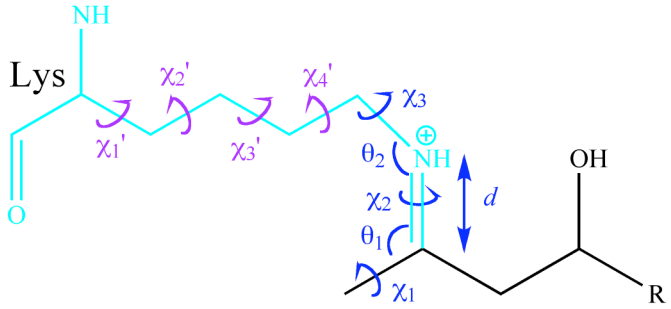
A. Superposition of the QM calculated transition states for lysine nucleophilic attack, carbinolamine intermediate, water elimination, imine intermediate, and C-C bond-breaking step. The reverse reactions have identical transition states, except for the naphthyl moiety now being absent. The TS models for carbinolamine intermediate (magenta) and C-C bond-breaking step (yellow) are highlighted. The carbinolamine intermediate is a reasonable approximation to the transition state for carbinolamine formation, intermediate, and water elimination steps; in combination with the C-C bond-breaking TS, the carbinolamine intermediate can cover most of the atomic motion along the pathway. Nζ and Cε of the catalytic lysine are also shown.



B. Composite transition state with water molecule for motif IV. The methyl group from the carbinolamine intermediate (purple) and carbon-carbon breaking step (yellow) are superimposed on the key atoms of the imine (Nζ of lysine and its Cα and Cβ atoms)- which are highlighted with the additional spheres. Nζ (blue) of the catalytic lysine is also shown. The yellow dashed lines indicate hydrogen bonds; the X's are positions of sidechain hydrogen bonding groups sought during matching.

Figure S3. Geometric parameters for catalytic residue placements used in motifs II, III and IV.

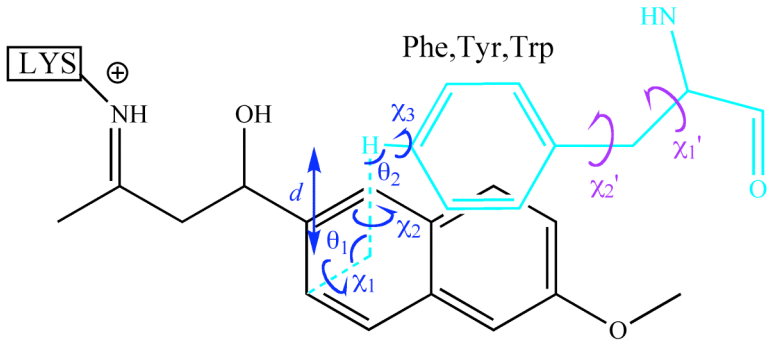
A. Geometric parameters for the imine-forming lysine



	ideal value	deviation	# of conformations
d	1.3Å	+/-0.2Å [a]	1
θ_1	125°	+/-5° [a]	1
θ_2	120°	+/-5° [a]	1
χ_1	180° [b]	+/-10°	3
χ_2	0°,180° [b]	+/-10°	6
χ_3	0° [c]	every 60°	6
χ_1'	64.3° [d]	+/-7.6°	
χ_2'	178.9° [d]	+/-8.1°	
χ_3'	177.5° [d]	+/-10.0°	
χ_4'	-179.2° [d]	+/-9.6°	
χ_1' - χ_4'			28x81[e]
all			244,944

[a]. this degree of freedom is not diversified during matching, but is allowed to vary during the subsequent optimization steps
[b]. planarity of the imine/enamine
[c]. freely rotatable bond
[d]. values for the highest frequency Lys rotamer from the Dunbrack backbone dependent library are shown for illustration; the highest frequency 28 rotamers were used in the matching calculations
[e]. for each of the 28 high frequency Lys rotamers from the Dunbrack library, 81 variants were generated with each chi angle perturbed by ± 1 standard deviation (3rd column) (3x3x3x3=81)

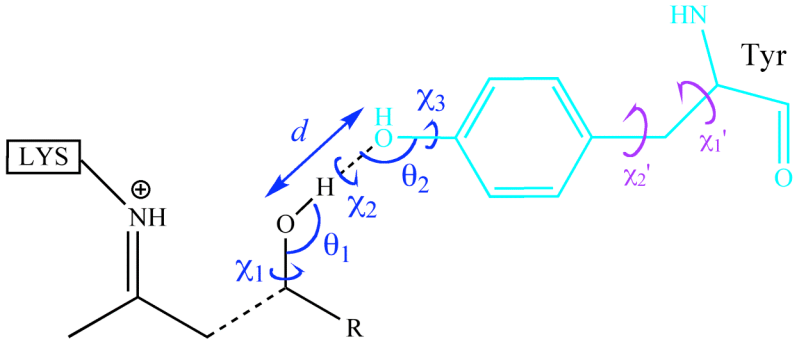
B. Geometric parameters for the π-stacking residue (Phe,Tyr,Trp) used in motif II and III.



	ideal value[a]	deviation	# of conformations
d	4.0Å	+0.6Å [b]	1
θ_1	90°	+/-20° [b]	1
θ_2	105°	+/-15°	3
χ_1	-90,90°	+/-20° [b]	2
χ_2		every 60°	6
χ_3	90°	+/-20°	3
symmetry about six-member ring			6 [c]
stacked on either ring center			2
Phe χ_1'	58.9° [d]	+/-8.1°	
Phe χ_2'	89.2° [d]	+/-9.2°	
Phe χ_1' - χ_2'			6x9
Tyr χ_1' - χ_2'			6x9
Trp χ_1' - χ_2'			9x9
all			244,944

[a]. idealized pi-stacking geometries (4)
[b]. this degree of freedom is not diversified during matching, but is allowed to vary during the subsequent optimization steps
[c]. the different choices for Cβ atom extension
[d]. values for a high frequency Phe rotamer are shown for illustration, a total of six (Phe,Tyr) or nine (Trp) rotamers are used including nine variants for each rotamer generated with each chi angle perturbed by ± 1 standard deviation (3rd column) (3x3)

C. Geometric parameters and diversification for placing the sidechain of Tyr in motif II

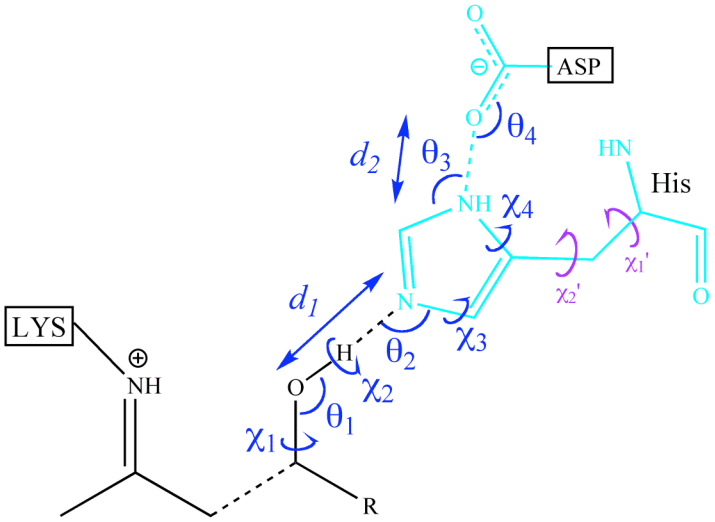


	ideal value	deviation	# of conformations
d	2.8Å	+/-0.2Å [a]	1
θ_1	110°	+/-10° [a]	1
θ_2	125°	+/-10° [a]	1
χ_1		every 60°	6
χ_2		every 60°	6
χ_3	0°,180°	+/-30°	6
χ_1'	53.5° [b]	+/-11.2°	
χ_2'	84.4° [b]	+/-12.0°	
χ_1' - χ_2'			6x9
all			11,664

[a]. this degree of freedom is not diversified during matching, but is allowed to vary during the subsequent optimization steps

[b]. values from a high frequency Tyr rotamer are shown for illustration, a total of six rotamers are used including nine variants for each rotamer generated with each chi angle perturbed by ± 1 standard deviation (3rd column) (3x3)

D. Geometric parameters and diversification for placing His-Asp dyad in motif III



His placement[a]

	ideal value	deviation	# conformations
d_1	2.8Å	+/-0.2Å [b]	1
θ_1	110°	+/-10° [b]	1
θ_2	125°	+/-10° [b]	1
χ_1		every 60°	6
χ_2		every 60°	6
χ_3	180° [c]	+/-20°	6
χ_1'	63.8° [d]	+/-7.0°	
χ_2'	-85.0° [d]	+/-14.7°	
χ_1' - χ_2'			9x9
all			17,496

Parameters for His-Asp dyad[a]

	ideal value	deviation
d_2	2.8Å	+/-0.2Å
θ_3	120°	+/-10°
θ_4	125°	+/-10°
χ_4	0° [c]	+/-20°

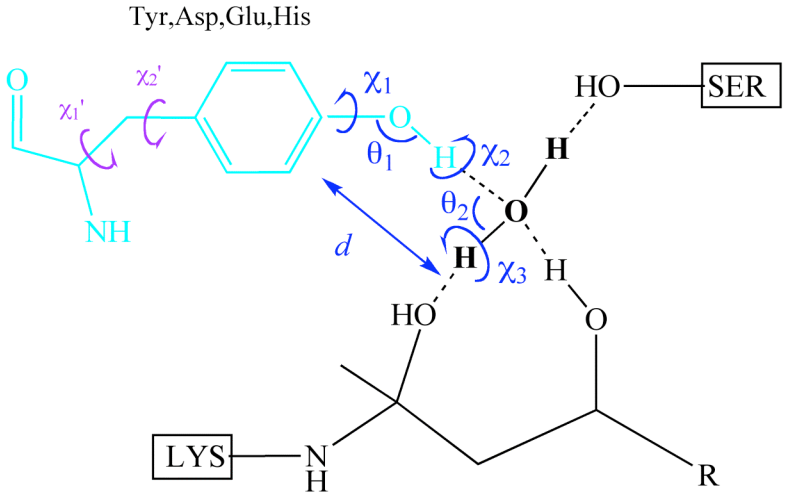
[a]. To place the His-Asp dyad, the placement of His rotamer is constrained by identifying, for each His rotamer in a scaffold, the set of Asp rotamers which can provide the supporting hydrogen bond

[b]. this degree of freedom is not diversified during matching, but is allowed to vary during the subsequent optimization steps

[c]. the value indicates the atoms are in the plane of the imidazole ring

[d]. values from a popular rotamer are shown for illustration, a total of nine rotamers are used including nine variants for each rotamer generated with each chi angle perturbed by ± 1 standard deviation (3rd column) (3x3)

E. Geometric parameters and diversification for placing basic residues (Tyr,Asp,Glu,His) in motif IV



	ideal value	deviation	# of conformations
d	2.8Å	+/-0.2Å [a]	1
θ_1	120°	+/-20° [a]	1
θ_2	110°	+/-10° [a]	1
χ_1	0°,180°	+/-30°	6
χ_2		every 60°	6
χ_3	90° [b]	+/-20°	3
Tyr χ_1'	53.5° [c]	+/-11.2°	
Tyr χ_2'	84.4° [c]	+/-12.0°	
Tyr $\chi_1' - \chi_2'$			6x9
Asp $\chi_1' - \chi_2'$ [d]			9x9
Glu $\chi_1' - \chi_3'$ [d]			21x27
His $\chi_1' - \chi_2'$ [d]			9x9
all			84,564

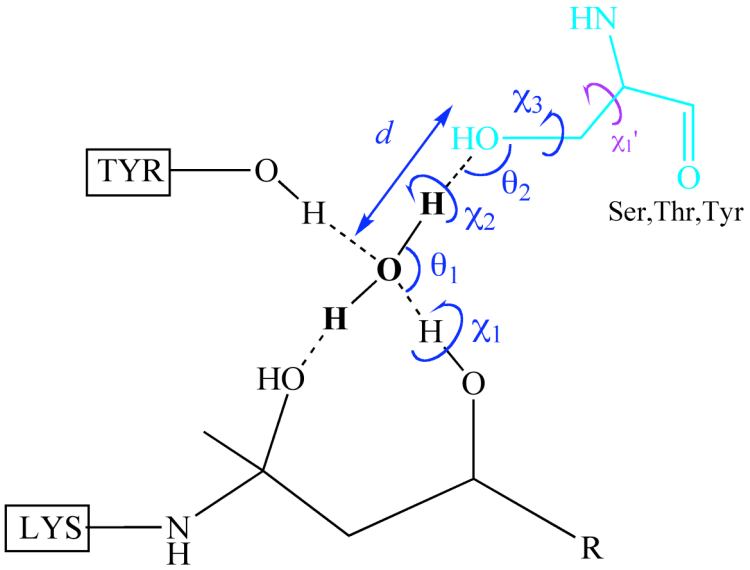
[a]. this degree of freedom is not diversified during matching, but is allowed to vary during the subsequent optimization steps

[b]. the value indicates tetrahedral geometry at the water molecule

[c]. values from a popular Tyr rotamer are shown for illustration, a total of six (Tyr), nine (Asp,His) or twenty-one (Glu) rotamers are used including nine (3x3;Tyr,Asp,His) or twenty seven (3^3;Glu) variants for each rotamer generated with each chi angle perturbed by ± 1 standard deviation (3^{rd} column)

[d]. Asp,Glu,His are placed by the geometric parameters following the simple rule for hydrogen bonding interactions described in our previous paper

F. Geometric parameters and diversification for placing an H-bonding group (Ser,Thr,Tyr) in motif IV



	ideal value	deviation	# of conformations
d	3.0Å	+/-0.2Å [a]	1
θ_1	110°	+/-10° [a]	1
θ_2	120°	+/-20° [a]	1
χ_1	-90° [b]	+/-20°	3
χ_2		every 60°	6
χ_3		every 60°	6
Ser χ_1'	69.6° [c]	+/-12.3°	
Ser χ_1'			3x3
Thr χ_1'			6x3
Tyr $\chi_1' - \chi_2'$			6x9
all			8,748

[a]. this degree of freedom is not diversified during matching, but is allowed to vary during the subsequent optimization steps

[b]. the value indicates tetrahedral geometry at the water molecule

[c]. values from a popular Ser rotamer are shown for illustration, a total of three (Ser) or six (Thr,Tyr) rotamers are used including three (Ser,Thr) or nine (Tyr) variants for each rotamer generated with each chi angle perturbed by ± 1 standard deviation (3rd column)

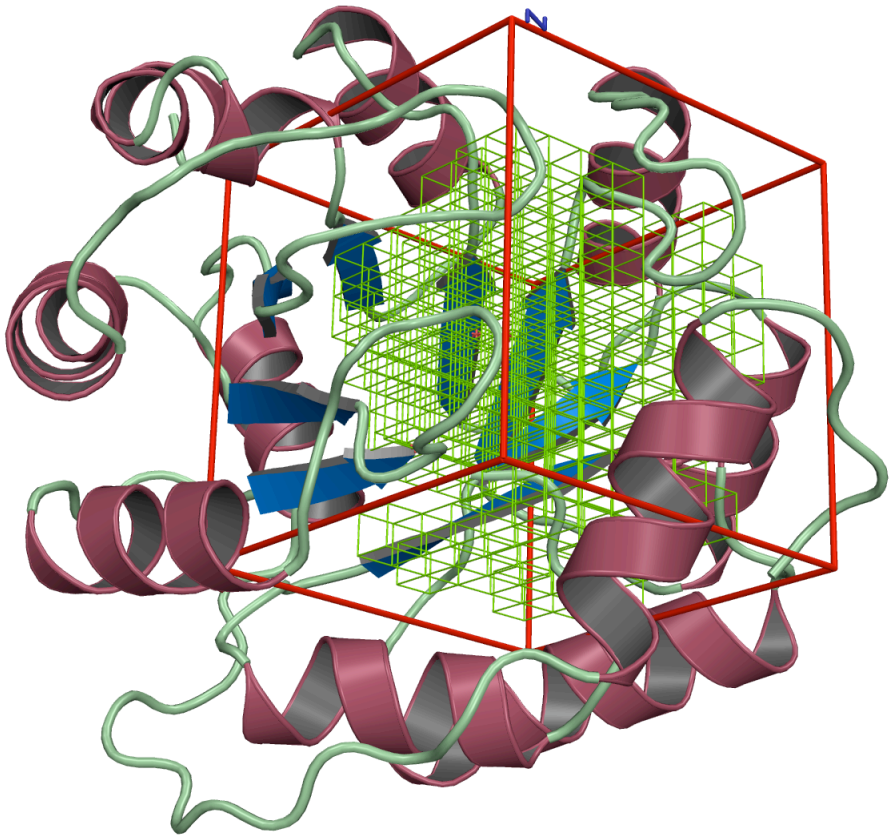
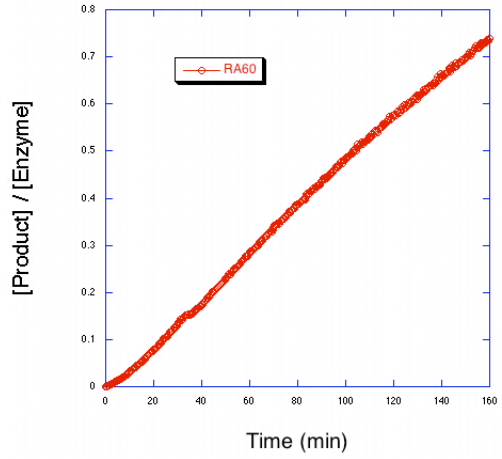
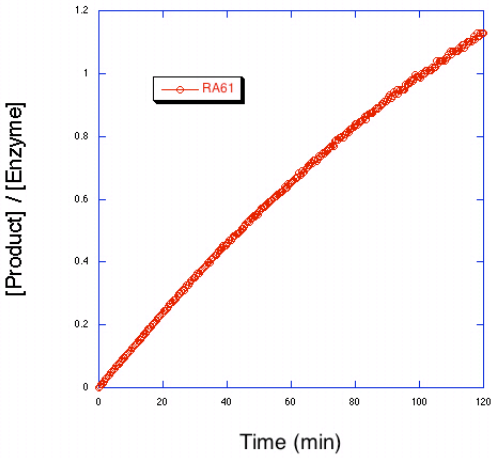


Figure S4. A pictoral representation of the grid used in RosettaMatch. A TIM barrel scaffold protein, 1a53, is shown as gray ribbons. The green grid is the three dimensional space allowed for localization of the substrate or composite active site. It has 0.25 Å resolution and is trimmed to remove clashes with the backbone and to encompass the pocket where enzyme active sites would generally occur within a given scaffold. A similar backbone grid, which is a precomputed three-dimensional representation of the backbone is used to quickly look up backbone clashes without the need to cycle through all atoms in the protein during the clash check.

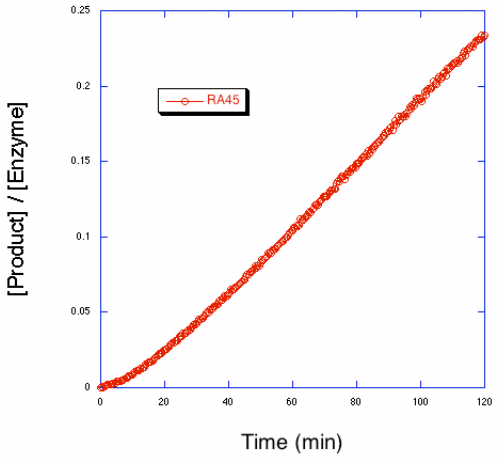
Figure 5. Progress curve for selected designs



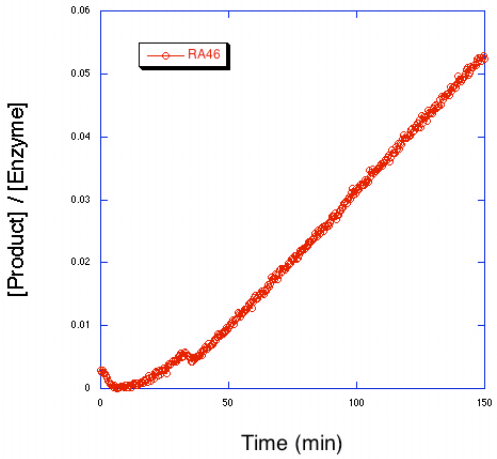
A. Progress curve for design RA60 incubated with 540 μM racemic substrate. RA60 was 8.6 μM in 25 mM HEPES, 100 mM NaCl, pH 7.5.



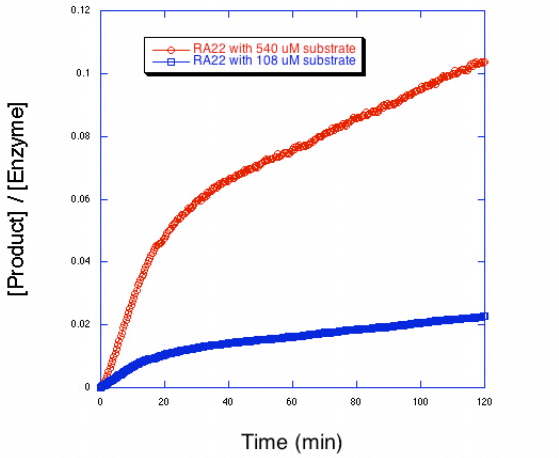
B. Progress curve for design RA61 incubated with 540 μM racemic substrate. RA61 was 13.8 μM in 25 mM HEPES, 100 mM NaCl, pH 7.5.



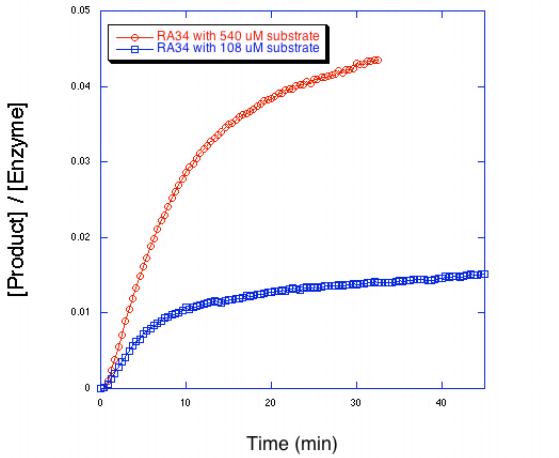
C. Progress curve for design RA45 incubated with 540 μM racemic substrate. RA45 was 18.0 μM in 25 mM HEPES, 100 mM NaCl, pH 7.5.



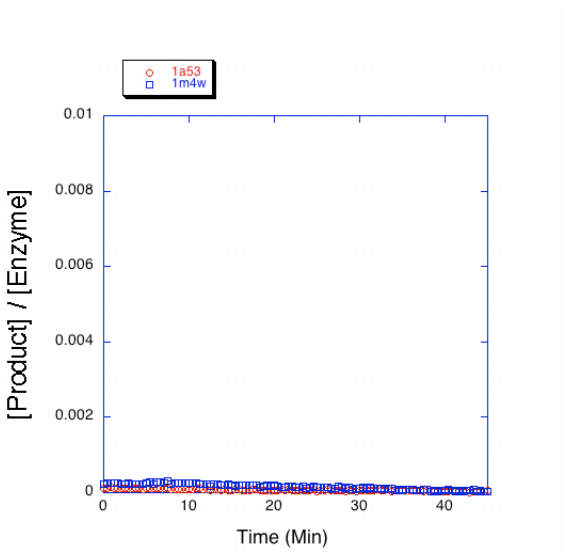
D. Progress curve for design RA46 incubated with 540 μM racemic substrate. RA46 was 11.0 μM in 25 mM HEPES, 100 mM NaCl, pH 7.5.



E. Progress curve for design RA22 incubated with 540 and 108 μM racemic substrate. RA22 was 16.0 μM in 25 mM HEPES, 100 mM NaCl, pH 7.5.



F. Progress curve for design RA34 incubated with 540 and 108 μM racemic substrate. RA34 was 12.3 μM in 25 mM HEPES, 100 mM NaCl, pH 7.5.



G. Progress curve for wild type scaffold proteins 1a53 and 1m4w incubated with 540 μM racemic substrate. 1a53 and 1m4w were 18.1 and 22.5 μM, respectively, in 25 mM HEPES, 100 mM NaCl, pH 7.5.

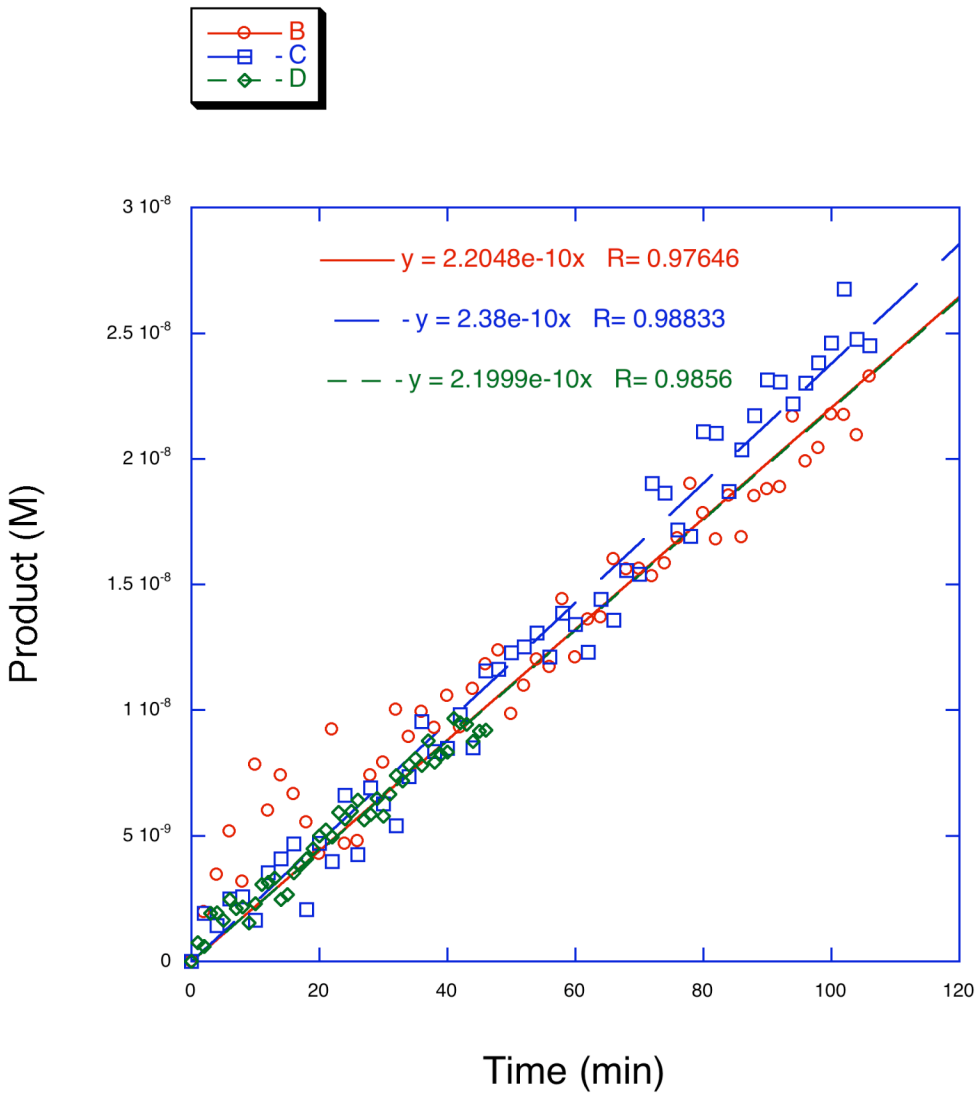


Figure S6. Uncatalyzed reaction for 560 μM substrate in 25 mM HEPES, 100 mM NaCl, pH 7.5, 2.7% acetonitrile. The uncatalyzed rate (the slope divided by the initial substrate concentration: 2.26 x10^{-10} M/min / 0.000560M) is thus 4.1 ± 0.2 x 10^{-7} min^{-1}. We use 3.9 x 10^{-7} min^{-1} to be consistent with previous literature(*18*).

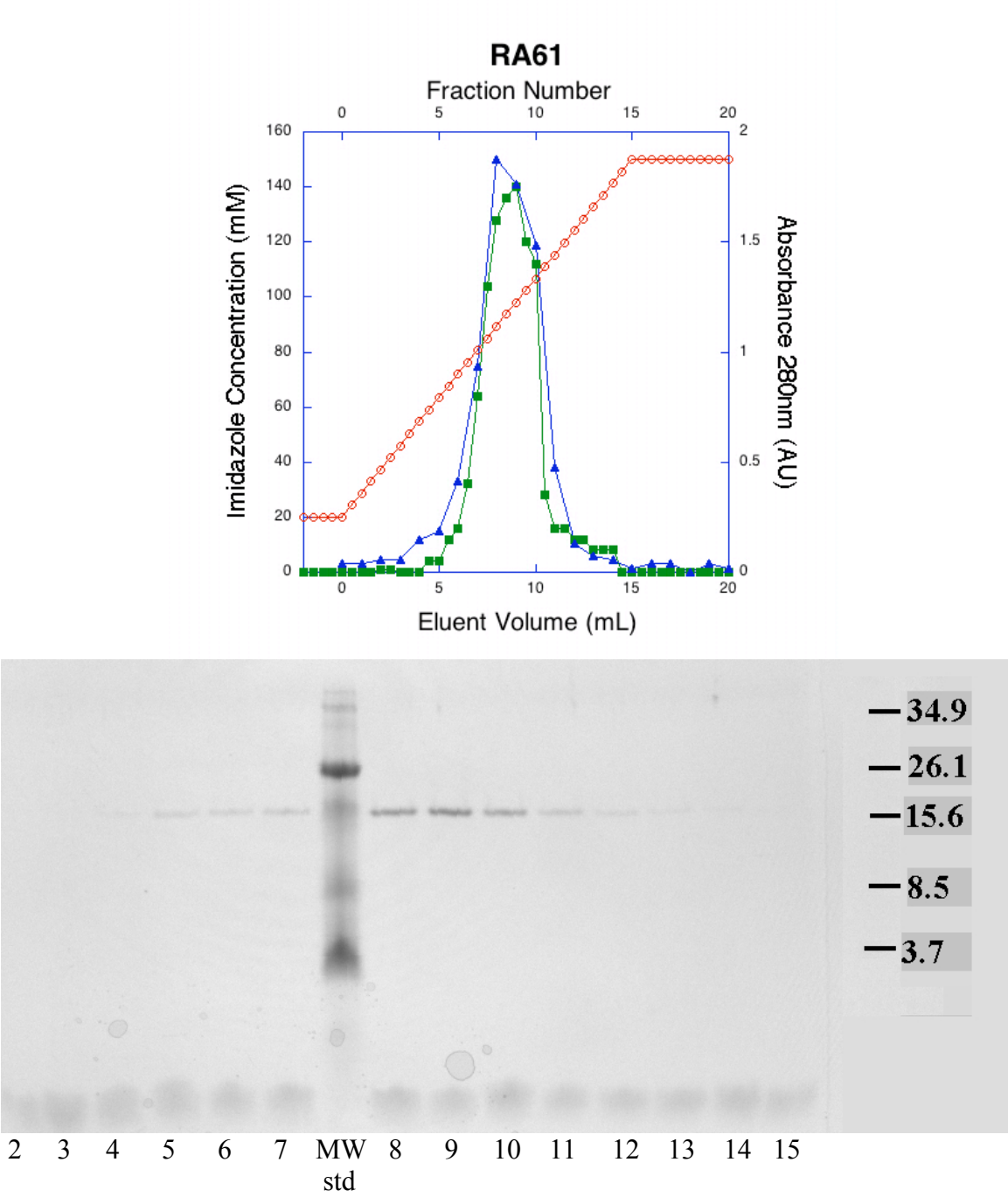


Figure S7. RA61 eluted with an imidazole gradient

Column fractions from a Ni-NTA column eluted with an imidazole gradient from 20 to 150 mM over 15 minutes at 1 mL/min. The gradient is shown in red and 1 mL fractions are collected (top panel). The absorbance at 280 nm for each fraction is shown in green and the relative retro-aldolase activity of each fraction, when assayed with 270 uM substrate, is shown in blue. The fractions were then analyzed by SDS PAGE (bottom panel). The mass of RA61 is 22.9 kDa.

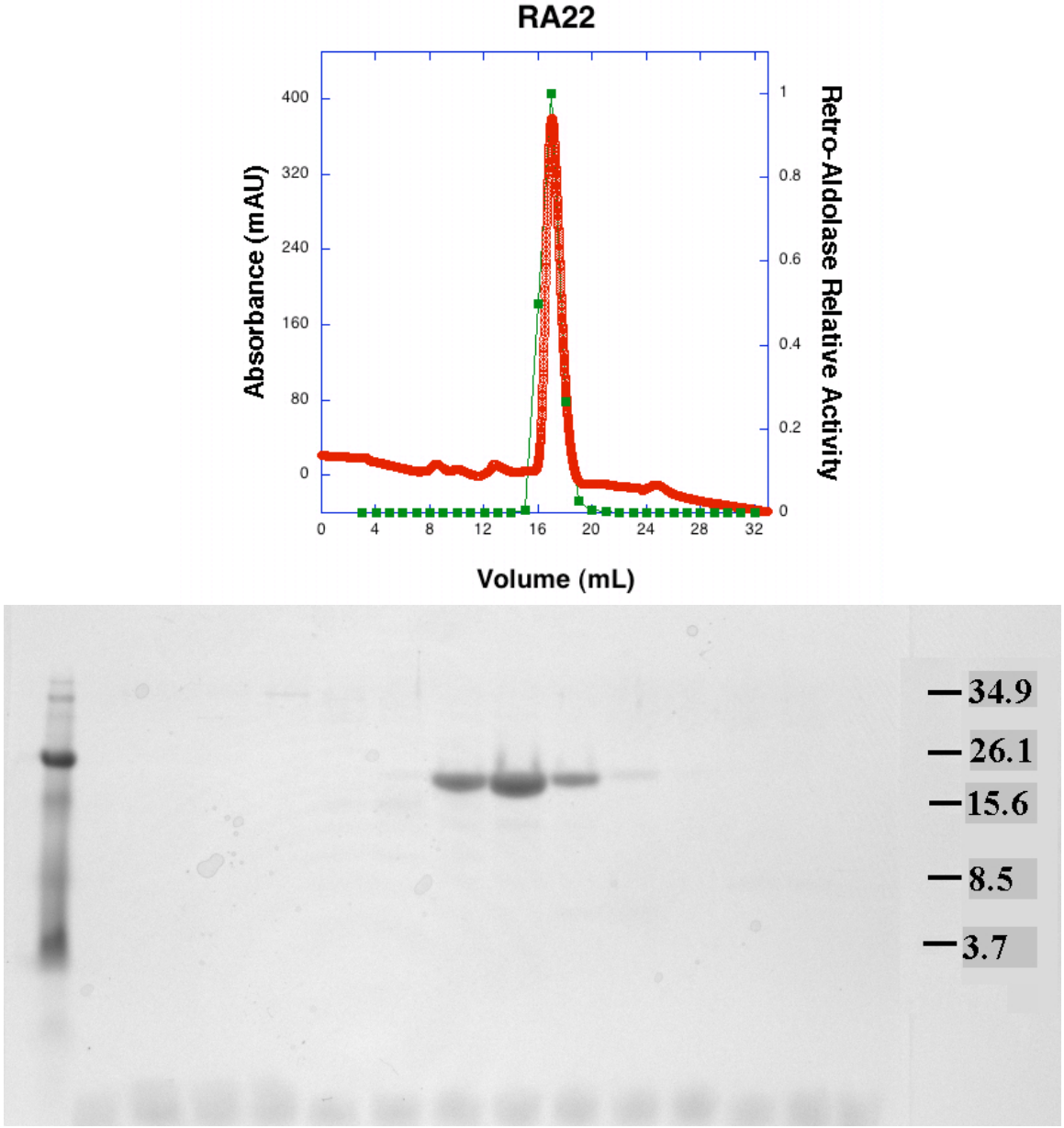


Figure S8A. RA22 size exclusion chromatography

1mL fractions were collected and tested for retro-aldolase activity. The absorbance at 280 nm of the eluent (red) and the relative retro-aldolase activity (green) of the fraction are shown versus to the elution volume. Fractions with significant absorbance at 280 nm were subjected to SDS PAGE (lower panel). The mass of RA22 is 29.5 kDa.

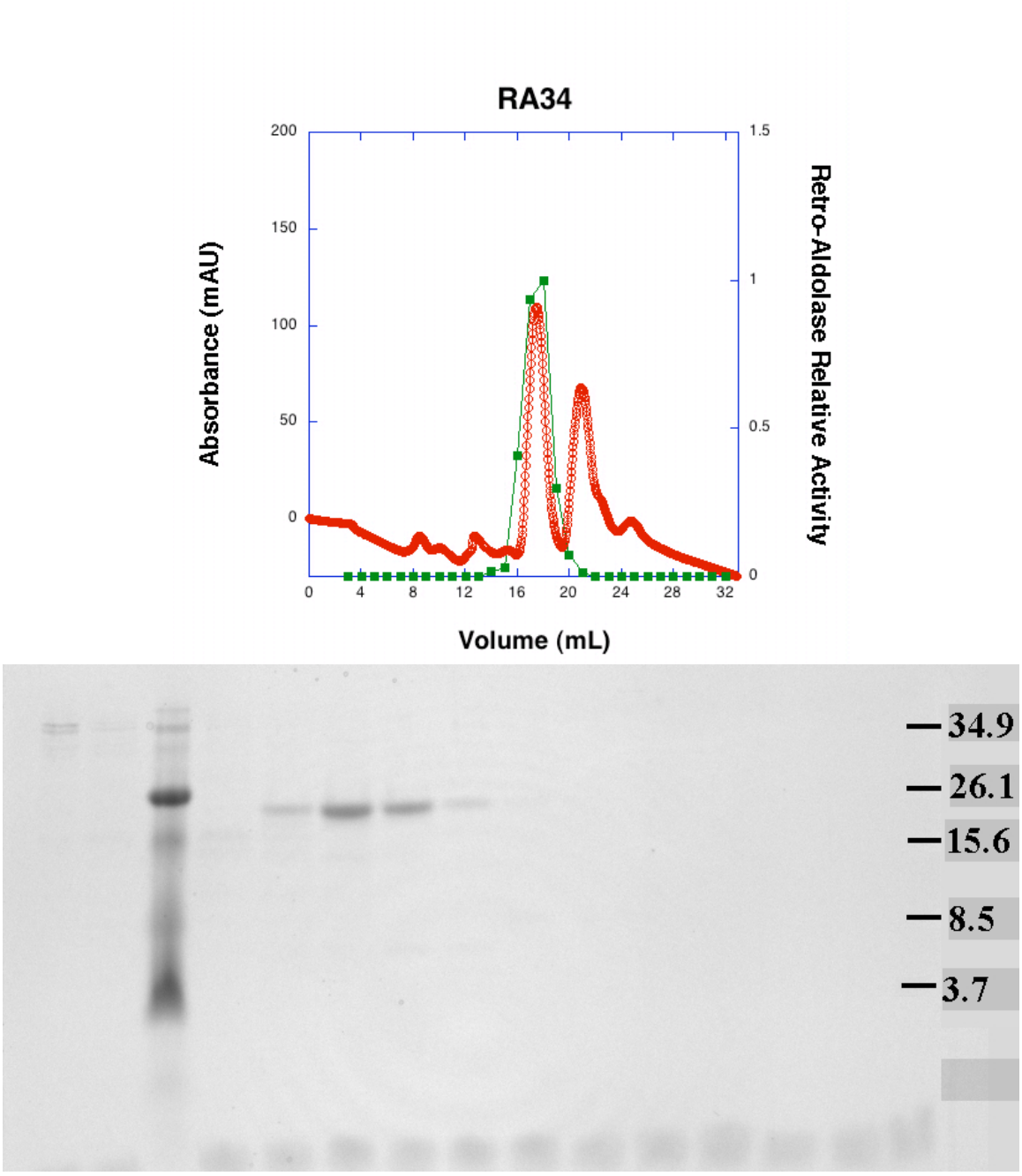


Figure S8B. RA34 size exclusion chromatography

1mL fractions were collected and tested for retro-aldolase activity. The absorbance at 280 nm of the eluent (red) and the relative retro-aldolase activity (green) of the fraction are shown versus to the elution volume. Fractions with significant absorbance at 280 nm were subjected to SDS PAGE (lower panel). The mass of RA34 is 29.6 kDa. The additional absorbance peak at 22mL is not protein (see the gel) and has no catalytic activity.

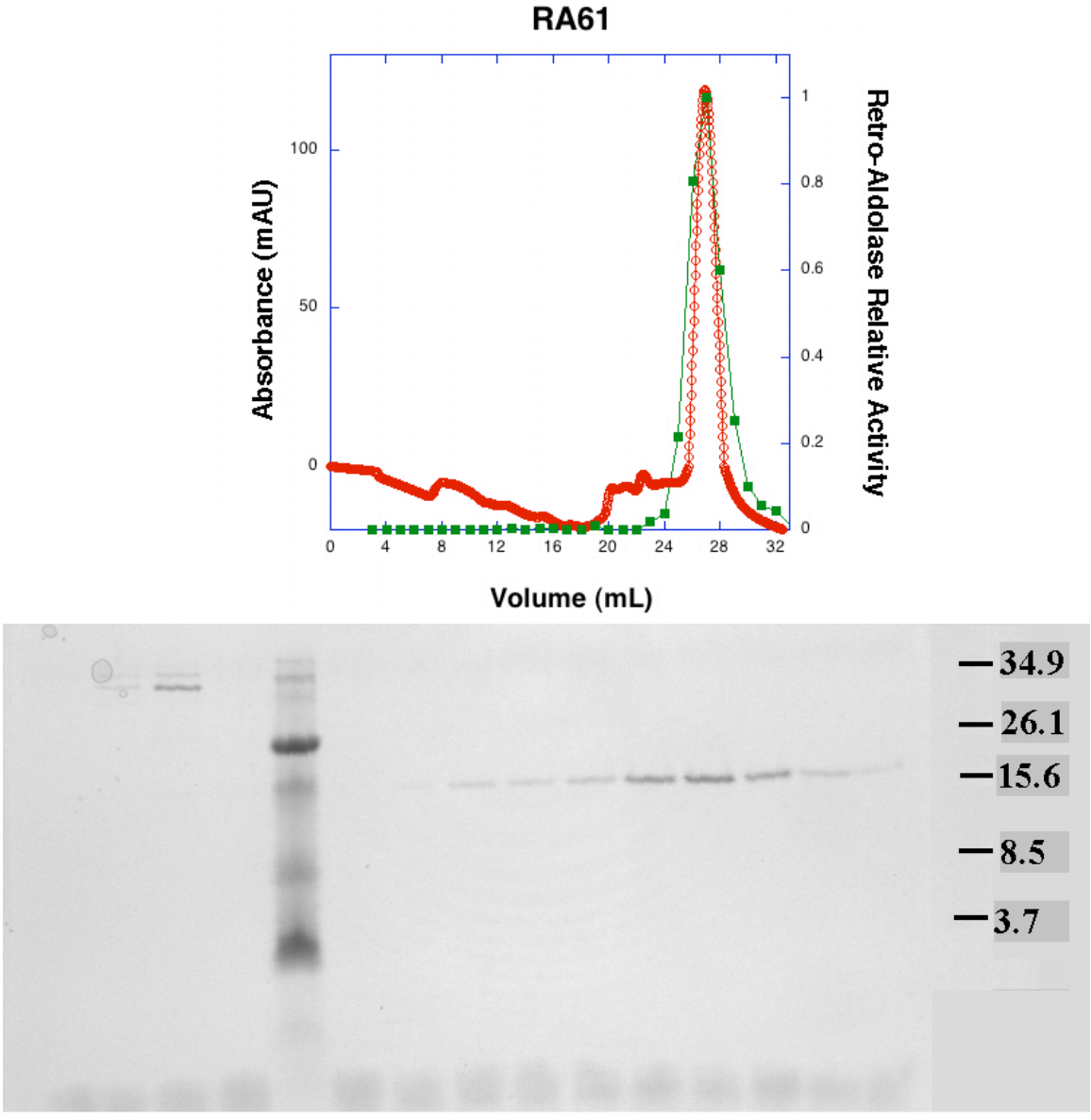


Figure S8C. RA61 size exclusion chromatography
1mL fractions were collected and tested for retro-aldolase activity. The absorbance at 280 nm of the eluent (red) and the relative retro-aldolase activity (green) of the fraction are shown versus to the elution volume. Fractions with significant absorbance at 280 nm were subjected to SDS PAGE (lower panel). The mass of RA61 is 22.9 kDa.

Table S3. Combinatorics of transition state and catalytic sidechain placement

	# of TS models	# of sidechain conformations			Total combinations
Motif I		Lys	Asp, Glu	Lys'	
	86	122,472	648	13,608	9.3×10^{13}
Motif II		Lys	Tyr	Pi-stack	
	1,296	244,994	11,664	244,994	9.1×10^{17}
Motif III		Lys	His	Pi-stack	
	1,296	244,994	17,496	244,994	1.4×10^{18}
Motif IV		Lys	H-bond	Basic residue	
	1,296	244,994	8,748	84,564	2.3×10^{17}

Table S4. Protein scaffold list

Fold	# scaffolds	PDB entry code (*19*)
Lipocalin	8	1cbs 1dc9 1icm 1icn 1ifc 1lic 1sa8 2ifb
Periplasmic binding protein	5	1abe 1gca 1hsl 1wdn 2dri
TIM beta/alpha-barrel	19	1dqx 1ebg 1eix 1ftx 1h61 1pii 1a53 1b9b 1btm 1dl3 1i4n 1igs 1lbf 1lbl 1lbm 1qo2 1thf 1tml 2btm
Jelly roll	5	1f5j 1h1a 1m4w 1pvx 1yna
Beta-propeller	34	1c9u 1cq1 1cru 1e2r 1eus 1gye 1h6l 1hzu 1hzv 1ms0 1n1s 1n1t 1n1v 1n1y 1pt2 1s1d 1sq9 1st8 1t2x 1uyp 1w8n 1w8o 1crz 1e1a 1poo 1q7f 1suu 1v04 2ber 2fhr 2fp9 2fpb 2fpc 2h13

Table S5. Protein fold distribution of *in silico* designs after active site identification and pocket design

Fold	# scaffolds	# matches per scaffold	# low energy designs per scaffold*
Lipocalin	8	4498	4
Periplasmic binding protein	5	1465	3
TIM beta/alpha-barrel	19	3595	10
Jelly roll	5	6169	15
Beta-propeller	34	849	2

*selection criterion: TS binding energy, total energy of the whole complex, packing around the lysine, top ranking on each key energy components, satisfaction on catalytic geometries.

Table S6. Distribution of rate enhancements of designed retro-aldolases

Rate enhancement (k_{cat}/k_{uncat})	# Designs
$1\sim10^1$	38
$10^1\sim10^2$	6
$10^2\sim10^3$	12
$10^3\sim10^4$	12
$10^4\sim10^5$	2
ALL	70

Table S7. Sequence information for active retro-aldolase designs

Jelly roll

Design	PDB scaffold	46	48	72	74	87	89	119	121	133	135	137	176	178
	1m4w	N	V	N	Y	E	Y	T	R	F	Q	W	E	Y
RA59	9	T	-	Y	-	S	-	G	Y	-	I	V	K	S
RA60	12	W	K	T	W	S	S	Y	W	Y	S	-	V	V
RA61	8	H	M	I	-	S	L	-	-	-	I	-	K	W

Design	PDB scaffold	16	43	45	81	90	180
	1f5j	E	N	L	Y	E	E
RA28	6	A	F	A	F	T	K

Design	PDB scaffold	9	11	48	78	101	103	126	128	130	169	171	176	201	222	224
	1thf	C	D	V	T	S	N	V	A	D	L	T	D	S	L	A
RA17	14	V	L	A	I	A	L	K	-	F	A	V	L	H	A	D
RA58	10	S	Y	T	-	-	H	-	S	W	K	V	W	H	-	-

TIM

Design	PDB scaffold	7	47	49	79	81	83	87	106	108	110	129	131	155	157	177	179	183	209	210	229
	1i4n	I	E	K	S	L	E	F	L	K	F	L	I	L	E	G	N	L	E	S	L
RA31	14	A	Q	S	L	D	A	-	-	H	I	V	L	-	L	-	K	-	L	W	-
RA32	13	-	Q	S	T	D	-	-	V	H	-	V	-	-	V	K	A	A	F	-	S
RA33	15	-	Q	S	T	D	A	A	V	H	-	-	-	V	V	K	A	A	L	-	S

Design	PDB scaffold	9	51	53	56	58	81	83	89	110	112	131	135	159	161	178	180	182	184	187	210	211	231	233
	1lbf	L	E	K	S	S	S	L	F	K	F	L	K	E	N	G	N	R	L	L	E	S	L	G
RA22	13	-	G	D	-	-	-	T	-	A	-	A	-	K	-	-	V	A	W	-	S	F	S	H
RA34	13	-	Y	M	-	W	-	T	A	A	-	A	-	K	-	-	V	-	W	-	S	Y	S	-

RA35	14	-	G	I	-	-	A	T	G	A	-	A	-	K	-	-	I	-	W	-	S	Y	S	Y					
RA36	17	-	G	V	G	W	A	A	W	T	-	A	-	K	-	-	V	A	W	-	S	F	A	Y					
RA39	14	-	G	S	-	T	-	-	-	T	-	A	-	V	T	K	L	-	W	G	S	-	S	Y					
RA41	17	A	V	S	-	-	T	Y	W	A	L	V	P	V	V	K	K	-	W	-	S	-	I	-					
RA47	15	-	G	M	-	W	A	T	A	A	-	A	-	K	-	-	V	-	W	-	S	Y	S	Y					

		8	51	53	56	58	81	83	85	89	110	112	131	133	135	157	159	161	178	180	182	184	187	210	211	231	233	234
	1lbl	W	E	K	S	S	S	L	E	F	K	F	L	I	K	L	E	N	G	N	R	L	L	E	S	L	G	S
RA6	15	-	M	S	-	W	V	-	-	A	V	W	V	V	-	-	V	-	K	S	E	-	-	A	H	-	-	-
RA42	15	-	L	S	-	W	V	A	-	T	W	-	-	-	-	-	V	-	K	S	M	Y	-	S	-	V	Y	-
RA45	13	-	L	S	H	-	-	W	-	V	V	-	-	F	-	-	V	-	-	K	-	W	-	M	-	-	E	Y
RA46	14	-	V	S	-	-	-	Y	-	W	V	L	V	-	S	V	V	-	L	K	-	-	-	S	W	-	-	-
RA48	16	-	M	S	-	L	A	F	-	W	L	-	K	S	L	-	S	Y	L	V	-	P	-	L	-	-	-	-
RA49	17	-	M	S	-	W	V	-	G	A	V	W	V	V	-	-	T	-	K	S	-	H	-	A	L	I	-	-
RA55	12	-	Y	T	-	W	-	-	-	-	T	-	A	-	-	-	K	-	-	V	-	W	-	S	-	S	V	I
RA56	16	R	G	I	H	-	A	Y	-	T	I	-	V	-	-	-	K	-	-	H	-	W	F	T	-	S	Y	-
RA57	13	-	G	M	-	-	-	W	-	S	L	-	V	-	-	-	K	-	-	V	-	W	-	S	H	T	Y	-

		51	53	58	110	131	133	135	159	180	184	210
	1igs	E	K	S	K	L	I	K	E	N	L	E
RA63	11	L	M	W	H	K	S	Y	S	V	P	I

		9	51	53	56	58	81	83	85	89	110	112	131	133	135	159	161	178	180	182	184	187	189	210	211	231	233
	1a53	L	E	K	S	S	S	L	E	F	K	F	L	I	K	E	N	G	N	R	L	L	I	E	S	L	G
RA26	10	-	-	S	-	W	L	-	-	-	H	-	K	-	-	A	H	-	S	-	P	-	-	L	-	-	-
RA40	17	V	V	M	-	-	T	Y	-	-	V	-	V	F	P	V	-	K	L	S	V	-	-	S	H	S	-
RA43	16	-	V	S	-	-	L	W	G	G	V	W	-	V	-	S	Y	K	A	-	P	-	-	S	-	V	-
RA53	16	-	V	S	H	-	-	Y	-	T	V	V	V	-	-	K	-	V	S	Y	W	-	-	S	G	S	-
RA68	10	-	-	-	-	T	-	S	-	H	-	-	-	-	-	L	-	-	-	-	Y	G	T	I	Y	-	S

Table S8. Crystallographic Data and Refinement Statistics

Protein Construct	**RA22 S210A**	**RA61 M48K**
Space Group	$P2_12_12$	$P2_12_12_1$
Unit Cell Dimensions (Å)	a=72.05, b=73.58, c=47.40	a=39.19, b=66.06, c=66.07
Resolution Limits (Total / (final shell)	36.0 - 2.20 (2.28 - 2.20)	33.7 - 1.80 (1.86 - 1.80)
# Reflections (total / unique)	45874 / 13187	56087 / 16082
Redundancy	3.48 (2.89)	3.49 (2.45)
Completeness	98.9 % (97.2 %)	97.5 % (83.3 %)
$I/\sigma(I)$	10.4 (3.8)	12.0 (3.1)
R_{merge}	0.068 (0.258)	0.059 (0.220)
R_{work} / R_{free}	0.236 / 0.293	0.203 / 0.248
rmsd (bond length (Å), bond angle(°))	0.009/1.305	0.007 / 1.069
Average B-factor ($Å^2$)	32.0	25.0
Ramachandran Distribution (% most favored, additional allowed, generously allowed, disallowed)	90.3, 9.3, 0.0, 0.4	88.5, 11.5, 0.0, 0.0