Supporting Information for "Computational Structure-Based Redesign of Enzyme Activity"

Cheng-Yu Chen,[1,†] Ivelin Georgiev,[2,†] Amy C. Anderson,[3] Bruce R. Donald[1,2,*]

[1]Department of Biochemistry, Duke University Medical Center

[2]Department of Computer Science, Duke University, Durham, NC 27708-0129, USA

[3]Department of Pharmaceutical Sciences, University of Connecticut, Storrs, Connecticut 06269

[†]These authors contributed equally to the work.

[*]To whom correspondence should be addressed.

Sec. S1 describes in more detail the protein redesign algorithms and the computational protocol, including computational experiment design and supporting data, used to redesign GrsA-PheA. Sec. S2 gives details about the structural analysis of the predicted mutants (Sec. S2.1) and the comparison to other methods and evolution (Sec. S2.2). Sec. S3 gives additional details about the experimental protocol and supporting data from the *in vitro* experiments. **SI** references (e.g., [6, 8]) are provided at the end of the **SI**.

S1 Algorithm

In Sec. S1.1, we describe the K^* algorithm, the associated computational protocol, and the application of K^* to redesign GrsA-PheA. In Sec. S1.2, we describe the computational protocol for identifying bolstering mutations outside of the enzyme active site, as well as the application of this protocol in a redesign of an active site GrsA-PheA mutant with Leu as substrate.

S1.1 K^*

First, we present an overview of the K^* ensemble-based protein redesign algorithm. We then outline the computational protocol for redesigning a given enzyme/substrate complex using K^*. Finally, we give details about the application of K^* to redesign the active site of GrsA-PheA for the target substrates Leu, Arg, Glu, Lys, and Asp. This section extends the relevant discussion in the *Materials and Methods* section in the main article.

S1.1.1 Description of the K^* Algorithm

This section describes the steps of the K^* algorithm. For a detailed description, as well as mathematical proofs of the algorithm's correctness, we refer the reader to [6, 8]. For a given protein-substrate complex, K^* computes a provably-accurate ε-approximation to the binding constant, up to the accuracy of the model (including the input structure, rotamer library, and energy function).

The input to K^* includes an initial protein structure in PDB format, a selection of residue positions to be redesigned (these residues are typically part of the binding/active site of the protein), a target substrate, and a rotamer library. All steps below are implemented in code and are fully automated. An initial sequence-space filter is applied to reduce the set of candidate mutations by restricting the allowed amino acid types for each of the redesigned positions. The user further specifies a value k, so that a k-point mutation search will be performed by the algorithm. In a k-point mutation search, at most k of the n flexible residues are allowed to simultaneously mutate, while the remaining flexible residues are allowed to change their side-chain conformation. This generates a set of candidate mutation sequences that must be evaluated by K^*. A volume filter is then used to prune sequences that are under- or over-packed relative to the wildtype (WT) enzyme/WT substrate complex. The remaining unpruned sequences are then supplied for evaluation to the partition function computation algorithm.

For a given protein sequence and the target substrate, K^* computes a provably-accurate ε-approximation to the binding constant (ε is a user-specified parameter). To achieve this, K^* computes ε-approximations to the partition functions for the bound protein-substrate complex, for the free protein, and for the free substrate. To compute a given partition function, K^* considers all rotamer-based conformations (e.g., for a protein-substrate complex, this may include all combinations of the rotamers for the active site residues as well as the substrate rotamers). For computational efficiency, K^* applies Dead-End Elimination (DEE)-based pruning [4, 10] to prune the majority of the rotameric conformations from further evaluation. Depending on the types of protein flexibility allowed, K^* can use the MinDEE [6] pruning criteria (for side-chain dihedral flexibility) and the BD [5] pruning criteria (for protein backbone flexibility) in this initial pruning stage. The remaining unpruned conformations are then enumerated by the A^* algorithm [6] in order of increasing lower bounds on the conformational energies. The enumerated conformations are energy-minimized and their energies are added to the computed partition function using Boltzmann probabilities. A provable halting condition [6] is used to guarantee that the desired ε-approximation is achieved. Since only low-energy conformations can contribute significantly to the partition function, typically only a very small subset of the remaining unpruned conformations must be enumerated by A^* before the provable halting condition is reached. An additional inter-mutation pruning algorithm is applied to efficiently evaluate and prune low-scoring mutation sequences [6].

S1.1.2 Outline of Protocol

Here, we briefly outline the sequence of steps in the protocol for enzyme redesign using K^*.

1. Obtain (or download) pdb structure for the enzyme/substrate complex to be redesigned;

2. Extract the substrate, active site residues, and a set of residues close to the active site (the *steric shell*) to use as the input structure for K^*. The steric shell constrains the movement of the active site residues and is used to speed up the computation;

3. Generate configuration files for the rotamer library, energy function and mutation search parameters;

4. Invoke the K^* mutation search on the generated input configuration files.

In particular, for the computational experiments with GrsA-PheA, steps 1 and 2 are described in the *Materials and Methods* section in the main article, and steps 3 and 4 are described in the *Materials and Methods* section in the main article and Sec. S1.1.3 below.

S1.1.3 Application to GrsA-PheA

This section gives additional details about the K^* mutation search parameters (Sec. S1.1.2, step 3) used in the computational redesigns of the active site of GrsA-PheA, and is a supplement to the relevant discussion in the *Materials and Methods* section in the main article.

A 2-point mutation search was performed for each of the target substrates. In a k-point mutation search, at most k of the n flexible residues are allowed to simultaneously mutate, while the remaining flexible residues are allowed to change their side-chain conformations. Out of all possible 2-point mutations, the sequence-space and volume filters were applied to select a set of mutants to be explicitly enumerated and evaluated by the partition function computation algorithm. 4452 sequences and 4.71×10^8 rotameric conformations were considered in our designs. For each of the redesign targets Leu, Arg, Glu, Lys, and Asp, sets of the top computational predictions were then visualized and selected for experimental validation. In all cases, the computational predictions selected to be tested experimentally were in the top ten as ranked by the algorithm.

In addition to the WT identity for the seven residue positions allowed to mutate (Ala236, Trp239, Thr278, Ile299, Ala301, Ala322, and Ile330), the following subsets of amino acid types were allowed in the redesigns for the different substrates:

- **Leu**: the set (Gly Ala Val Leu Ile Tyr Phe Trp Met) was allowed for all seven residue positions;

- **Arg and Lys**: 236 (Gly Ala Val Leu Ile Tyr Phe Trp Met), 239 (Gly Ala Val Leu Ile Tyr Phe Trp Met Ser Thr His Asn Gln Lys Arg Asp Glu), 278 (Gly Ala Val Leu Ile Tyr Phe Trp Met Ser Thr His Asn Gln Lys Arg Asp Glu), 299 (Gly Ala Val Leu Ile Tyr Phe Trp Met Ser Thr His Asn Gln Lys Arg Asp Glu), 301 (Gly Ala Cys), 322 (Gly Ala Val Leu Ile Tyr Phe Trp Met), 330 (Gly Ala Val Leu Ile Tyr Phe Trp Met);

- **Asp and Glu**: 236 (Gly Ala Val Leu Ile Met), 239 (Gly Ala Val Leu Ile Met Ser Thr His Asn Gln Lys Arg), 278 (Gly Ala Val Leu Ile Met Ser Thr His Asn Gln Lys Arg), 299 (Gly Ala Val Leu Ile Met Ser Thr His Asn Gln Lys Arg), 301 (Gly Ala), 322 (Gly Ala Val Leu Ile Met Ser Thr His Asn Gln Lys Arg Asp Glu), 330 (Gly Ala Val Leu Ile Met).

From the set of all sequence combinations resulting from the allowed sets of amino acids, only sequences with up to two mutations from the WT active site were considered (note that the WT and all single-point mutations were also included). The number of candidate sequences (with the total number of conformations shown in parenthesis) after this stage were as follows:

- Leu: 1450 (6.44×10^7)
- Arg: 2511 (5.06×10^8)
- Glu: 1633 (1.72×10^8)
- Lys: 2511 (4.05×10^8)
- Asp: 1633 (1.14×10^8)

For each target substrate, the volume filter was applied to further prune sequences that were more than 30 Å^3 (for Leu) or 40 Å^3 (for all other substrates) from the WT enzyme/WT substrate volume. The number of remaining sequences (with the total number of conformations in parenthesis) that were enumerated and evaluated by K^* were as follows:

- Leu: 505 (1.12×10^7)
- Arg: 1259 (1.54×10^8)
- Glu: 776 (8.44×10^7)
- Lys: 1237 (1.53×10^8)
- Asp: 675 (6.86×10^7)

The ligand substrate was also modeled using continuously-flexible rotamers and was allowed to rotate/translate [6]. A value of 0.03 was used for ε, in order to guarantee that the computed partition functions were at least 0.97 of the respective full partition functions (when all rotamer-based conformations are included). The inter-mutation pruning filter (see **SI** S1.1) was used to guarantee that ε-approximation scores were computed for all sequences whose score was within two orders of magnitude from the best score for the given mutation search. The K^* algorithm distributes the computation of the K^* approximation scores for the different mutation sequences to a cluster of compute nodes, such that a single sequence is distributed to a single processor. For the different substrates, the average time in minutes required to compute the K^* score for a single sequence (along with the max time per sequence in parenthesis) was as follows: 3 (114) [Leu]; 10.9 (205) [Arg]; 9.6 (225.6) [Glu]; 10.8 (262.3) [Lys]; 8.3 (202.1) [Asp]. The top ten K^* 2-point mutation predictions for each of the target substrates are shown in Table S1.

As an orthogonal check for the computational predictions with Leu as substrate, additional K^* runs that allowed flexible backbones were performed. These runs applied the BD algorithm [5] in the DEE pruning stage and continuous backbone minimization in the A^* enumeration stage, but were otherwise identical to the continuously-flexible rotamers K^* runs using MinDEE (**SI** S1.1).

Table S1: Top ten 2-point active site mutants predicted by K^* for each of the five target substrates (Leu, Arg, Glu, Lys, and Asp)

	Leu	**Arg**	**Glu**	**Lys**	**Asp**
1	278L/301G	278D/301G	239R/322R	278D/299D	239M/278R
2	278I/301G	299D/301G	278H/301G	278D/299E	278H/301G
3	299W/301G	278E/301G	239R/278H	278E/299D	278K/301G
4	299F/301G	278A/301G	322Q/301G	278D/301G	239K/278R
5	236M/301G	299E/301G	278K/301G	278E/299E	278K/330M
6	236L/301G	236G/278A	239K/278H	278D/299M	239R/278K
7	330F/301G	322G/301G	278N/322K	278D/299Q	239K/322K
8	278M/301G	278S/301G	278N/301G	278D/299W	239K/322R
9	322V/301G	299L/301G	278H/299K	278D/299F	236M/278K
10	278F/301G	236G/301G	278H/299R	278D/322G	278K/299M

The goal of these additional runs was to determine whether the top Leu mutants predicted by the MinDEE K^* runs (**SI** S1.1) and selected for experimental validation, would also be among the top predictions when the backbone was allowed to flex. Indeed, 278L/301G was again ranked $\mathbf{1}^{st}$, 278M/301G - $\mathbf{10}^{th}$, and 322V/301G - $\mathbf{4}^{th}$. The fact that these three mutants were still within the top ten predictions even when backbone flexibility was allowed, further increased the confidence in the feasibility of our predictions.

S1.2 Bolstering Mutation Prediction

This section gives details about the design of the bolstering mutation prediction computational experiments. First, we present an overview of the Self-Consistent Mean Field (SCMF) entropy-based method for estimating residue positions susceptible to beneficial mutations [13]. We then describe how our MinDEE/A^* algorithm [6] can be combined with SCMF as part of a computational protocol for identifying mutations outside of the enzyme active site for further improvement in the target substrate specificity. Finally, we give details about the application of our SCMF with MinDEE/A^* protocol in the GrsA-PheA redesigns for Leu. This section extends the relevant discussion in the *Materials and Methods* section in the main article.

S1.2.1 Description of the SCMF Approach

In [13], a SCMF entropy-based method for estimating residue positions susceptible to beneficial mutations was applied as a preprocessing step for focusing directed evolution experiments. Using SCMF, the probabilities for different amino acid types and, consequently, the residue entropy at each position in a protein can be estimated. For a given protein with n residue positions, the residue entropy S_i for each residue position i can be computed as:

$$S_i = -\sum_{a \in A_i} p(a|i) \ln p(a|i), \tag{S1}$$

where A_i is the set of amino acid types allowed at residue position i, and $p(a|i)$ is the probability of having amino acid type a at residue position i. Here,

$$p(a|i) = \sum_{r \in R_a} p(r|i), \tag{S2}$$

where R_a is the set of rotamers (as given by the rotamer library) for amino acid type a and $p(r|i)$ is the probability of having rotamer r for amino acid a at residue position i. The probabilities $p(r|i)$ are computed using SCMF [13].

Since high residue entropy implies the existence of multiple amino acid types with reasonably high probabilities, this method was used as a means to identify residue positions that are tolerant to mutations.

S1.2.2 Outline of Protocol

We extended and modified the SCMF method to make it applicable as part of a five-step protocol for determining mutations both close to and far away from the enzyme active site. The steps of our protocol are as follows:

1. Apply the K^* algorithm [6] to compute active site mutations with the desired target substrate specificity;
2. Experimentally test a set of top-ranked K^* predictions;
3. A computationally-predicted active-site mutant that is experimentally-verified to have a reasonably high specificity toward the target substrate is then selected for further redesign;
4. SCMF is applied to select residue positions anywhere in the protein (both close to and far away from the active site) that are to be redesigned;
5. TheMinDEE/A^* algorithm [6] is applied to predict mutations to the set of residue positions identified in step 4 for further improvement in the substrate specificity.

The active site mutations (step 2) and bolstering mutations (step 5) are then combined, and the resulting mutants are experimentally tested.

S1.2.3 Application to GrsA-PheA

This section gives additional details about the design of the bolstering mutation prediction computational experiments for GrsA-PheA, and is a supplement to the relevant discussion in the *Materials and Methods* section in the main article.

For all non-Pro residues in the structure 1amu (residues with missing heavy atoms in the crystal structure were removed from the input structure), the residue entropy was computed using SCMF. Residue positions with too few neighboring (in space) residues were removed from further consideration, since we did not have sufficient confidence in the entropy estimates for residue positions

with a small number of residue interactions. Let x_i be the number of neighboring residues (within a distance cutoff d) for residue position i, and let X be the set of numbers x_i for all residue positions i. We then discarded all residues that had fewer neighboring residues than the σ^{th} percentile threshold for the distribution of the elements in X. In all our computational experiments, we used $d = 5$ Å (for any pair of rotamers for any two residue positions) and $\sigma = 75$ (in effect, this mostly discards surface positions).

The remaining residue positions were then ranked in order of decreasing residue entropy. The eight residue positions with the highest computed residue entropy S_i (45, 187, 207, 210, 238, 239, 277, and 447) were selected for the MinDEE/A^* mutation search (Fig. S1, *top*). The lowest-energy rotamer conformation for the highest-activity K^* mutant for Leu (T278L/A301G) was used as input to the MinDEE/A^* redesign algorithm [6]. The MinDEE/A^* input structure included the residue positions in 1amu within 8 Å of the ligand or the eight high-entropy positions identified in the entropy step. Residues 45, 187, 207, 210, 238, 277, and 447 were modeled using continuously-flexible rotamers and allowed to mutate; residue 239 was modeled using continuously-flexible rotamers but was not allowed to mutate, since it is part of the enzyme active site.

In the MinDEE/A^* run, only the amino acid types with probabilities above a certain cutoff threshold (as computed by SCMF) were allowed for each of the seven mutable residue positions. We used the 75^{th} percentile of the distribution of all amino acid probabilities for all residue positions as the cutoff threshold. In addition to the WT identity, the allowed amino acid types in the MinDEE/A^* experiments were as follows: 45 (Ala Leu Ile Phe Asn Gly), 187 (Ala Val Leu Ile Ser Thr Asp Glu Asn Gly), 207 (Ala Leu Ile Ser Thr Asn Gly), 210 (Ala Leu Ile Phe Tyr Asn Gly), 238 (Ala Leu Ile Ser Thr Asn Gly), 277 (Ala Val Leu Ile Ser Thr Asn Gly), 447 (Ala Val Leu Ile Ser Thr Asn Gly). For computational efficiency, we applied a heuristic halting condition for the MinDEE/A^* search, different from the provably-accurate halting condition described in [6]. Let b_m be the computed lower bound on the conformational energy of the first rotameric conformation generated by A^*, and let b_c be the computed lower bound on the conformational energy of the current rotameric conformation generated by A^*. The MinDEE/A^* search was then halted when $b_c > b_m + \lambda$; in our experiments, we used a value of 2.0 for λ.

Up to 3-point bolstering mutation search (in addition to the initial 2 active site mutations) was performed for the mutable positions, and the top mutants were visualized and selected (Table S2). At that point, mutant proteins were created using site-directed mutagenesis by adding to the active site double mutant 1-, 2-, and 3-point bolstering mutations comprising the $\leq$ 3-point MinDEE/A^*-predicted mutants. The lowest-energy S447N conformation is shown in Fig. S1(*bottom*). Some modifications to the energy function were introduced for the entropy and MinDEE/A^* steps. A vdW radii scaling factor of 0.95 and a solvation-energy scaling factor of 0.5 were used; hydrogens were not used in the vdW energy computation. In addition, amino acid reference energies were computed and used (similarly to [7]) to limit the number of times a particular amino acid type is selected within a given redesign. A simpler version of our software that did not include various enhancements and modifications to the algorithm (as compared to the K^* experiments) was used for the SCMF and MinDEE/A^* runs. The bolstering mutation computation required approximately a week of computational time.

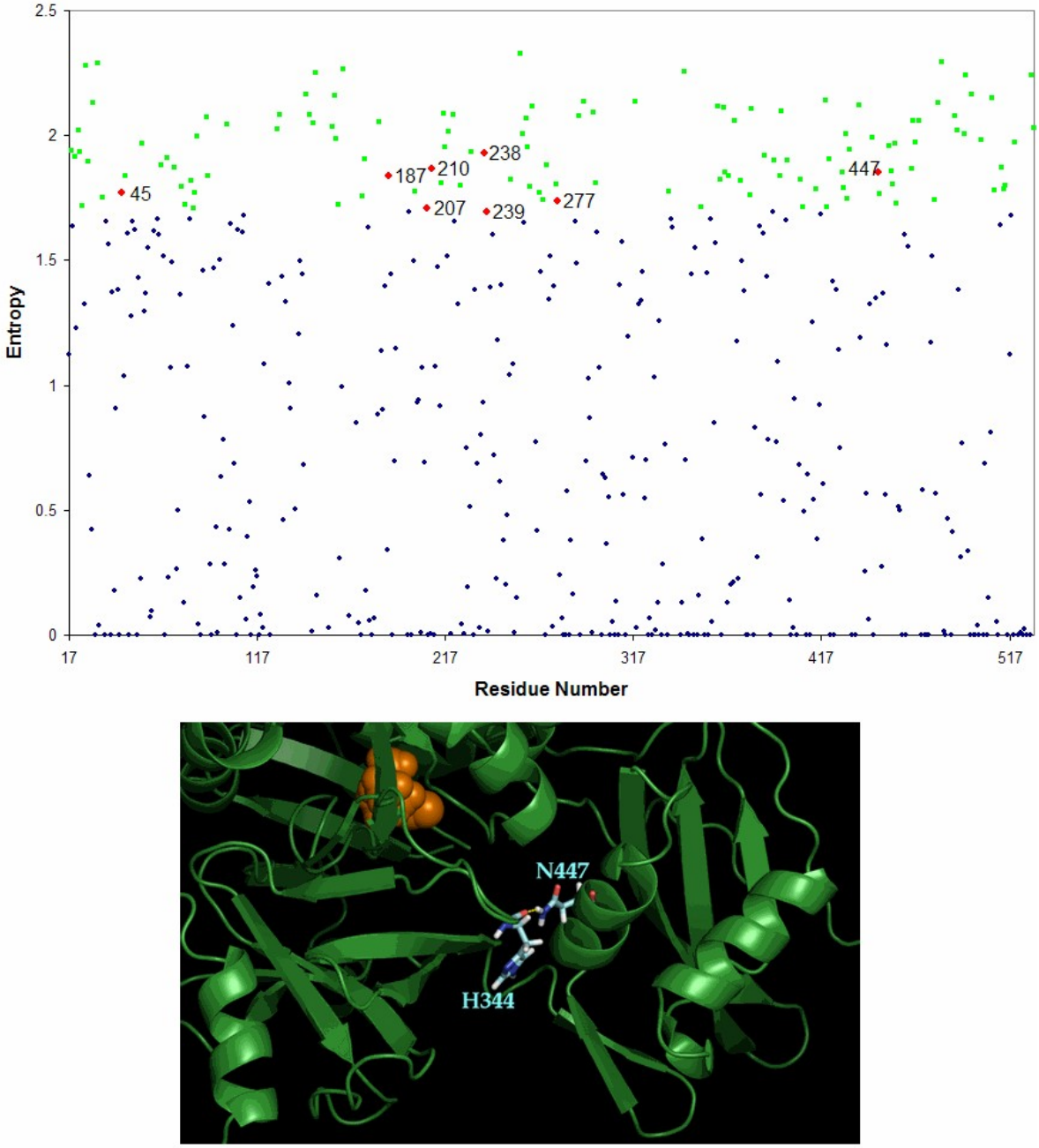

Figure S1: (*top*) Residue entropy computed by the SCMF approach. The eight positions selected for the MinDEE/A^* mutation search are labeled and shown in red. The positions discarded by the 'neighboring residues' filter are shown in green. (*bottom*) The lowest-energy S447N conformation predicted by MinDEE/A^*. The hydrogen bond between H344 and N447 (the distance between N_{δ_2} (Asn) and the backbone carbonyl oxygen of His is 2.94 Å) is shown with a dashed yellow line. The Leu ligand is shown in orange.

Table S2: Top bolstering mutations from the 3-pt MinDEE/A^* mutation search. These mutations are in addition to the 2-point active site mutation 278L/301G

Rank	Mutant
1	187L/238I/447L
2	187L/238I/447N
3	187L/238I/447I
4	187L/277L/447L
5	187L/238I/447A
6	187L/447L
7	187L/207L/447L
8	187L/277L/447N
9	187L/277L/447I
10	187L/447N
11	187I/238I/447L

S2 Analysis of the Computational Results

In this section, we give details about the structural analysis of the predicted mutants (Sec. S2.1) and the comparison to other methods and evolution (Sec. S2.2).

S2.1 Structural Analysis

Structural comparison between the predicted mutant structures and the WT can reveal insights into the reasons for the switch of specificity in the Leu redesigns. We thus generated and visualized the structures in the K^* bound-state ensemble for the T278L/A301G mutant. When overlayed with the WT structure, all conformations of the Leu substrate found in the K^* ensemble clash sterically with the side-chain of Ala at position 301. The mutation A301G appears to free up the space necessary to accommodate the Leu side-chain. In addition, the mutation T278L fills up the enzyme pocket to partially compensate for the change in the substrate size from the bulkier Phe to the smaller Leu. The lowest-energy T278L/A301G structure with Leu as substrate is shown in Fig. 1 in the main article.

To analyze the effect of the double mutation T278D/A301G on the enzyme specificity for Arg, we generated and visualized the five lowest-energy structures from the K^* ensemble for this double mutant, as well as the five lowest-energy structures for the point mutations T278D and A301G. To facilitate the structural comparison, we used MolProbity [3] (Table S3). The comparison between the structures for A301G (which includes the WT Thr at position 278) and T278D/A301G suggests that the addition of the T278D mutation allows the side-chain of the Arg substrate to participate in stronger hydrogen bonding and/or electrostatics interactions with the charged carboxyl group of Asp at 278. The Arg substrate conformations in all five T278D/A301G structures fill the space that is otherwise occupied by the Ala side-chain at 301. In the T278D structures, the Arg atoms are pushed away from the A301 side-chain mainly through changes in the Arg side-chain dihedrals.

This mostly alleviates the Arg-A301 steric clashes (although in some structures of the single-point mutant these clashes are still significant), at the cost of introducing new significant clashes with other residues in the active site. Similarly to the Leu redesigns, the mutation A301G therefore appears to free up the space necessary to accommodate the Arg substrate. The second lowest-energy T278D/A301G structure from the K^* ensemble with Arg as substrate is shown in Fig. 2 in the main article.

Table S3: MolProbity [3] analysis for the [a]five lowest-energy T278D and T278D/A301G structures from the respective K^* ensembles with Arg as substrate. The [b]clashscore for a given structure is defined as the number of atom pairs with van der Waals overlap greater than 0.4 Å, per 1000 atoms; lower clashscores are better since they correspond to fewer significant clashes. For each structure, the list of [g]steric overlaps > 0.4 Å between an atom from the [e]Arg substrate and [f]atoms from residues in the enzyme active site are shown. For each structure, the [d]Arg side-chain dihedrals (in parenthesis) are shown along with the corresponding closest Arg rotamer [9].

Mutant	**Structure Rank**[a]	**Clashscore**[b]	**Arg Clashes**[c]			**Arg Rotamer**[d]
			Arg Atom[e]	*Other Atom*[f]	*Overlap* [g]	
T278D	1	18.02	NH2	CD1 (W239)	0.544	mmt180 (307,286.4,189,189)
			NH1	CB (A236)	0.417	
	2	22.52	NH1	HA (A236)	0.555	mmt-85 (289.6,283,189,284)
			HE	HB3 (A301)	0.550	
			HH21	CG (D278)	0.492	
			NH1	CG2 (I330)	0.484	
			NH1	HG23 (I330)	0.422	
	3	18.02	NH2	CD1 (W239)	0.544	mmt180 (307,286.4,189,189)
			NH1	CB (A236)	0.416	
	4	22.52	NH1	HA (A236)	0.551	mmt-85 (289.8,283,189,284)
			HE	HB3 (A301)	0.549	
			NH1	CG2 (I330)	0.487	
			HH21	CG (D278)	0.486	
			NH1	HG23 (I330)	0.427	
	5	24.02	HH1	CB (A301)	0.862	mmt85 (306.9,288.3,171,76)
			NH1	CB (A301)	0.610	
			NH2	CD2 (W239)	0.526	
			NH2	CG (W239)	0.520	
			NH2	CD1 (W239)	0.439	
			NH2	CE2 (W239)	0.414	
T278D/A301G	1	15.08	–	–	–	mtt85 (284.3,189,171,80.1)
	2	16.59	HH12	CG (D278)	0.401	mtt-85 (301.7,171,173,284)
	3	15.08	–	–	–	mtt85 (284.7,189,171,79.6)
	4	16.59	HH12	CG (D278)	0.402	mtt-85 (301.7,171,173.1,284)
	5	16.59	HH11	HA3 (G301)	0.406	mmt85 (301.5,284.7,179.3,77)

S2.2 Comparison to Other Methods and Evolution

It is interesting to compare our structure-based redesign predictions both to enzymes selected by evolution and to the predictions from alternative redesign methods. With this motivation in mind, we obtained the set of 1230 NRPS adenylation domains of both known and unknown specificity and their sequence alignment from [12]. We then compared the active sites of our experimentally-validated mutants (Table 1 in the main article) to the active sites of the 1230 domains [12], and determined that although the amino acid identities at mutated positions were found as constituents of longer signature sequences, none of our exact mutant active sites could be found in that domain set. This suggests that our structure-based method can successfully identify mutant GrsA-PheA active sites (specific for given target substrates) that have not been selected by natural evolution.

We further compared our structure-based predictions (Table 1 in the main article) to two sequence-based methods: the support vector machine (SVM) method of [12] and the phylogenetic method of [2] (Tables S4 and S5, respectively). The SVM predictor clusters sequences with similar substrate specificities into composite specificities. Both large clusters (a larger number of specificities grouped together) and small clusters (fewer specificities grouped together) were used by this method [12, Table 2]. The small-cluster predictor assigned a Phe-Trp specificity to all six mutants in Table 1(i) in the main article, and was thus unable to recognize the switch in specificity (confirmed by our experimental results) for the Leu-specific mutants. The large-cluster SVM predictor identified a composite specificity that contains Leu (as well as 6 other substrates) for T278L/A301G and T278M/A301G, and Asp for T278K/A301G, but failed to make a prediction about the other three mutants in Table 1(i) in the main article.

When using the method by Challis *et al.* [2], the top prediction of that method coincided with the measured target substrate specificity for three of the six mutants in Table 1(i) in the main article (T278L/A301G, T278M/A301G, and T278H/A301G). For A322V/A301G, the top prediction by this method was Phe, while a Leu domain ranked fifth in the prediction list. For T278D/A301G, this method identified Trp as its top prediction, while two Phe domains were ranked second and third, an Arg domain was ranked fourth in the list, and Lys was not predicted at all. For T278K/A301G, the top prediction was a Trp domain, while there was no Asp domain in the top forty predictions. Thus, although the method by Challis *et al.* performed reasonably well for the given set of mutations, our structure-based method was still able to predict substrate specificities missed by this sequence-based method.

Even when allowing only a small number of mutations to the enzyme active site, our structure-based redesign algorithm predicted novel mutants of GrsA-PheA (with improved target specificity) that are not yet found in nature and that were not predicted by alternative redesign methods. Our structure-based approach aims at identifying protein sequences that are predicted to have improved target substrate specificity (we refer to this as *positive design*). In addition to positive design, explicit *negative design* may be required, so that the design aims not only to improve the specificity for the target substrates, but also to destabilize the protein interactions with other substrates. In some cases, a serendipitous switch of specificity may be obtained without explicit negative design (e.g., our Leu mutants). In general, however, the lack of a negative design procedure may only yield an improvement in (but not a switch of) the target substrate specificity (e.g., our redesigns for charged substrates). However, the systematic incorporation of negative design poses

Table S4: Predictions from the SVM-based method using the [a]large clusters and [b]small clusters methods [12, 11]. *NP*, no prediction; Abu, 2-amino-butyric acid; Iva, isovaline; Aad, 2-amino-adipic acid.

Mutant	Large Clusters[a]	Small Clusters[b]
278L/301G	gly-ala-val-leu-ile-abu-iva	phe-trp
278M/301G	gly-ala-val-leu-ile-abu-iva	phe-trp
322V/301G	*NP*	phe-trp
278D/301G	*NP*	phe-trp
278H/301G	*NP*	phe-trp
278K/301G	asp-asn-glu-gln-aad	phe-trp

both significant computational and additional modeling challenges (e.g., backbone flexibility [5]) for structure-based approaches. In contrast, the sequence-based redesign methods compared here can be easily and efficiently applied to predict the specificity of a given protein sequence for a variety of substrates. It would be interesting to apply a hybrid approach that combines the strength of structure-based approaches for performing positive design with the ability of methods like [2, 12] to explicitly generate negative designs for a given sequence. Such an approach, combined with extensive experimental validation, could enhance our understanding of the natural selection processes for NRPS adenylation domains, and help further improve the accuracy of the *in silico* predictions.

S3 Experimental Protocol

S3.1 Primers

The mutagenesis primers used for the site-directed mutagenesis of mutant PheA are summarized in Table S6.

S3.2 Protein Expression and Purification

Fig. S2 shows the SDS-PAGE for the WT and mutant PheA.

S3.3 Steady-state Kinetics

Figures S3-S6 show the representative steady-state kinetics curves for the wild-type and mutant PheA. Each data point represents the initial rate v measured at one substrate concentration. The initial rate v was obtained by monitoring the increase of absorbance at 340 nm as a function of time (sec). The rate was plotted against amino acid concentration and fit directly into the Michaelis-Menten equation to derive k_{cat}, K_{M} and $k_{\mathrm{cat}}/K_{\mathrm{M}}$. Amino acid concentrations covering 0.2-5.0 K_{M} were used to obtain the hyperbolic curve. Each single reaction was repeated at least three times to obtain the standard deviation shown as the error bar of each point.

Table S5: Comparison of the active sites of our experimentally-validated mutants using the phylogenetic method of Challis *et al.* [2, 1]. This method tries to predict the specificity of an active site mutation by comparing its sequence to a database. For each of the [a]mutants, the top five matches are shown (ranked): the [b]protein and [c]module in the gene with the respective [d]activated amino acid substrate.

Mutant[a]	**Protein**[b]	**Module**[c]	**Substrate**[d]
278L/301G	Cyclosporine synthetase, CssA	3	Leu
	Cyclosporine synthetase, CssA	10	Leu
	Cyclosporine synthetase, CssA	2	Leu
	Cyclosporine synthetase, CssA	8	Leu
	Gramicidin synthetase A, GrsA	1	Phe
278M/301G	Cyclosporine synthetase, CssA	3	Leu
	Cyclosporine synthetase, CssA	10	Leu
	Cyclosporine synthetase, CssA	2	Leu
	Cyclosporine synthetase, CssA	8	Leu
	Cyclosporine synthetase, CssA	4	Val
322V/301G	Gramicidin synthetase A, GrsA	1	Phe
	tyrocidine synthetase 1, TycA	1	Phe
	CDA peptide synthetase I, Cda1	3	Trp
	enniatin sythetase, Esyn1	2	Val
	Microcistin synthetase B, McyC	1	Leu
278D/301G	CDA peptide synthetase I, Cda1	3	Trp
	Gramicidin synthetase A, GrsA	1	Phe
	tyrocidine synthetase 1, TycA	1	Phe
	Pyoverdin synthetase, PvdD	2	Arg
	tyrocidine synthetase 3, TycC	2	Gln
278H/301G	fengycin synthetase, FenA	2	Glu
	Fengycin synthetase, Pps4	2	Glu
	Fengycin synthetase, FenC	1	Glu
	Fengycin synthetase, Pps1	1	Glu
	Fengycin synthetase, FenE	1	Glu
278K/301G	CDA peptide synthetase I, Cda1	3	Trp
	Gramicidin synthetase A, GrsA	1	Phe
	tyrocidine synthetase 1, TycA	1	Phe
	tyrocidine synthetase 3, TycC	2	Gln
	Cyclosporine synthetase, CssA	3	Leu

Table S6: Mutagenesis primers.

T278L-f 5’- GTTATTTTGTTACCACCTACCTATGTAG -3’
T278L-r 5’- GGTAACAAAATAACAGTGATTTCCTTTTG G -3’
T278M-f 5’- GTTATTATGTTACCACCTACCTATGTAG -3’
T278M-r 5’- GGTAACATAATAACAGTGATTTCCTTTTGG -3’
T278H-f 5’- CTGTTATTCACTTACCACCTACCTATGTAG -3’
T278H-r 5’- GTGGTAAGTGAATAACAGTGATTTCCTTTTGG -3’
T278D-f 5’- CTGTTATTGACTTACCACCTACCTATGTAG -3’
T278D-r 5’- GTGGTAAGTCAATAACAGTGATTTCCTTTTG G -3’
T278K-f 5’- CTGTTATTAAGTTACCACCTACCTATGTAG -3’
T278K-r 5’- GTGGTAACTTAATAACAGTGATTTCCTTTTGG -3’
A322V-f 5’- CATAAATGTCTATGGCCCTACGGAAAC -3’
A322V-r 5’- GCCATAGACATTTATGTAAGTTAC -3’
S447N-f 5’- GAAGTTGAGAATATTCTTCTAAAGCATATG -3’
S447N-r 5’- GAATATTCTCAACTTCTTCTAGTTCAACTCG -3’
I277L-f 5’- GTTCTTTTGTTACCACCTACCTATGTAG -3’
I277L-r 5’- GGTAACAAAAGAACAGTGATTTCCTTTTGG -3’
V187L-f 5’- GCTTATCTTATTTATACTTCTGGTACAACAGGC -3’
V187L-r 5’- GTATAAATAAGATAAGCAAGATCGGTTGATTTACTTGG -3’

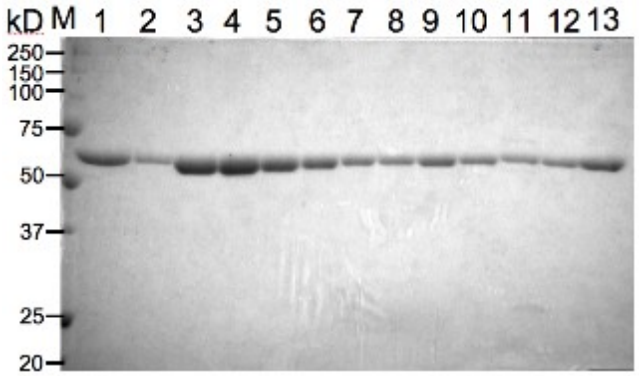

Figure S2: SDS-PAGE showing the homogeneity of the WT and mutant PheA. 1-2 μg of protein was loaded for each sample lane. M: Marker, 1: WT, 2: A301G, 3: T278L, 4: A301G/A322V, 5: T278M/A301G, 6: T278L/A301G, 7: T278L/A301G/S447N, 8: T278L/I277L/A301G, 9: V187L/T278L/A301G, 10: I277L/T278L/A301G/S447N, 11: T278D/A301G, 12: T278H/A301G, 13: T278K/A301G.

S3.4 Specificity Redesign

Figures S7 and S8 show, respectively, the relative and absolute substrate specificities for WT and mutant PheA with Phe and Leu. These figures are based on the data from Table 1 in the main article.

S3.5 Other Experimentally-tested Mutants

Table S7 presents the experimental results for several mutants not shown in Table 1 in the main article. These mutants are divided into three categories: (1) Computationally-predicted mutants that do not exhibit the desired specificity for a target substrate (Table S7i); (2) Single-point mutations that, by themselves, are not predicted by our algorithms, but that are components of other computationally-predicted mutants (Table S7ii); (3) Mutants predicted by our algorithms and selected for experimental verification that were difficult to purify due to solubility issues (Table S7iii). The double mutants T278E/A301G and T278D/I299E designed to bind Arg and Lys, respectively, both went into inclusion bodies in all expression conditions and could thus not be studied.

S3.6 Free Energy Calculation for WT and Mutant PheA

According to transition state theory, the activation energy, $\Delta \mathrm{G}_{\mathrm{T}}^{\ddagger}$ has two terms, an energetically unfavorable term $\Delta \mathrm{G}^{\ddagger}$, due to the chemical steps of bond making and breaking, and a compensating energetically favorable term $\Delta \mathrm{G}_{\mathrm{s}}$, due to the realization of the binding energy. That is,

$$\Delta \mathrm{G}_{\mathrm{T}}^{\ddagger} = \Delta \mathrm{G}^{\ddagger} + \Delta \mathrm{G}_{\mathrm{s}} \, . \tag{S3}$$

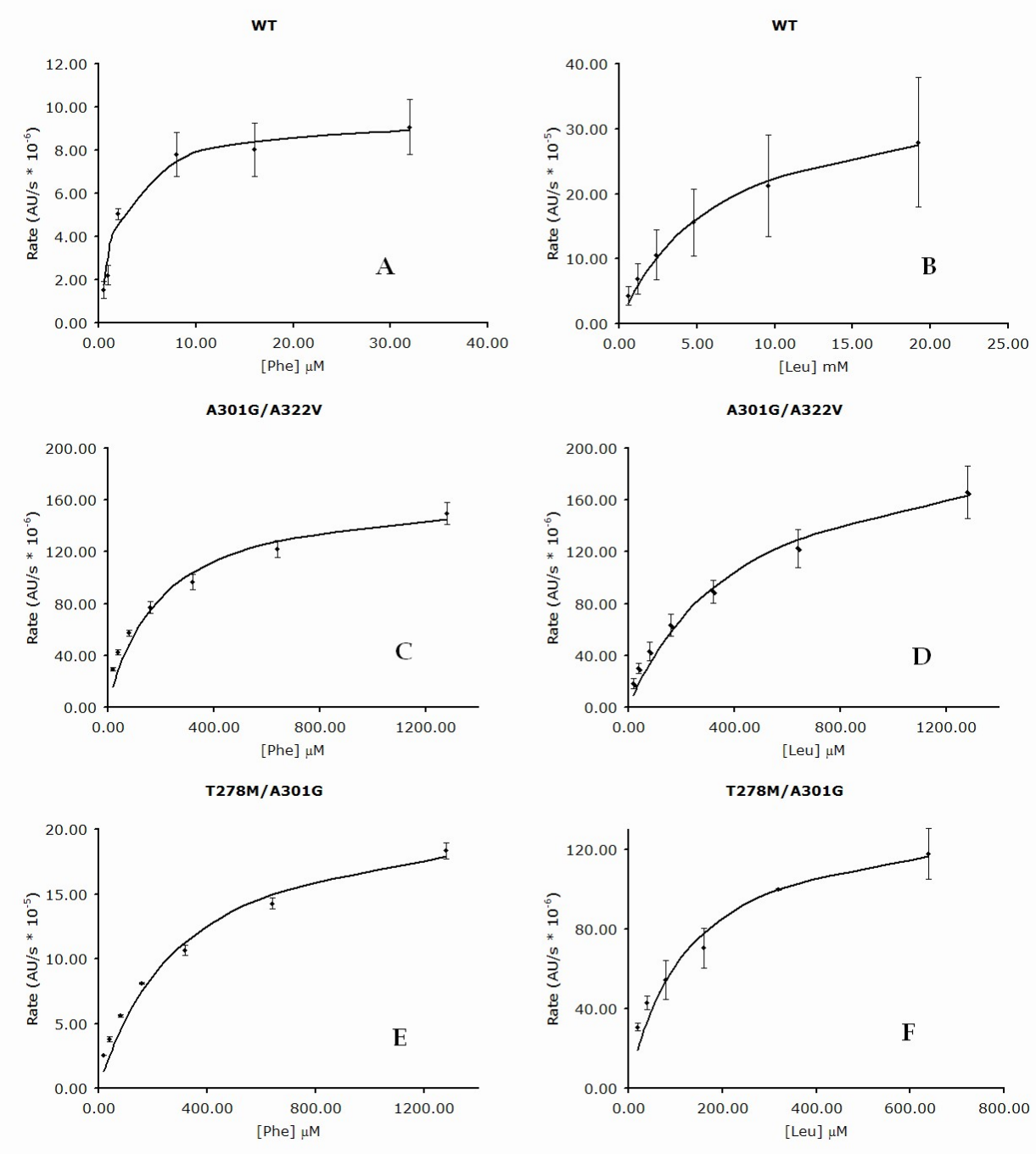

Figure S3: Representative steady-state kinetics curves for: WT PheA with Phe (**A**) and Leu (**B**), A301G/A322V with Phe (**C**) and Leu (**D**), and T278M/A301G with Phe (**E**) and Leu (**F**).

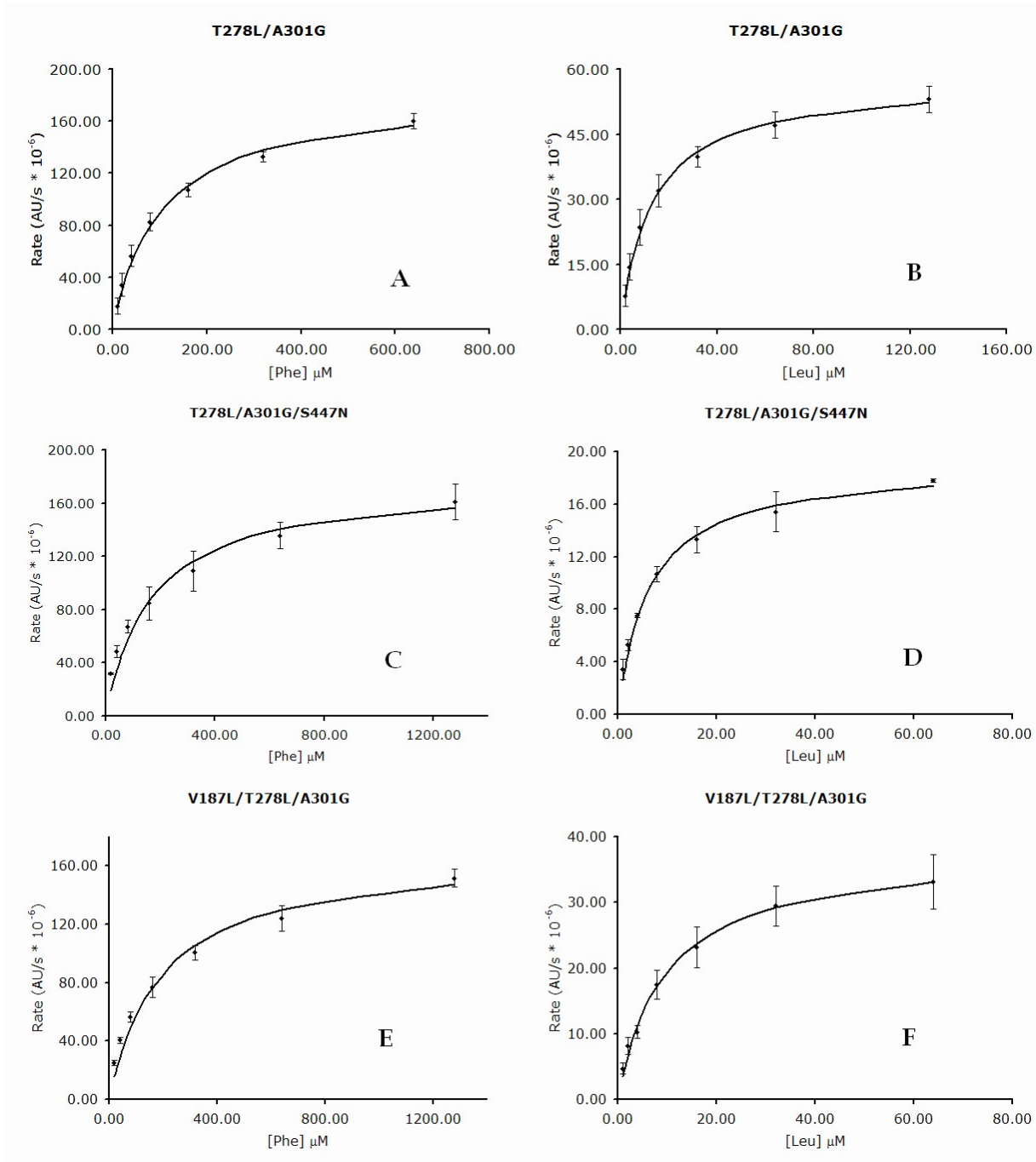

Figure S4: Representative steady-state kinetics curves for: T278L/A301G with Phe (**A**) and Leu (**B**), T278L/A301G/S447N with Phe (**C**) and Leu (**D**), and V187L/T278L/A301G with Phe (**E**) and Leu (**F**).

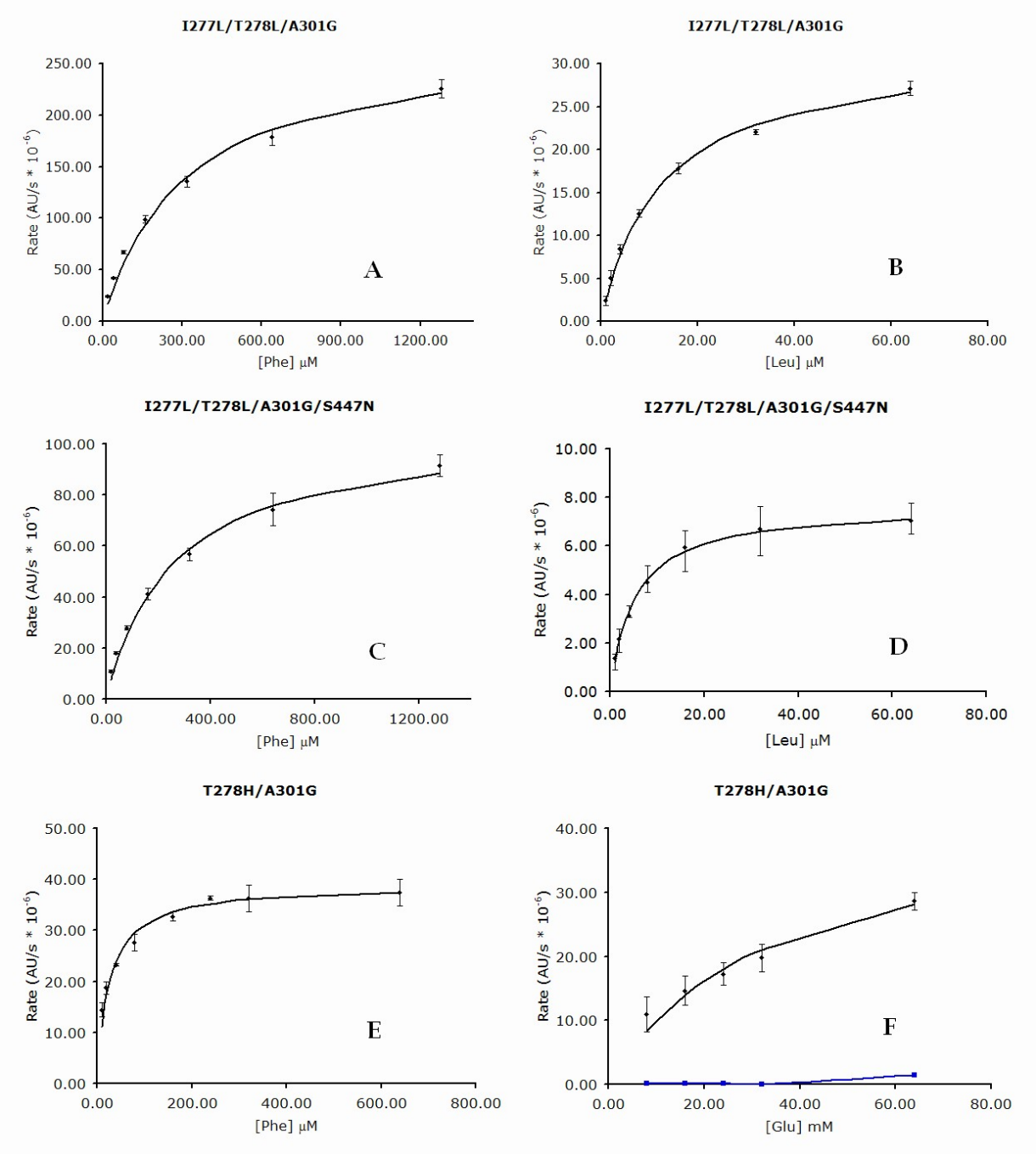

Figure S5: Representative steady-state kinetics curves (black) for I277L/T278L/A301G with Phe (**A**) and Leu (**B**), I277L/T278L/A301G/S447N with Phe (**C**) and Leu (**D**), and T278H/A301G with Phe (**E**) and Glu (**F**). For comparison, the steady-state kinetics curve for the WT enzyme with Glu (F, blue) is also shown.

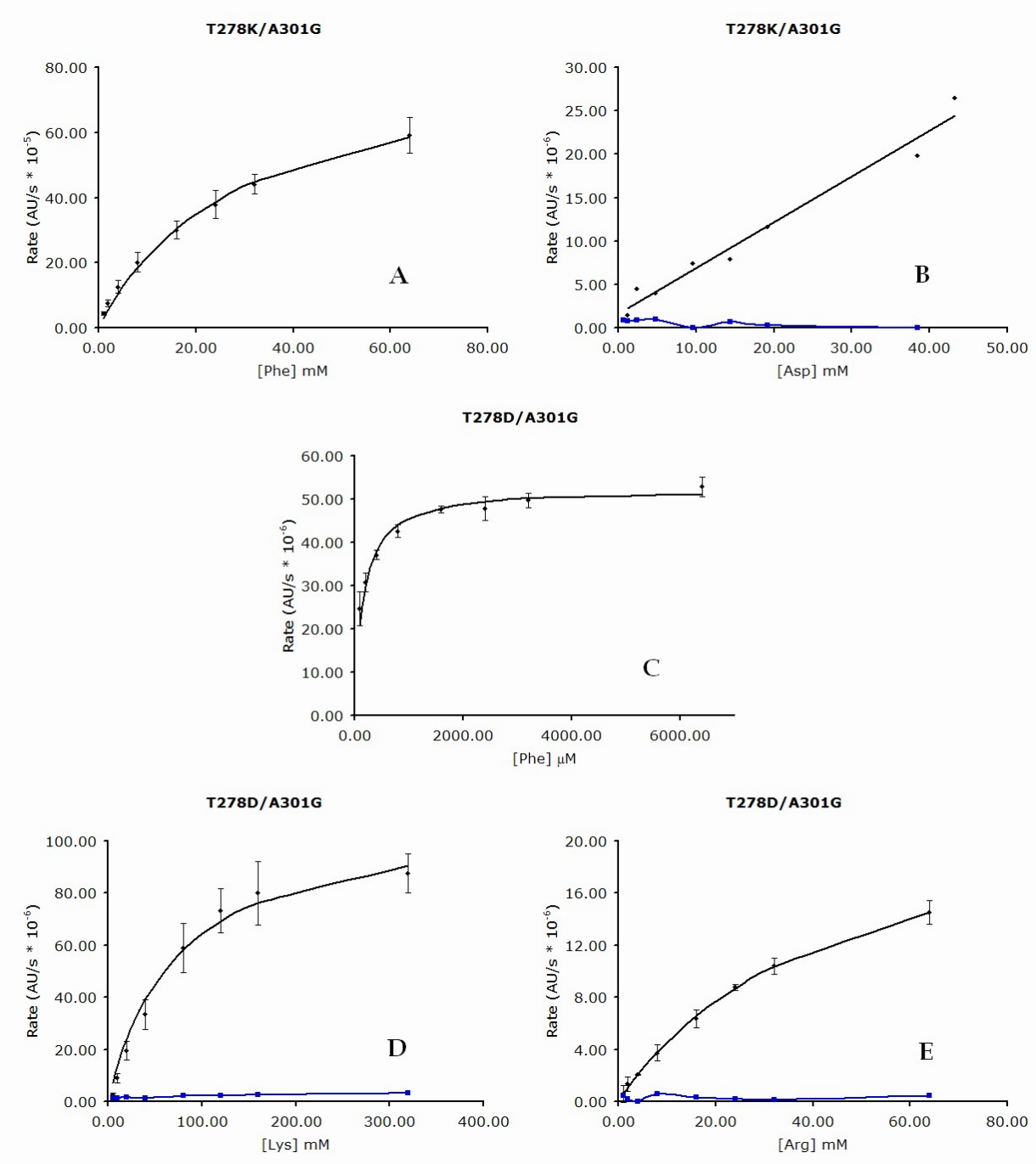

Figure S6: Representative steady-state kinetics curves (black) for T278K/A301G with Phe (**A**) and Asp (**B**), and T278D/A301G with Phe (**C**), Lys (**D**), and Arg (**E**). For comparison, the steady-state kinetics curve for the WT enzyme with Asp (B, blue), Lys (D, blue), and Arg (E, blue) are also shown.

Table S7: **Other experimentally-tested [a]mutants.** (i) The [d]k_{cat}, [e]K_M, and [f]k_{cat}/K_M for a set of mutants that showed little or no specificity for the target (Leu) vs. the WT (Phe) [b]substrate. The K^* [c]ranks for the respective mutants with Leu; [¶]this mutant was a top-ranked prediction by the algorithm of [8], but was pruned by the version of K^* described here. [§]Not detectable. [‡]K_M and k_{cat}/K_M cannot be accurately determined because the solubility of Phe ($\sim$180 mM in water) limits the reaction velocity under the experimental condition, in which the velocity remains linearly dependent on the concentration of the substrate. (ii) Single-point [a]mutants that are components of K^*-predicted double-mutants or of predictions from a GMEC-based approach [6]. The [d]k_{cat}, [e]K_M, and [f]k_{cat}/K_M for each mutant with the respective [b]substrate are shown. (iii) [a]Mutants difficult to purify due to [g]solubility issues are shown with their K^* [c]ranks and the target substrates.

	Mutant[a]	**Substrate**[b]	**Rank**[c]	k_{cat}[d] (min^{-1})	K_M[e] (mM)	k_{cat}/K_M[f] (mM^{-1} min^{-1})
	A301G/I330F	Leu	7	3.66	122.24	0.03
(i)	A301G/I330F	Phe		3.1	59.0	0.05
	A301G/I330W	Leu	¶	§	§	§
	A301G/I330W	Phe		$> 0.32^{‡}$		

	Mutant[a]	**Substrate**[b]	k_{cat}[d] (min^{-1})	K_M[e] (mM)	k_{cat}/K_M[f] (mM^{-1} min^{-1})
	A236M	Phe	8.1	0.0834	97.12
	A236M	Leu	21.1	217.2	0.097
(ii)	A322M	Phe	20.2	23.2	0.87
	A322M	Leu	9.4	640	0.015
	T278E	Arg	1.4	80.0	0.018

	Mutant[a]	**Rank**[c]	**Soluble**[g]
(iii)	T278E/A301G	3 (Arg)	no
	T278D/I299E	2 (Lys)	no

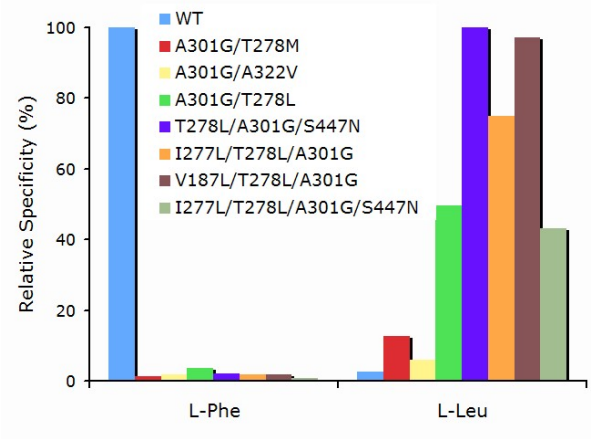

Figure S7: Relative substrate specificity for Phe (left) and Leu (right) of WT and mutant PheA. The specificity of all enzymes for each of the two substrates are normalized relative to the enzyme with the highest specificity for that substrate.

With the assumption that the active site residues and the side-chain of the amino acid substrate are not directly involved in the chemistry of the reaction, the difference of the activation energy is considered to be the difference in binding energy of the enzyme and transition state. Let $\Delta\Delta G_{\mathrm{Phe-Leu}}$ be the difference in binding energy between substrate Phe and Leu for WT and mutant PheA; let $\Delta\Delta G_{\mathrm{WT-Mut}}$ be the difference in binding energy between WT and mutants for Phe and Leu; and let $\Delta\Delta G_{\mathrm{int}}$ be the coupling energy (interaction energy) between T278L and A301G for Phe and Leu. Calculation of the free energy difference was done using equations [c], [d], and [e] in the caption of Table S8; the results are summarized in Table S8.

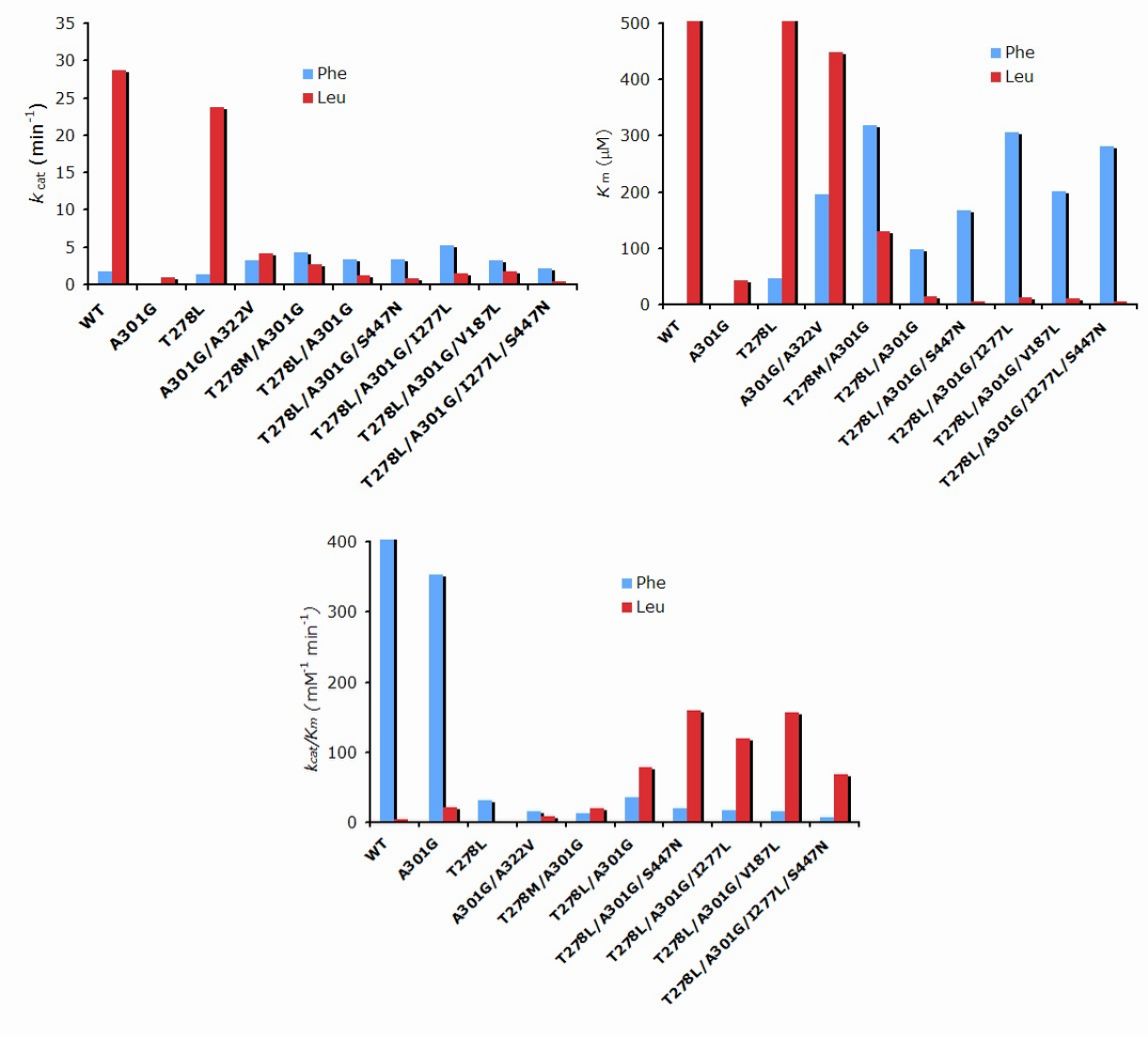

Figure S8: Absolute values of k_{cat} (top left), K_{M} (top right), and $k_{\mathrm{cat}}/K_{\mathrm{M}}$ (bottom) for Phe (cyan) and Leu (red) of WT and mutant PheA. Values of K_{M} for WT with Leu (6980 μM) and T278L with Leu (26280 μM) and value of $k_{\mathrm{cat}}/K_{\mathrm{M}}$ for WT with Phe (951 mM^{-1} min^{-1}) exceed the range displayed in the plot.

Table S8: Kinetic constants and free energy difference for [b]WT and mutant PheA with [a]Phe and Leu as substrates. [c]$\Delta\Delta G_{Phe-Leu} = -RT \ln \frac{(k_{cat}/K_M)_{Phe}}{(k_{cat}/K_M)_{Leu}}$; [d]$\Delta\Delta G_{WT-Mut} = -RT \ln \frac{(k_{cat}/K_M)_{WT}}{(k_{cat}/K_M)_{Mutant}}$; [e]$\Delta\Delta G_{int} = (\Delta\Delta G_{WT-A301G} + \Delta\Delta G_{WT-T278L}) - \Delta\Delta G_{WT-A301G/T278L}$, where R is the gas constant and T is the absolute temperature at 298.15 K. [§]For clarity, the values of $\Delta\Delta G_{Phe-Leu}$ for each enzyme are shown only once.

Substrate[a]	Enzyme[b]	k_{cat} (min^{-1})	K_M (mM)	k_{cat}/K_M (mM^{-1} min^{-1})	$\Delta\Delta G_{Phe-Leu}$[c] (kcal mol^{-1})	$\Delta\Delta G_{WT-Mut}$[d] (kcal mol^{-1})	$\Delta\Delta G_{int}$[e] (kcal mol^{-1})
Phe	WT	1.73 ± 0.29	0.0018 ± 0.0004	951.4 ± 111.2	-3.22[§]		
	A301G	0.20 ± 0.03	0.0005 ± 0.0001	353.7 ± 29.4	-1.66[§]	-0.59	
	T278L	1.41 ± 0.13	0.046 ± 0.009	30.74 ± 3.43	-2.09[§]	-2.03	
	A301G/T278L	3.37 ± 0.08	0.097 ± 0.013	34.94 ± 4.76	0.49[§]	-1.96	-0.67
Leu	WT	28.74 ± 1.58	6.98 ± 1.00	4.15 ± 0.36			
	A301G	0.89 ± 0.05	0.042 ± 0.008	21.48 ± 3.22		0.97	
	T278L	23.68 ± 2.30	26.28 ± 1.62	0.90 ± 0.04		-0.91	
	A301G/T278L	1.16 ± 0.10	0.015 ± 0.002	79.49 ± 13.67		1.75	-1.69

Supporting Information References

[1] G. Challis, J. Ravel, and C. Townsend. NRPS Predictive Blast Server. `http://www.tigr.org/jravel/nrps/blast/index2.html`, 2000. [Online; accessed 1 Sep 2008].

[2] G. Challis, J. Ravel, and C. Townsend. Predictive, structure-based model of amino acid recognition by nonribosomal peptide synthetase adenylation domains. *Chem. Biol.*, 7:211–224, 2000.

[3] I. Davis, A. Leaver-Fay, V. Chen, J. Block, G. Kapral, X. Wang, L. Murray, W. Arendall, J. Snoeyink, J. Richardson, and D. Richardson. MolProbity: all-atom contacts and structure validation for proteins and nucleic acids. *Nucleic Acids Res.*, 35:W375–383, Jul 2007.

[4] J. Desmet, M. De Maeyer, B. Hazes, and I. Lasters. The dead-end elimination theorem and its use in protein side-chain positioning. *Nature*, 356:539–542, 1992.

[5] I. Georgiev and B. R. Donald. Dead-end elimination with backbone flexibility. *Bioinformatics*, 23(13):i185–94, 2007.

[6] I. Georgiev, R. Lilien, and B. R. Donald. The minimized dead-end elimination criterion and its application to protein redesign in a hybrid scoring and search algorithm for computing partition functions over molecular ensembles. *J Comput Chem*, 29(10):1527–42, 2008.

[7] B. Kuhlman and D. Baker. Native protein sequences are close to optimal for their structures. *PNAS*, 97:10383–10388, 2000.

[8] R. Lilien, B. Stevens, A. Anderson, and B. R. Donald. A novel ensemble-based scoring and search algorithm for protein redesign, and its application to modify the substrate specificity of the Gramicidin Synthetase A phenylalanine adenylation enzyme. *J Comp Biol*, 12(6–7):740–761, 2005.

[9] S. C. Lovell, J. Word, J. Richardson, and D. Richardson. The penultimate rotamer library. *Proteins*, 40:389–408, 2000.

[10] N. Pierce, J. Spriet, J. Desmet, and S. Mayo. Conformational splitting: a more powerful criterion for dead-end elimination. *J. Comput. Chem.*, 21:999–1009, 2000.

[11] C. Rausch, T. Weber, O. Kohlbacher, W. Wohlleben, and D. H. Huson. NRPSPredictor - Algorithms in Bioinformatics. `http://www-ab.informatik.uni-tuebingen.de/software/NRPSpredictor/welcome.html`, 2005. [Online; accessed 1 Sep 2008].

[12] C. Rausch, T. Weber, O. Kohlbacher, W. Wohlleben, and D. H. Huson. Specificity prediction of adenylation domains in nonribosomal peptide synthetases (NRPS) using transductive support vector machines (TSVMs). *Nucleic Acids Res*, 33(18):5799–808, 2005.

[13] C. A. Voigt, S. L. Mayo, F. H. Arnold, and Z. G. Wang. Computational method to reduce the search space for directed protein evolution. *Proc Natl Acad Sci U S A*, 98(7):3778–83, 2001.