Article

Directed Evolution Reveals the Functional Sequence Space of an Adenylation Domain Specificity Code

Kurt Throckmorton, Vladimir Vinnik, Ratul Chawdhury, Taylor Cook, Marc G Chevrette, Costas Maranas, Brian Pfleger, and Michael George Thomas

ACS Chem. Biol., **Just Accepted Manuscript** • DOI: 10.1021/acschembio.9b00532 • Publication Date (Web): 20 Aug 2019

Just Accepted

is published by the American Chemical Society. 1155 Sixteenth Street N.W., Washington, DC 20036

DIRECTED EVOLUTION REVEALS THE FUNCTIONAL SEQUENCE SPACE OF AN ADENYLATION DOMAIN SPECIFICITY CODE

Kurt Throckmorton[1], Vladimir Vinnik[1], Ratul Chowdhury[2], Taylor Cook[3], Marc G. Chevrette[1,4], Costas Maranas[2], Brian Pfleger[3], Michael George Thomas[1#]

[1]Department of Bacteriology, University of Wisconsin-Madison, Madison, WI, USA 53706

[2]Department of Chemical Engineering, The Pennsylvania State University, University Park, PA, USA 16802

[3]Department of Chemical and Biological Engineering, University of Wisconsin-Madison, Madison, WI, USA 53706

[4]Department of Genetics, University of Wisconsin-Madison, Madison, WI, USA 53706

#Address correspondence to Michael George Thomas, michael.thomas@wisc.edu

K.T. and V.V. contributed equally to this work

ABSTRACT

Nonribosomal peptides are important natural products biosynthesized by nonribosomal peptide synthetases (NRPSs). Adenylation (A) domains of NRPSs are highly specific for the substrate they recognize. This recognition is determined by ten residues in the substrate-binding pocket, referred to as the specificity code. This finding led to the proposal that nonribosomal peptides could be altered by specificity code swapping. Unfortunately, this approach has proven, with few exceptions, to be unproductive; changing the specificity code typically results in broadened specificity or poor function. To enhance our understanding of A domain substrate selectivity, we carried out a detailed analysis of the specificity code from the A domain of EntF, an NRPS involved in enterobactin biosynthesis in *Escherichia coli*. Using directed evolution and a genetic selection, we determined which sites in the code have strict residue requirements and which are tolerant of variation. We showed that the EntF A domain, and other L-Ser-specific A domains, have a functional sequence space for L-Ser recognition, rather than a single code. This functional space is more expansive than the aggregate of all characterized L-Ser-specific A domains: we identified 152 new L-Ser specificity codes. Together, our data provide essential insights into how to overcome the barriers that prevent rational changes to A domain specificity.

KEY WORDS

Directed evolution, enterobactin, siderophore, substrate specificity, nonribosomal peptide synthetase, adenylation domain, natural product

INTRODUCTION

Nonribosomal peptides (NRPs) are common natural products of tremendous medical and agricultural importance. Assembly of these natural products by nonribosomal peptide synthetases (NRPSs) involves enzymology that functions analogously to an assembly line. Each amino acid is recognized and incorporated into the NRP by a set of enzymatic domains. Nature has derived the diversity of biological activities of NRPs by changing the number, order, and amino acid specificities of these domains.

The adenylation (A) domain of each module controls amino acid specificity, although the determinants of this specificity are not yet fully clear (*1*). The crystal structure of an L-Phe-bound A domain revealed 10 residues that interact with the substrate in the binding pocket (*2,3*). Analysis of the primary sequence of other A domains revealed a 'specificity code' which is highly conserved between A domains that recognize the same substrate (*3,4*). Later, it was shown that the specificity code is located within a subregion of the A domain, referred to hereafter as the recognition subdomain (RS), transfer of which may be responsible for differences in specificity between otherwise close homologs (*5, 6*). Through the identification of the RS and specificity code, substrate specificities can be inferred from the primary sequence of an A domain (*3,7*).

The discovery of the specificity code suggested that exchanging the code of one A domain with another would alter substrate selection allowing the production of new NRPs. While this approach seemed promising initially, early successes were likely enabled by the similarities of the A domains, specificity codes, and substrates that were exchanged (*3,8–10*). Attempts at less conservative changes resulted in variants with

substantially reduced catalytic efficiency (*11,12*). To date, structural biology and bioinformatics have failed to identify a clear path for reliably switching specificity.

Directed evolution has shown promise in overcoming limitations of rational specificity code design. Successive single-residue randomization and an *in vitro* enzymatic screen were used to identify three single-residue substitutions in the specificity code that result in the activation of a non-native substrate (*12*). Simultaneous randomization of three specificity code sites was used to alter the structure of the NRPS-derived antibiotic andrimid (*13*). Directed evolution has also been used with yeast cell surface display to switch substrate recognition (*14,15*). Additionally, others have used computational methods to guide changes to the substrate-binding pocket to switch amino acid specificity (*11,16*).

In this study, we used directed evolution of EntF, an NRPS involved in enterobactin (ENT) siderophore biosynthesis in *Escherichia coli* (Figure 1), to gain a deeper understanding of how the A domain specificity code confers substrate selectivity (*17–20*). Because siderophores are essential for Fe^{3+} acquisition in iron-limited conditions, this model system is ideal for performing selections or screens to process large DNA libraries. ENT has been used previously in directed evolution screens to identify functional chimeric A domains as well as to understand protein-protein interactions (*21–24*). Using a genetic selection and saturation mutagenesis to randomize the EntF specificity code, we discovered that there is an expansive functional sequence space, rather than a small number of discrete codes, for recognition of L-Ser. Additionally, we established that this sequence space is likely shared by other L-Ser-specific A domains. We characterized 157 unique EntF variants to assess the tolerance for residue variability at each site in the

specificity code and determined that only a few specificity code residues are required for substrate recognition. Our data are consistent with the conclusion that A domains have a strong bias for their native substrate that may complicate targeted specificity code swaps and require directed evolution approaches.

RESULTS AND DISCUSSION

Identification of novel L-Ser specificity codes. In Nature, the specificity code for an amino acid may fall into several distinct groups, e.g., L-Leu has 4 different code groups, (*3,27,28*). However, whether there is sequence flexibility, i.e. tolerance for variation, in the specificity code of a single A domain is not known. This knowledge gap has contributed to the lack of successful specificity code changes in efforts to reprogram NRPS enzymology. To address A domain specificity code flexibility, we targeted residues two through nine of the 10-site EntF specificity code for site-saturation mutagenesis (Figure 1D; Figure S1A). Asp649 and Lys952, in EntF notation, and hereafter referred to in specificity code notation as sites 1 and 10, were not randomized because they interact with the amino and carboxyl groups of the amino acid substrates and are highly conserved. For the remaining eight sites, we designed libraries 1 (L1) and 2 (L2) to randomize sites 2, 3, 4, 6, and 7 and sites 3, 4, 6, 7, and 9, respectively (Figure S1A). Based on sequence alignments of 82 characterized L-Ser-specific codes (Figure 1E; Table S1), site 5 appeared to be the most variable while site 8 is in almost always Val or Ile. Both sites are considered “wobble” sites for L-Ser-specific codes and are the only two sites without a ~90% predominant residue (*27*). Thus, to maximize our chances of identifying sequence variation, sites 5 and 8 (dark green in Figure 1C, 1D, and 1E) were

excluded from the initial libraries. We did not target residues outside the original ten-residue code (*3*) since these specificity code residues are the most likely to have direct interactions with the substrate and intermediates during catalysis, therefore having a higher probability of impacting substrate recognition. Additionally, L-Ser is the only amino acid competent to form the trilactone core of ENT thereby restriction our analysis to specificity codes that are competent for recognizing this amino acid at some level.

The *entF* specificity code mutant libraries were electroporated into an *E. coli* Δ*entF* strain and clones enabling ENT production were selected on iron-limiting media containing the iron chelator ethylenediamine-di(o-hydroxyphenylacetic acid), EDDHA. As little as 1 μM EDDHA inhibited growth of a Δ*entF* strain, thus, to capture EntF specificity code variants, including those significantly reduced in function, we used 1 μM EDDHA in all selections. After confirmation that the growth phenotype was conferred by the plasmid, the RS-encoding region of each ENT producer strain was sequenced (Table S2). The results from L1 and L2 suggested that His4, Ser6, and Asp9 (yellow, purple, and blue, respectively, in Figure 1C, 1D, and 1E) are nearly invariant, with one Ser4 exception (Figure 2).

To explore the possibility that a change at site 5 or site 8 is required for variation at site 4, 6, or 9, we designed library 3 to randomize sites 4, 5, 6, 8, and 9 (Figure S1A). Selections from L3 yielded only a single EntF variant, EntF 3-58. EntF 3-58 has His4, Ser6, and Asp9, reinforcing the requirement for these residues. We discovered that site 8 also has limited potential for variability; thus, this library contains only one highly variable site, dramatically reducing the proportion of viable EntF variants. In contrast, library 4 was designed to exclude sites 4, 6, and 9 and focus on the remaining, more flexible sites 2, 3,

5, 7, and 8. Consequently, L4 yielded the largest number of functional EntF variants. Considering all libraries, we identified 225 DNA-unique functional clones corresponding to 157 unique EntF-specificity code variants that included 26 residues not previously observed at their particular sites when compared to characterized L-Ser-specific A domains (*27*, Table S1).

Our dataset had a site-residue consensus for His4-Ser6-Asp9, with one Ser4 exception. Likewise, in Nature, among all characterized A domains, His4-Ser6-Asp9 is characteristic of L-Ser specificity; approximately 97% of A domains that have His4-Ser6-Asp9 are L-Ser specific. However, 11 other three-site-residue combinations, e.g. Val2-Phe5-Ser6, are just as, or more, characteristic of L-Ser specificity among A domains in Nature (*27*). Yet, for EntF, we experimentally found that only His4-Ser6-Asp9 are strictly required. The importance of His, Ser, and Asp is supported by a recent structure of EntF, crystallized with the catalytic-intermediate mimic, serine adenosine vinylsulfonamide (Ser-AVS), showing these residues in close proximity to the seryl moiety of the substrate (*29*). At site 8 we also observed residue usage similarities between the EntF variants and the characterized serine-specific codes: both contain Val and Ile in a roughly 55:45 proportion (Figure 2).

The residue usage between the EntF variants and the characterized L-Ser-specific A domains is less similar at sites 5 and 7 (dark green and light green, respectively, in Figure 1C, 1D, and 1E). In both datasets, site 5 is tolerant of a wide range of residues, yet more so in the EntF variants which have 15 allowed residues compared to seven for the codes found in Nature. Likewise, site 7 is more diverse in the EntF variants with 12 observed residues for the EntF variants compared to five in the natural codes (Figure 2).

The tolerance for amino acid variability at these two sites is consistent with the observation that both are distal to the substrate, particularly site 5 (Figure 1D). However, based only on the characterized codes, the tolerance for diversity at site 7 would appear approximately as strict as for sites 2, 3, 4, 6, and 9; thus, the variability observed among the EntF variants at site 7 was unexpected.

At sites 2 and 3 (light green in Figure 1C, 1D, and 1E), the EntF variants deviate significantly from the characterized A domains. In Nature, Val2 and Trp3 predominate, with diversity similar to sites 4, 6, 7, and 9. However, in the EntF variants, site 3 is highly variable with 12 allowed residues, and site 2 has an approximately even distribution of Val- and Pro-containing variants. The latter is highly unexpected because Pro is observed only once (at site 3) in the codes of naturally occurring L-Ser-specific A domains and is the second least-used code residue among all characterized A domains, after Arg (*27*). Furthermore, in the EntF structure, site 2 is proximal to the substrate; thus, tolerance of Pro2 was surprising.

Characterization of *in vivo* function and substrate specificity of EntF variants. To discriminate between the isolated ENT-producing strains *in vivo*, we screened each for growth in M9 liquid minimal media with EDDHA concentrations ranging from 50 μM to 850 μM. The maximum tolerated concentration of EDDHA for each strain is shown in Table S2 and is referred to as an MIC_{EDDHA}. Variants with more code differences from wild-type typically resulted in a lower MIC_{EDDHA}. However, this was not always the case as many strains had a low MIC_{EDDHA} despite only one or two residue differences, and several strains had a high MIC_{EDDHA} despite three to four differences. Thus, we found a

broad range of *in vivo* ENT production, approximated by MIC_{EDDHA}, among the isolated strains (Table S2).

To test whether the observed differences in MIC_{EDDHA} were due to broadened A domain specificity, we analyzed a subset of EntF variants spanning all libraries and all MIC_{EDDHA} tiers using the ATP/PP_i exchange assay (Table S3). A broadening of specificity would reduce ENT production due to the activation or aminoacylation of other amino acids not competent for ENT formation. Even the least active EntF variants were specific for L-Ser (Figure S2), though this may be due to our minimum threshold of growth at 50 μM EDDHA for detailed characterization. It is possible that variants initially isolated on iron-limited media with only 1μM EDDHA, but then failed to grow at 50 μM EDDHA, had broader substrate specificity. Future characterization of these variants may provide insights into how to overcome the inherent L-Ser specificity of the EntF A domain. Thus, the differences in MIC_{EDDHA} observed for the characterized ENT-producer strains is not due to broadened A domain specificity for non-serine substrates. Additionally, the levels of the co-purified MbtH-like (MLP) protein, YbdZ, important for *in vivo* ENT production and *in vitro* A domain activity (*20*), were very similar among the purified EntF variants, determined by immunoblotting done as previously reported (*30*). Therefore, the amino acid substitutions did not disrupt the EntF/YbdZ interactions that influence A domain function (*20,30,31*).

Next, we examined whether the EntF variants were impacted kinetically for L-Ser activation. ATP/PP_i exchange assays were used to determine the apparent K_m for L-Ser-binding and apparent V_{max} of a subset of variants (Table S3, Figure S3). With two exceptions, the variants ranged from 2% to 21% of the catalytic efficiency (V_{max}/K_m) of

wild-type EntF with predominantly a K_m effect. Overall, we found no correlation between either apparent K_m or apparent V_{max} and *in vivo* ENT production as measured by MIC_{EDDHA}. Two variants, 4-136 and 3-58, had higher catalytic efficiency than wild-type EntF. Interestingly, a strain expressing 3-58 only grew to 600 μM EDDHA, while one carrying 4-136 grew at the highest tested EDDHA concentration, as did the wild-type. The difference in MIC_{EDDHA} between strains carrying 3-58 and 4-136, as well as the lack of correlation between kinetic parameters and MIC_{EDDHA}, suggest that some other aspect of ENT biosynthesis beyond L-Ser recognition is impacted by the specificity code substitutions.

Molecular modeling of substrate binding. One such aspect is the binding of the Ser-AMP intermediate that is formed prior to transfer to the 4'-phosphopantetheinyl group on the thiolation domain (Figure 1B). To address this question, we used the recently solved crystal structure of EntF with a non-reactive Ser-AMP-phosphopantetheinyl intermediate mimic, Seryl-AVS (*29*), to model binding of Ser-AMP and L-Ser. Modeling of L-Ser in the binding pocket of the *in vitro*-characterized EntF variants revealed a correlation between the K_m and *in silico* CHARMM-based interaction energy scores, referred to as IPRO energy scores, for L-Ser, which account for non-covalent forces of interaction (van der Waals, electrostatics, and solvation). Furthermore, the K_m values correlated more strongly with the IPRO energy scores for L-Ser than for Ser-AMP, the catalytic intermediate (Figure 3A). This was expected since the K_m measures the binding of L-Ser, not Ser-AMP. However, the MIC_{EDDHA} correlated much more strongly with the IPRO energy scores for Ser-AMP than for L-Ser (Figure 3B). This correlation was also observed when considering all the variants (Figure 3C). These data suggest that EntF-Ser-AMP binding

is important for the function of EntF and that specificity code substitutions may impact this binding. For example, variant 4-19 has the highest IPRO energy score of all EntF variants and a high associated MIC_{EDDHA} of 600 μM, despite the third lowest V_{max}/K_m among the *in vitro*-characterized enzymes. Similarly, the relatively low Ser-AMP IPRO energy score of 3-58 could be the reason why its associated MIC_{EDDHA} is lower than both the wild-type and 4-136, despite similar kinetic parameters (Table S3).

Using the modeled EntF-Ser-AMP-bound complexes, we observed that most of the high-functioning variants (by MIC_{EDDHA}) have strong electrostatic interactions with the Ser-AMP intermediate. On the other hand, the low-functioning variants either exhibit weaker interactions or lack similar interactions with Ser-AMP altogether. Additionally, the low-functioning variants have more intra-enzyme electrostatic and hydrophobic interactions, typically distal to Ser-AMP (Figure 4A and 4B).

Characterization of residue usage in variant specificity codes. To test whether the diversity of codes was influenced by bias in library construction, we analyzed ~100 unselected clones from each library and determined that significant nucleotide usage biases occurred with a trend toward overrepresentation of Cs in sites 2–7 and Gs in sites 8 and 9. This is reflected in a significant skew in the predicted amino acid distributions for all sites except 7 and 9 (Table S4) with, at most, two-fold up or down changes relative to NNK proportions. These biases did not prevent underrepresented residues from emerging from the selection, with some of the most frequently observed residues at sites 3, 5, and 7, being underrepresented.

Using a chi-squared test, we compared the amino acid input frequencies in the libraries (based on the determined nucleotide usage bias) to the outputs of the selection

to determine whether selection occurred at each site and to determine whether the relative proportions of the observed residues matched the input. This comparison showed that selective pressure was exerted on each site, including the highly variable sites 3 and 5 (Figure 5A) and that in all cases the observed residue proportions deviate from expectation, indicating favor or disfavor by the selection (Figure 5B). We saw enrichment of Val2, Trp3, Val5, Leu7, and Ile8, which, except for Val5 and Ile8, are the wild-type residues. WT residues Ser5, His7, Phe7, and Val8, however, were all unenriched.

Potential explanations for enrichment include effects on *in vivo* production of ENT (MIC_{EDDHA}) or co-variation between sites. Using ANOVA, we examined whether certain residues were associated with higher or lower MIC_{EDDHA} on average. Across all libraries, only one site-residue combination, Trp3, had an average associated MIC_{EDDHA} significantly different than any other. It is likely that most codes function better with a bulky hydrophobic residue at the bottom of the substrate-binding pocket (Figure 4B). Notably, variants with Pro, among the rarest residues in all specificity codes (*27*), at site 2, were not outperformed by variants with Val2, the WT residue.

We analyzed all pairwise combinations of sites for deviation from an even distribution of the residues observed at one site amongst those observed at another, using a chi-squared test. This analysis revealed a skewed distribution, suggesting covariation, between sites 2 and 3, 2 and 5, and 3 and 8. (Table S5). Covariance between sites 2 and 3 and sites 3 and 8 can be rationalized due to their proximity in the binding pocket, however, covariance between sites 2 and 5 is more surprising (Figure 1D). We observed no correlation between covariance and MIC_{EDDHA} or residue enrichment. For example, both Val2 and Trp3 are enriched, however, the Val2/Trp3 pair is

underrepresented in the covariance. Patterns of residue frequency may simply reflect the extent to which different residues allow for possibilities at other sites, i.e., how many functional specificity code "solutions" exist given a particular set of residues at other sites.

Variant EntF codes function in non-EntF protein contexts. To determine whether the specificity codes we identified occur in other A domains, we searched 146,187 sequence-unique A domains from GenBank using SANDPUMA (*27*). 11,026 were predicted by multiple methods to be specific for L-Ser, providing a total of 23 unique L-Ser specificity codes. Among these, five different specificity codes match those found in the EntF variants. Two of these five codes, DVWHLSLIDK (3-58) and DVWHLSLVDK (4-213), are found in A domains that have been characterized and confirmed to be L-Ser specific (*32*–*34*). Three of these five codes, matching 4-136, K16A, and 4-54, had not been previously characterized in any A domain. The remaining 152 EntF variant codes, despite being biologically functional, do not match the specificity code of any A domain sequence in GenBank. To investigate whether this large number of unobserved codes is only relevant to EntF, we phylogenetically compared the A domains with codes matching the EntF variants to those of characterized L-Ser-specific A domains. We found that A domains with codes matching the EntF variants are not confined to the clades most closely related to EntF (Figure S4), suggesting that we were able to identify diverse L-Ser specificity codes.

We were interested in determining whether any uncharacterized A domains that have a specificity code that matches one of our EntF variant codes is specific for L-Ser in their native context. The EntF specificity code variant DVWHYSLVDK (4-136) is found in the A domain of DltA from *Paenibacillus donghaensis*. The A domain of DltA is 50.3%

identical to EntF across the RS region and 42.4% identical overall. We overproduced and purified the A-PCP from DltA in *E. coli*, assayed it by ATP/PP_i exchange, and determined that it activated only L-Ser (Figure 6). Thus, the EntF variants greatly expand the number of characterized unique L-Ser specificity codes, adding up to 155 to the previously known 23 (*27*), and can be used to confirm *in silico* A domain specificity predictions.

The five naturally-occurring codes that match the EntF variants differ from wild-type EntF at one or two sites. To test whether a less similar code could function in a non-EntF context, we changed the code of the DltA A domain to match the EntF variant 4-16, which differs from both EntF and DltA at five sites. We assayed the DltA 4-16 variant by ATP/PP_i exchange and found that it is specific for L-Ser (Figure 6). The activity of DltA is 27.2% that of EntF 4-136, while the activity of DltA 4-16 is 19.8% of EntF 4-16. This similarity suggests that the reduction in activity is primarily due to the differences between the two proteins rather than the 4-16 code in DltA. Thus, all of the EntF variant codes have the potential to activate L-Ser outside of an EntF context and the functional sequence space of EntF may also extend to other L-Ser-specific A domains.

In conclusion, our findings show that the EntF specificity code, and possibly any other, has the potential for variation greatly exceeding that which occurs in Nature. Despite the presumably relaxed selective pressure in a laboratory setting, the identification of a sequence space was surprising since all EntF proteins found to date in bacteria have the same specificity code. Even a recently identified group of EntF homologs from yeast, diverged over 60 million years, with just 56% identity with *E. coli* EntF (*35*), differ by only one code residue.

The variable tolerance for residue diversity among the specificity code sites along with the broad functional sequence space for L-Ser present several interesting possibilities for directed code swaps. First, our data suggest that specificity changes could be accomplished with minimal perturbations to the code by targeting key residues that confer specificity, e.g. His4, Ser6, and Asp9, in the case of EntF. Other residues, at more variable sites, could be left unchanged to preserve intra-protein interactions or be adjusted to best fit the target substrate. On the other hand, the sequence space of the EntF specificity code is consistent with the conclusion that some level of L-Ser activation, rather than of the desired substrate, would persist despite rational code swaps. Instead, to recognize a non-native substrate, a selection or screen would be necessary to overcome the improbability of choosing a code that is, first, outside of the sequence space for the native substrate, and, second, inside the sequence space for the non-native substrate. In summary, our data provide essential insights and suggest strategies that can be leveraged to overcome the barriers preventing rational changes to A domain substrate specificity.

METHODS

Plasmids and bacterial strains. Bacterial strains, plasmids, and primers are listed in Table S6. BW27749 Δ*entF E. coli* strain was constructed using pMAK705-*entF* as previously described (*36*). Plasmid pACYC184*entF*-ES was constructed with *entF* from *E. coli* MG1655 and *Eag*I and *Sac*I restriction sites flanking the RS-encoding region (Figure S1). Plasmids pCR-BluntII-TOPO*entF*-RS-F1–5 were constructed with the RS-encoding gene fragments, each cloned using the TOPO method (Invitrogen). Plasmid

pACYC184*entF*-ES-RS_{sub} was constructed by replacing the RS-encoding region of pACYC184*entF*-ES with placeholder *E. coli* DNA. Plasmids pACYC184*entF*-ES-L1–4 were constructed by amplification of the RS-encoding fragments in pCR-BluntII-TOPO*entF*-RS-F1–5 using NNK mutagenic primers, overlap-extension PCR, and ligation into pACYC184*entF*-ES-RS_{sub} (Figure S1). Selection-isolated plasmids were designated as "pACYC184*entF*-ES-variant#" corresponding to the strain number (Table S2). Plasmids pET28b*entF*-ES-variant#, pET28b*entF*-ES-wildtype, pET28b*dltA*-wildtype, and pET28b*dltA*-4-16 were constructed using polymerase incomplete primer extension, PIPE (*37*). Plasmid pACYC-duet-1 containing the *E. coli* MLP-encoding gene, *ybdZ*, was previously constructed (*20*).

EntF library creation. The RS-encoding fragments in pCR-BluntII-TOPO*entF*-RS-F1–5 were amplified using primers containing NNK codons corresponding to the residues targeted for mutagenesis. The mutagenized RS-encoding fragments were combined by overlap extension PCR and ligated into pACYC184*entF*-ES-RS_{sub} (Figure S1C). Ligations were electroporated into NEB 10β cells. Transformants were pooled and pACYC184*entF*-ES-L1–4 plasmids (Figure S1C), consisting of 2.5, 3.5, 2.5, and 7.5 million transformants, respectively, were recovered.

Selection and isolation of ENT producers. Plasmids pACYC184*entF*-ES-L1–4 were electroporated into BW27749 Δ*entF* cells and incubated on M9 minimal media noble agar plates with 0.4% (v/v) glycerol, 1 μM EDDHA (Complete Green Company), and antibiotic (chloramphenicol [34 μg mL^{-1}] or streptomycin [100 μg mL^{-1}]) (Figure S1D). Chloramphenicol was used for pACYC184*entF*-ES-L1–2, pACYC184*entF*-ES-L3–4 were switched to streptomycin to eliminate a chloramphenicol-resistant pACYC184*entF*-

wildtype contaminant. For technical reasons, libraries were not saturated. Most notably, the frequency of observation of wildtype contamination was much higher than that of successful transformants for library 3, which had the most restrictive combination of sites targeted for mutagenesis. This wildtype contamination was traced to the reversion of the *recA1* allele in commercially purchased competent cells, followed by recombination with the wildtype *entF* in the chromosome of these cells during library construction. ENT producer colonies were streaked for isolation on M9 plates with 0.4% (v/v) glycerol, 1 μM EDDHA, and antibiotic. From each, four colonies were incubated in a 96-well plate containing M9 with 0.4% glycerol (v/v), 50 μM EDDHA, and antibiotic. Plasmid preps were done from colonies that grew. Plasmids were screened by digest to eliminate pACYC184*entF*-wildtype (non-*Eag*I or *Sac*I). Each correct construct was re-transformed into BW27749 Δ*entF* and selected in liquid M9 with 50 μM EDDHA. Plasmid DNA was recovered, and the RS-encoding region was sequenced (Figure S1D).

Phenotypic characterization. MIC_{EDDHA} assays were done in technical triplicate. BW27749 Δ*entF* strains were grown overnight in LB, sub-cultured into LB, grown until an OD_{600} of ~0.5, normalized to an OD_{600} of 0.5, and 0.5 μL was inoculated into 200 μL of M9 with 0.4% glycerol (v/v), antibiotic, and 250, 350, 400, 450, 500, or 600 μM EDDHA. Cultures were inoculated through an Excel Scientific AeraSeal sterile membrane, covered with a second membrane, and incubated at 37 °C and 250 RPM for 45 hrs. If the OD_{600} rose to ≥ 0.2, from a calculated starting OD_{600} of 0.01, the strain was considered to have grown. Strains that grew at 600 μM EDDHA were tested at 600, 650, 700, 750, 800, and 850 μM EDDHA. Strains that did not grow at 250 μM EDDHA were tested at 50, 100, 150, and 200 μM EDDHA.

Sequence analysis of *entF* mutants. Each library was transformed into DH5α and the RS-encoding region of ~100 transformants per library was sequenced. A chi-squared test was used to check for statistically significant skews away from perfect randomization of the NNK sites in terms of the usage of nucleotides, codons, and the corresponding amino acids from. The nucleotide usage frequencies, combined across libraries but not sites, were used to make an adjusted genetic code. This genetic code was used to formulate the expectations for comparison to the selection output. For the remaining statistical analyses, the dataset of 225 DNA-unique *entF* mutants was considered. To determine if selection influenced the residues allowed at each site, the amino acid residue usage at each site from this dataset was compared to the adjusted genetic code, further adjusted for the exclusion of the stop codon, using a chi-squared test. To determine if selection influenced the relative proportions of the amino acids observed in the output, codes containing residues used fewer than five times at a given site were removed. This dataset was compared to the adjusted genetic code, adjusted for the exclusion of any unobserved residues, in addition to the stop codon, using a chi-squared test.

The amino acid diversity at each mutagenized specificity code site was compared between libraries and to a dataset of 82 characterized, L-Ser-specific A domains (*27*, Figure 2, and Table S1). SANDPUMA (*27*) was used to search the 146,187 sequence-unique A domains available in GenBank and to identify 11,026 A domains predicted to be L-Ser-specific using Active Site Motif (ASM), Support Vector Machine (SVM), and profile Hidden Markov Model (pHMM) methods. Codes extracted from this set were searched for the codes of the EntF variants (Table S2). To identify any covariance between specificity code sites, the DNA-unique L4 dataset, with codes containing rare residues or

non-code mutations removed, was used. For pairwise combinations of sites, this dataset was compared to the expectation of a proportional distribution of the possible substitutions at one site amongst those at each other site, using a chi-squared test. To detect the effect of specific residues at each site on the MIC_{EDDHA} of the associated strain, ANOVA was used in the Excel XLSTAT package with default parameters on the DNA-unique L4 dataset with codes containing rare residues or non-code mutations removed.

Construction of phylogenetic tree. MAFFT v7.310 (*38*) was used for multiple sequence alignment of the 82 characterized L-Ser-specific A domains (*27*, Table S1), several with specificity for L-Ser analogs or β-Ala (included as an outgroup), as well as 70 found in GenBank with predicted specificity codes matching several of the variants using default parameters. The alignment was manually trimmed and realigned, and a phylogenetic tree was constructed with FastTreev2.1.9 (*39*) with default parameters, rooted on AAG02364.1 and edited using FigTree v.1.4.2 (http://tree.bio.ed.ac.uk/software/figtree/).

Overproduction and purification of EntF and DltA variants. EntF variants (Table S3) were co-overproduced in *E. coli* BL21(DE3) with a C-terminal hexahistidine tag with the MLP, YbdZ. DltA proteins were overproduced in *E. coli* BL21(DE3) *ybdZ::acc(3)IV* with a C-terminal hexahistidine tag. All protein purifications were done as previously described (*20,31,40*). Protein concentration was determined by BCA assay (Pierce).

Radiolabeled ATP/PP_i assays of the EntF and DltA variants. ATP/PP_i exchange assays were performed as previously described (*31,40*). Variants (Table S3) were assayed in duplicate for substrate activation against four pools of five amino acids, Pool 1: SATPG, Pool 2: VLIMC, Pool 3: NDEQF, Pool 4: YWHKR, and then with individual amino acids from Pool 1. The apparent K_m and apparent V_{max} of the variants were

determined in the linear range for product formation and 8 min reaction time (Figure S3), and calculated using non-linear regression analysis (GraphPad Prism ver. 6.0h).

IPRO energy score calculations. Computational models of wild-type EntF and 157 variants were constructed in complex with the L-Ser or Ser-AMP (substrates). The reported structure of EntF (*29*, PDB 5JA1) with serine adenosine vinylsulfonamide (Ser-AVS) was used as the model. Energy-minimized structures of the EntF variants were generated using the Mutator module of the Iterative Protein Redesign and Optimization Suite of programs, IPRO (*41*). The complexes were energy minimized using CHARMM-force field (*42*). The CHARMM-based interaction energy scores (or IPRO energy scores) between a variant and substrate was computed as a sum of pairwise additive, non-bonded energy terms accounting for (a) van der Waal's, (b) electrostatics, and (c) implicit solvation using the Generalized-Born implicit solvation method (*43*). This is conceptually akin to RosettaLigand (*44*) which reports Rosetta scores for the non-covalent forces between enzyme and ligand at the binding pocket as an *in silico* analogue of substrate affinity. Following side-chain conformation alterations done in Mutator, the location of substrates was re-adjusted using improved rigid-body docking (*45*) by randomly perturbing substrates along and around the *X, Y,* and *Z* axes using a Gaussian distribution centered at zero, with standard deviations of 0.2, 0.2, and 2.0 Å, respectively. 500 iterations are performed with subsequent interaction energy (binding score) re-calculation.

SUPPORTING INFORMATION

The supporting information is available free of charge at http://pubs.acs.org. The supporting information contains the following: a graphical summary of the experimental procedures, additional data from *in vitro* enzyme assays, phylogenetic analysis, lists of all naturally occurring characterized L-Ser-specific A domains as well as all isolated ENT producer strains, library amino acid usage bias, covariance, and all information regarding strains, primers, and plasmids.

AUTHOR CONTRIBUTIONS

K.T., V.V., B.P., and M.G.T. designed the project and analyzed the data. K.T., V.V., and T.C. carried out selection and *in vitro* experiments. K.T. performed statistical analyses. R.C. performed molecular modeling experiments and R.C. and C.M. analyzed the data. K.T. and M.G.C. performed bioinformatics analyses and M.G.C. provided A domain sequence information. K.T. and V.V. wrote the manuscript. All authors contributed to proofreading the manuscript.

ACKNOWLEDGEMENTS

This work was supported in part by the National Science Foundation (NSF 1716594 to M.G.T. and B.P.), the National Institutes of Health (GM100346 to M.G.T.), and the Great Lakes Bioenergy Research Center, U.S. Department of Energy, Office of Science, Office of Biological and Environmental Research (DOE BER Office of Sciences DE-SC0018409). V.V. was supported in part by the Jerome J. Stefaniak Predoctoral Fellowship. K.T. was supported in part by funds supplied by the E. B. Fred Professorship

(M.G.T). T.C. is the recipient of a National Institutes of Health Biotechnology Training Program (NIGMS 5 T32 GM08349). We thank C. Ané for statistical consultation.

DECLARATION OF INTERESTS

The authors declare no competing interests.

FIGURE LEGENDS

Figure 1 Enterobactin (ENT) formation and the structure, specificity code, and function of the A domain of EntF. A) Diagram of the ENT biosynthetic pathway. EntE tethers 2,3-dihydroxybenzoic acid (DHB) to EntB. The condensation (C) domain of EntF condenses DHB and L-Ser, previously activated and bound to the thiolation (T) domain of EntF. After one turnover, the DHB-L-Ser monomer is stored on the thioesterase (TE) domain. After three iterations of this process, the final product is cyclized and released. B) The two half-reactions of the EntF A domain. In the first, the A domain activates L-Ser as Seryl-AMP (1) and transfers the seryl group to the T domain (2) in the second. C) Ribbon representation of the EntF (PDB5JA1) A domain with the recognition subdomain, in red, and specificity code residues, highlighted in colors matching D and E. D) Binding pocket residues that form the specificity code and the reaction intermediate-mimic inhibitor, seryl-AVS (*25*), used for crystallization. E) WebLogo (*26*) of specificity code sites 1–10 for 82 characterized L-Ser codes (*27*).

Figure 2 Residue usage by code site in the EntF variants and characterized L-Ser-specific A domains. Pie charts describing, by site in the specificity code, the amino acid usage across all libraries. The wild-type EntF residues are shown in the top row followed by the 82 characterized L-Ser-specific A domain residues, and, subsequently, each of the four libraries. Circles with diagonal lines designate non-mutagenized sites. The color of each sector corresponds to the key at the bottom; both are organized alphabetically, clockwise and left to right. For each library, n designates the number of isolated strains containing DNA-unique *entF* mutants, provided they had no non-code residue substitutions. All Libraries n= 200, 216, 50, 168, 50, 216, 168, and 17 for sites 2–9, respectively.

Figure 3 *in silico* IPRO energy scores of EntF variants for L-Ser correlate with K_m, binding for Ser-AMP correlates with MIC_{EDDHA}. Correlation between the IPRO energy score with L-Ser or Ser-AMP and A) K_m and B) the associated MIC_{EDDHA} of *in vitro*-characterized EntF variants (Table S3). A more negative value for IPRO energy score indicates stronger binding. C) Correlation between IPRO energy score with Ser-AMP by each EntF variant (blue) and the MIC_{EDDHA} of the corresponding *entF* mutant strain (green). IPRO energy scores and MIC_{EDDHA} are normalized to wild-type.

Figure 4 Substrate-binding pocket interactions differ between high- and low-functioning variants. Diagrams of the substrate-binding pockets of (A) WT EntF, EntF variant 4-71, and (B) EntF variant 106 highlighting differences between their energy-

minimized structures, based on PDB 5JA1, in the electrostatic interactions of the code residues and Ser-AMP (dashed line; side view – top row) and the intra-enzyme interactions (green line; bottom view – bottom row). Each diagram also contains specificity code residues 2–9 (light pink with blue font), the Ser-AMP intermediate (yellow), and residues differing from WT (black font).

Figure 5 Preferential residue usage is observed at all sites. A) Comparison of amino acid residue proportions at sites 3 and 5 between the input and selection output. Each sector color corresponds to an amino acid residue according to the key at the bottom and sectors are sorted, clockwise from the top, by contribution to the chi-squared statistic. B) Comparison of the residue proportions at each mutagenized site in the L4 dataset to the input. The color of each sector corresponds to the key at bottom, both of which are organized alphabetically, clockwise and left to right, respectively.

Figure 6 A variant EntF specificity code, differing at all five mutagenized sites, functions in a non-EntF context. Comparison of the *in vitro* activity of purified proteins EntF WT, EntF variant 4-16, DltA WT, and DltA variant 4-16 with pooled amino acid substrates (A) and the individual amino acids contained in pool 1 (B) as determined by ATP-PP_i exchange assay. The color key in panel A applies to both panels. Pool 1: SATPG, 2: VILMC, 3: NDEQF, 4: YWHKR.

REFERENCES

[1] Gevers, W., Kleinkauf, H., and Lipmann, F. (1968) The activation of amino acids for biosynthesis of gramicidin S. *Proc. Natl. Acad. Sci. U. S. A. 60*, 269–276.

[2] Conti, E., Stachelhaus, T., Marahiel, M. A., and Brick, P. (1997) Structural basis for the activation of phenylalanine in the non-ribosomal biosynthesis of gramicidin S. *EMBO J. 16*, 4174–4183.

[3] Stachelhaus, T., Mootz, H. D., and Marahiel, M. A. (1999) The specificity-conferring code of adenylation domains in nonribosomal peptide synthetases. *Chem. Biol. 6*, 493–505.

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

[5] Crusemann, M., Kohlhaas, C., and Piel, J. (2013) Evolution-guided engineering of nonribosomal peptide synthetase adenylation domains. *Chem. Sci. 4*, 1041–1045.

[6] Kries, H., Niquille, D. L., and Hilvert, D. (2015) A subdomain swap strategy for reengineering nonribosomal peptides. *Chem. Biol. 22*, 640–648.

[7] Blin, K., Shaw, S., Steinke, K., Villebro, R., Ziemert, N., Lee, S. Y., Medema, M. H., and Weber, T. (2019) antiSMASH 5.0: updates to the secondary metabolite genome mining pipeline. *Nucleic Acids Res*. *47*, W81–W87.

[8] Eppelmann, K., Stachelhaus, T., and Marahiel, M. A. (2002) Exploitation of the selectivity-conferring code of nonribosomal peptide synthetases for the rational design of novel peptide antibiotics. *Biochemistry 41*, 9718–9726.

[9] Thirlway, J., Lewis, R., Nunns, L., Al Nakeeb, M., Styles, M., Struck, A. W., Smith, C. P., and Micklefield, J. (2012) Introduction of a non-natural amino acid into a nonribosomal peptide antibiotic by modification of adenylation domain specificity. *Angew. Chem. Int. Ed. Engl. 51*, 7181–7184.

[10] Kries, H., Wachtel, R., Pabst, A., Wanner, B., Niquille, D., and Hilvert, D. (2014) Reprogramming nonribosomal peptide synthetases for "clickable" amino acids. *Angew. Chem. Int. Ed. Engl. 53*, 10109–10108.

[11] Chen, C. Y., Georgiev, I., Anderson, A. C., and Donald, B. R. (2009) Computational structure-based redesign of enzyme activity. *Proc. Natl. Acad. Sci. U. S. A. 106*, 3764–3769.

[12] Villiers, B., and Hollfelder, F. (2011) Directed evolution of a gatekeeper domain in nonribosomal peptide synthesis. *Chem. Biol. 18*, 1290–1299.

[13] Evans, B. S., Chen, Y., Metcalf, W. W., Zhao, H., and Kelleher, N. L. (2011) Directed evolution of the nonribosomal peptide synthetase AdmK generates new andrimid derivatives in vivo. *Chem. Biol. 18*, 601–607.

[14] Zhang, K., Nelson, K. M., Bhuripanyo, K., Grimes, K. D., Zhao, B., Aldrich, C. C., and Yin, J. (2013) Engineering the substrate specificity of the DhbE adenylation domain by yeast cell surface display. *Chem. Biol. 20*, 92–101.

[15] Niquille, D. L., Hansen, D. A., Mori, T., Fercher, D., Kries, H., and Hilvert, D. (2018) Nonribosomal biosynthesis of backbone-modified peptides. *Nat. Chem. 10*, 282–287.

[16] Stevens, B. W., Lilien, R. H., Georgiev, I., Donald, B. R., and Anderson, A. C. (2006) Redesigning the PheA domain of gramicidin synthetase leads to a new understanding of the enzyme's mechanism and selectivity. *Biochemistry 45*, 15499–15504.

[17] Pettis, G. S., and McIntosh, M. A. (1987) Molecular characterization of the Escherichia coli enterobactin cistron entF and coupled expression of entF and the fes gene. *J. Bacteriol. 169*, 4154–4162.

[18] Reichert, J., Sakaitani, M., and Walsh, C. T. (1992) Characterization of EntF as a serine-activating enzyme. *Protein Sci. 1*, 549–556.

[19] Gehring, A. M., Mori, I., and Walsh, C. T. (1998) Reconstitution and characterization of the Escherichia coli enterobactin synthetase from EntB, EntE, and EntF. *Biochemistry 37*, 2648–2659.

[20] Felnagle, E. A., Barkei, J. J., Park, H., Podevels, A. M., McMahon, M. D., Drott, D. W., and Thomas, M. G. (2010) MbtH-like proteins as integral components of bacterial nonribosomal peptide synthetases. *Biochemistry 49*, 8819–8817.

[21] Fischbach, M. A., Lai, J. R., Roche, E. D., Walsh, C. T., and Liu, D. R. (2007) Directed evolution can rapidly improve the activity of chimeric assembly-line enzymes. *Proc. Natl. Acad. Sci. U. S. A. 104*, 11951–11956.

[22] Lai, J. R., Fischbach, M. A., Liu, D. R., and Walsh, C. T. (2006) A protein interaction surface in nonribosomal peptide synthesis mapped by combinatorial mutagenesis and selection. *Proc. Natl. Acad. Sci. U. S. A. 103*, 5314–5319.

[23] Zhou, Z., Lai, J. R., and Walsh, C. T. (2006) Interdomain communication between the thiolation and thioesterase domains of EntF explored by combinatorial mutagenesis and selection. *Chem. Biol. 13*, 869–879.

[24] Zhou, Z., Lai, J. R., and Walsh, C. T. (2007) Directed evolution of aryl carrier proteins in the enterobactin synthetase. *Proc. Natl. Acad. Sci. U. S. A. 104*, 11621–11626.

[25] Drake, E. J., Miller, B. R., Shi, C., Tarrasch, J. T., Sundlov, J. A., Allen, C. L., Skiniotis, G., Aldrich, C. C., and Gulick, A. M. (2016) Structures of two distinct conformations of holo-non-ribosomal peptide synthetases. *Nature 529*, 239–238.

[26] Crooks, G. E., Hon, G., Chandonia, J. M., and Brenner, S. E. (2004) WebLogo: a sequence logo generator. *Genome Res. 14*, 1188–1190.

[27] Chevrette, M. G., Aicheler, F., Kohlbacher, O., Currie, C. R., and Medema, M. H. (2017) SANDPUMA: ensemble predictions of nonribosomal peptide chemistry reveal biosynthetic diversity across Actinobacteria. *Bioinformatics 33*, 3202–3210.

[28] Khayatt, B. I., Overmars, L., Siezen, R. J., and Francke, C. (2013) Classification of the adenylation and acyl-transferase activity of NRPS and PKS systems using ensembles of substrate specific hidden Markov models. *PLoS One 8*, e62136.

[29] Miller, B. R., Drake, E. J., Shi, C., Aldrich, C. C., and Gulick, A. M. (2016) Structures of a nonribosomal peptide synthetase module bound to MbtH-like proteins support a highly dynamic domain architecture. *J. Biol. Chem. 291*, 22559–22571.

[30] Schomer, R. A., Park, H., Barkei, J. J., and Thomas, M. G. (2018) Alanine scanning of YbdZ, an MbtH-like protein, reveals essential residues for functional Interactions with its nonribosomal peptide synthetase partner EntF. *Biochemistry 57*, 4129–4134.

[31] Schomer, R. A., and Thomas, M. G. (2017) Characterization of the functional variance in MbtH-like protein interactions with a nonribosomal peptide synthetase. *Biochemistry 56*, 5380–5390.

[32] Guenzi, E., Galli, G., Grgurina, I., Gross, D. C., and Grandi, G. (1998) Characterization of the syringomycin synthetase gene cluster. A link between prokaryotic and eukaryotic peptide synthetases. *J. Biol. Chem. 273*, 32857–32863.

[33] Li, W., Rokni-Zadeh, H., De Vleeschouwer, M., Ghequire, M. G., Sinnaeve, D., Xie, G. L., Rozenski, J., Madder, A., Martins, J. C., and De Mot, R. (2013) The antimicrobial compound xantholysin defines a new group of *Pseudomonas* cyclic lipopeptides. *PLoS One 8*, e62946.

[34] Kodani, S., Bicz, J., Song, L., Deeth, R. J., Ohnishi-Kameyama, M., Yoshida, M., Ochi, K., and Challis, G. L. (2013) Structure and biosynthesis of scabichelin, a novel tris-hydroxamate siderophore produced by the plant pathogen *Streptomyces scabies* 87.22. *Org. Biomol. Chem. 11*, 4686–4694.

[35] Kominek, J., Doering, D. T., Opulente, D. A., Shen, X. X., Zhou, X., DeVirgilio, J., Hulfachor, A. B., Groenewald, M., Mcgee, M. A., Karlen, S. D., Kurtzman, C. P.,

Rokas, A., and Hittinger, C. T. (2019) Eukaryotic acquisition of a bacterial operon. *Cell 176*, 1356–1366.e1310.

[36] Hamilton, C. M., Aldea, M., Washburn, B. K., Babitzke, P., and Kushner, S. R. (1989) New method for generating deletions and gene replacements in *Escherichia coli*. *J. Bacteriol. 171*, 4617–4622.

[37] Klock, H. E., Koesema, E. J., Knuth, M. W., and Lesley, S. A. (2008) Combining the polymerase incomplete primer extension method for cloning and mutagenesis with microscreening to accelerate structural genomics efforts. *Proteins 71*, 982–994.

[38] Katoh, K., and Standley, D. M. (2013) MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol. Biol. Evol. 30*, 772–780.

[39] Price, M. N., Dehal, P. S., and Arkin, A. P. (2010) FastTree 2--approximately maximum-likelihood trees for large alignments. *PLoS One 5*, e9490.

[40] McMahon, M. D., Rush, J. S., and Thomas, M. G. (2012) Analyses of MbtB, MbtE, and MbtF suggest revisions to the mycobactin biosynthesis pathway in *Mycobacterium tuberculosis*. *J. Bacteriol. 194*, 2809–2818.

[41] Pantazes, R. J., Grisewood, M. J., Li, T., Gifford, N. P., and Maranas, C. D. (2015) The Iterative Protein Redesign and Optimization (IPRO) suite of programs. *J. Comput. Chem. 36*, 251–263.

[42] Brooks, B. R., Brooks, C. L., Mackerell, A. D., Nilsson, L., Petrella, R. J., Roux, B., Won, Y., Archontis, G., Bartels, C., Boresch, S., Caflisch, A., Caves, L., Cui, Q., Dinner, A. R., Feig, M., Fischer, S., Gao, J., Hodoscek, M., Im, W., Kuczera, K., Lazaridis, T., Ma, J., Ovchinnikov, V., Paci, E., Pastor, R. W., Post, C. B., Pu, J. Z., Schaefer, M., Tidor, B., Venable, R. M., Woodcock, H. L., Wu, X., Yang, W., York, D. M., and Karplus, M. (2009) CHARMM: the biomolecular simulation program. *J. Comput. Chem. 30*, 1549–1614.

[43] Lu, B., Zhang, D., and McCammon, J. A. (2005) Computation of electrostatic forces between solvated molecules determined by the Poisson-Boltzmann equation using a boundary element method. *J. Chem. Phys. 122*, 214102.

[44] Meiler, J., and Baker, D. (2006) ROSETTALIGAND: protein-small molecule docking with full side-chain flexibility. *Proteins 65*, 538–548.

[45] Chowdhury, R., Ren, T., Shankla, M., Decker, K., Grisewood, M., Prabhakar, J., Baker, C., Golbeck, J. H., Aksimentiev, A., Kumar, M., and Maranas, C. D. (2018) PoreDesigner for tuning solute selectivity in a robust and highly permeable outer membrane pore. *Nat. Commun. 9*, 3661.

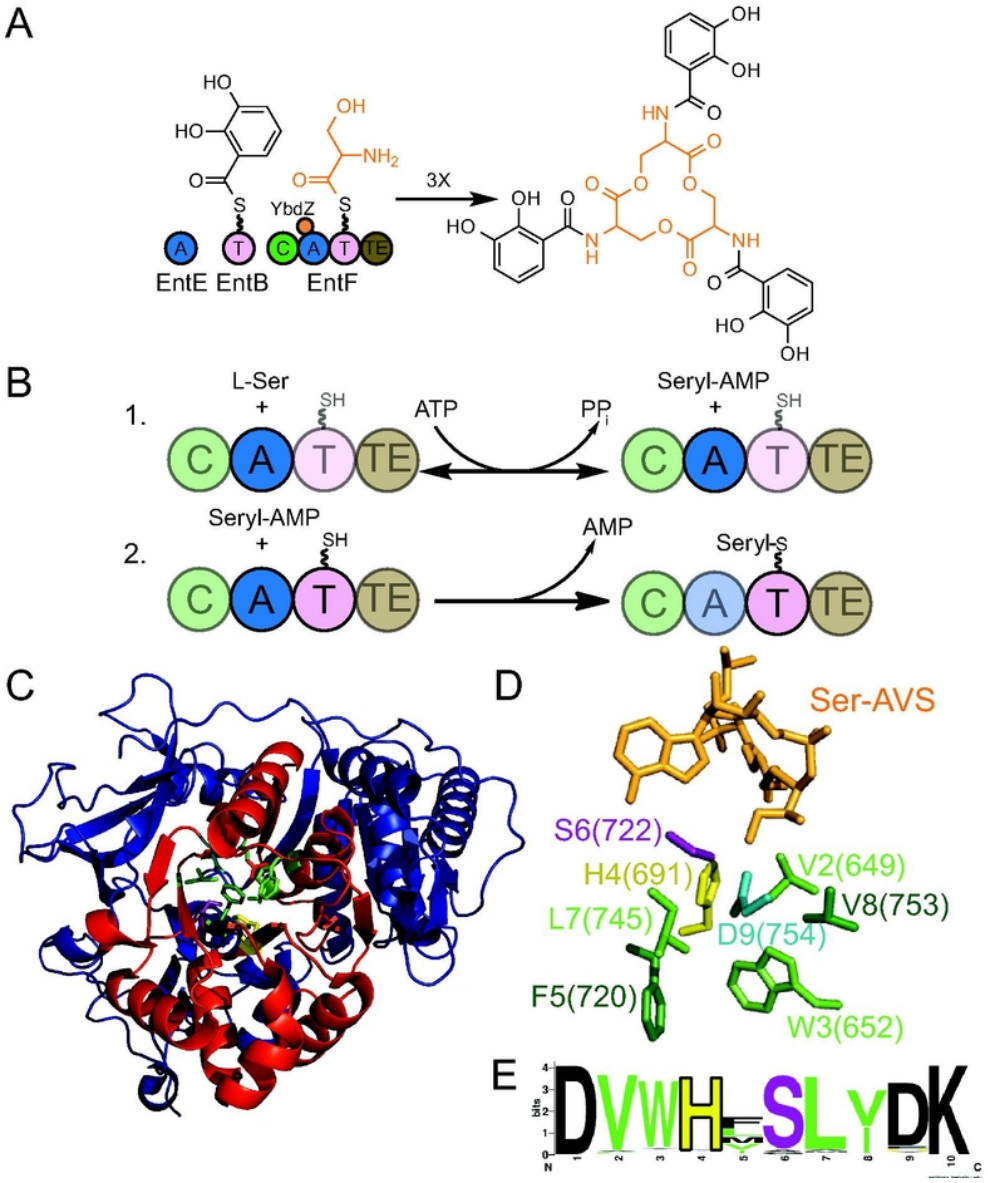

Figure 1. Enterobactin (ENT) formation and the structure, specificity code, and function of the A domain of EntF. A) Diagram of the ENT biosynthetic pathway. EntE tethers 2,3-dihydroxybenzoic acid (DHB) to EntB. The condensation (C) domain of EntF condenses DHB and L-Ser, previously activated and bound to the thiolation (T) domain of EntF. After one turnover, the DHB-L-Ser monomer is stored on the thioesterase (TE) domain. After three iterations of this process, the final product is cyclized and released. B) The two half-reactions of the EntF A domain. In the first, the A domain activates L-Ser as Seryl-AMP (1) and transfers the seryl group to the T domain (2) in the second. C) Ribbon representation of the EntF (PDB5JA1) A domain with the recognition subdomain, in red, and specificity code residues, highlighted in colors matching D and E. D) Binding pocket residues that form the specificity code and the reaction intermediate-mimic inhibitor, seryl-AVS (25), used for crystallization. E) WebLogo (26) of specificity code sites 1–10 for 82 characterized L-Ser codes (27).

67x80mm (300 x 300 DPI)

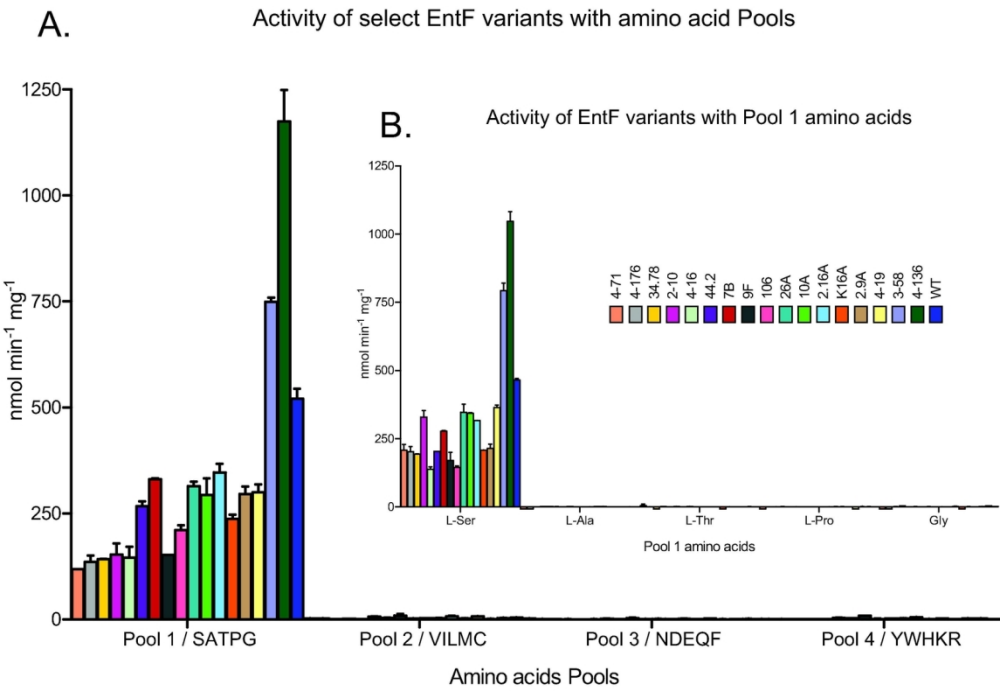

Figure 2. Residue usage by code site in the EntF variants and characterized L-Ser-specific A domains. Pie charts describing, by site in the specificity code, the amino acid usage across all libraries. The wild-type EntF residues are shown in the top row followed by the 82 characterized L-Ser-specific A domain residues, and, subsequently, each of the four libraries. Circles with diagonal lines designate non-mutagenized sites. The color of each sector corresponds to the key at the bottom; both are organized alphabetically, clockwise and left to right. For each library, n designates the number of isolated strains containing DNA-unique entF mutants, provided they had no non-code residue substitutions. All Libraries n= 200, 216, 50, 168, 50, 216, 168, and 17 for sites 2–9, respectively.

139x95mm (300 x 300 DPI)

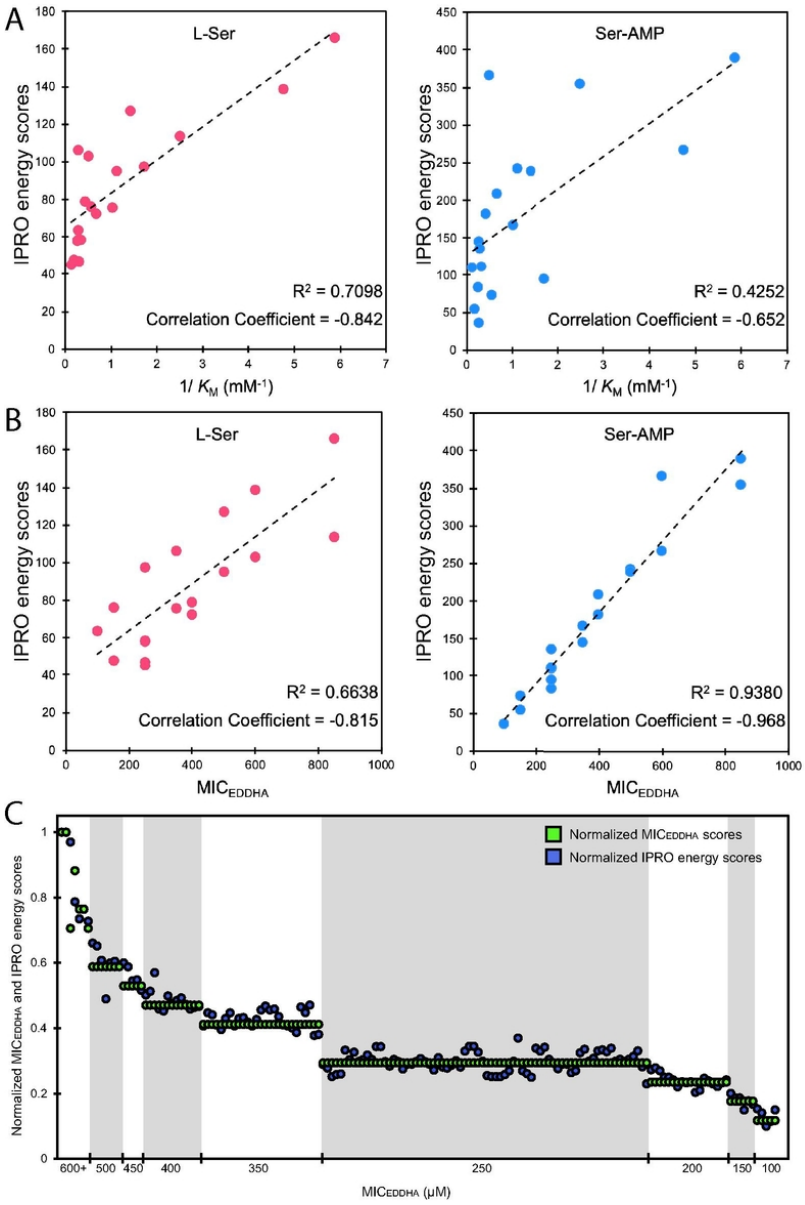

Figure 3 in silico IPRO energy scores of EntF variants for L-Ser correlate with Km, binding for Ser-AMP correlates with MICEDDHA. Correlation between the IPRO energy score with L-Ser or Ser-AMP and A) Km and B) the associated MICEDDHA of in vitro-characterized EntF variants (Table S3). A more negative value for IPRO energy score indicates stronger binding. C) Correlation between IPRO energy score with Ser-AMP by each EntF variant (blue) and the MICEDDHA of the corresponding entF mutant strain (green). IPRO energy scores and MICEDDHA are normalized to wild-type.

66x99mm (300 x 300 DPI)

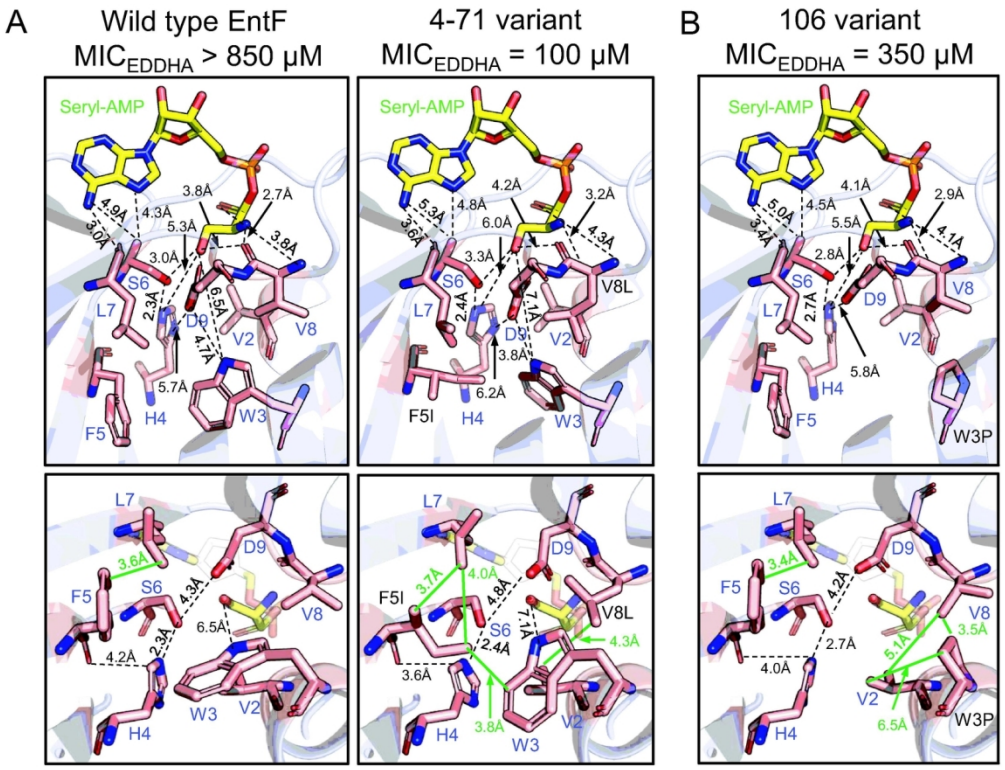

Figure 4. Substrate-binding pocket interactions differ between high- and low-functioning variants. Diagrams of the substrate-binding pockets of (A) WT EntF, EntF variant 4-71, and (B) EntF variant 106 highlighting differences between their energy-minimized structures, based on PDB 5JA1, in the electrostatic interactions of the code residues and Ser-AMP (dashed line; side view – top row) and the intra-enzyme interactions (green line; bottom view – bottom row). Each diagram also contains specificity code residues 2–9 (light pink with blue font), the Ser-AMP intermediate (yellow), and residues differing from WT (black font).

139x106mm (300 x 300 DPI)

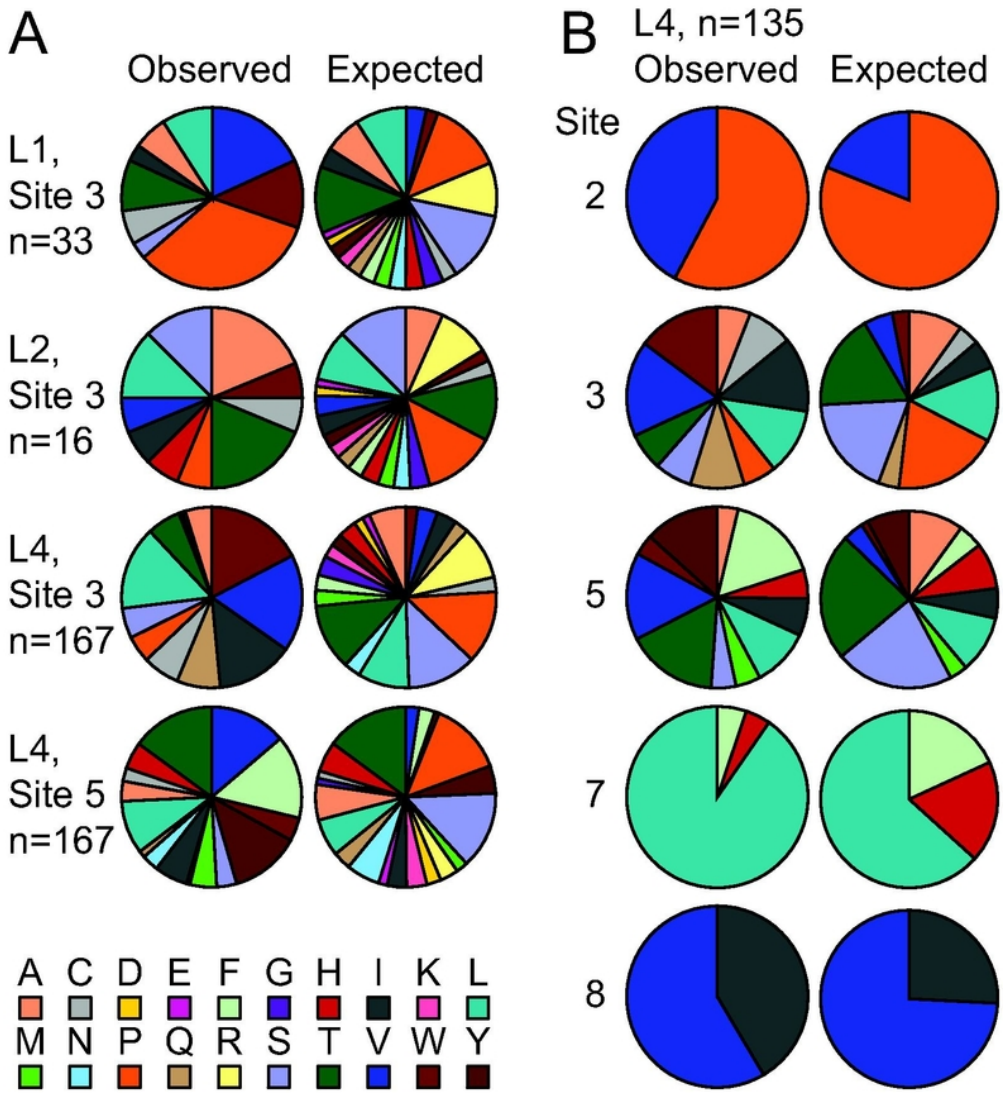

Figure 5. Preferential residue usage is observed at all sites. A) Comparison of amino acid residue proportions at sites 3 and 5 between the input and selection output. Each sector color corresponds to an amino acid residue according to the key at the bottom and sectors are sorted, clockwise from the top, by contribution to the chi-squared statistic. B) Comparison of the residue proportions at each mutagenized site in the L4 dataset to the input. The color of each sector corresponds to the key at bottom, both of which are organized alphabetically, clockwise and left to right, respectively.

66x73mm (300 x 300 DPI)

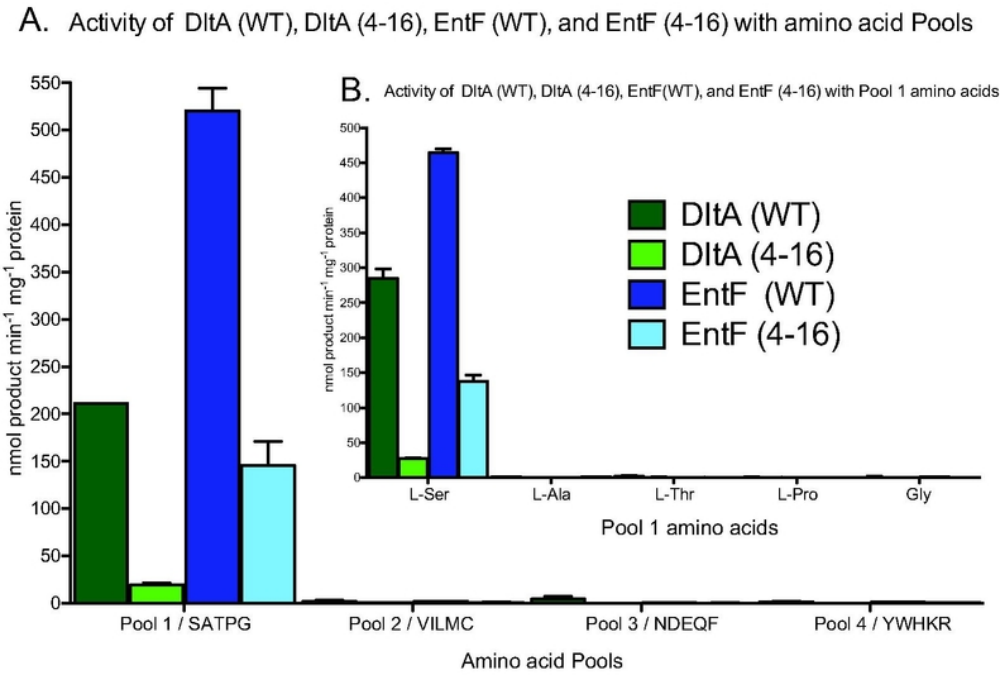

Figure 6. A variant EntF specificity code, differing at all five mutagenized sites, functions in a non-EntF context. Comparison of the in vitro activity of purified proteins EntF WT, EntF variant 4-16, DltA WT, and DltA variant 4-16 with pooled amino acid substrates (A) and the individual amino acids contained in pool 1 (B) as determined by ATP-PPi exchange assay. The color key in panel A applies to both panels. Pool 1: SATPG, 2: VILMC, 3: NDEQF, 4: YWHKR.

66x45mm (300 x 300 DPI)

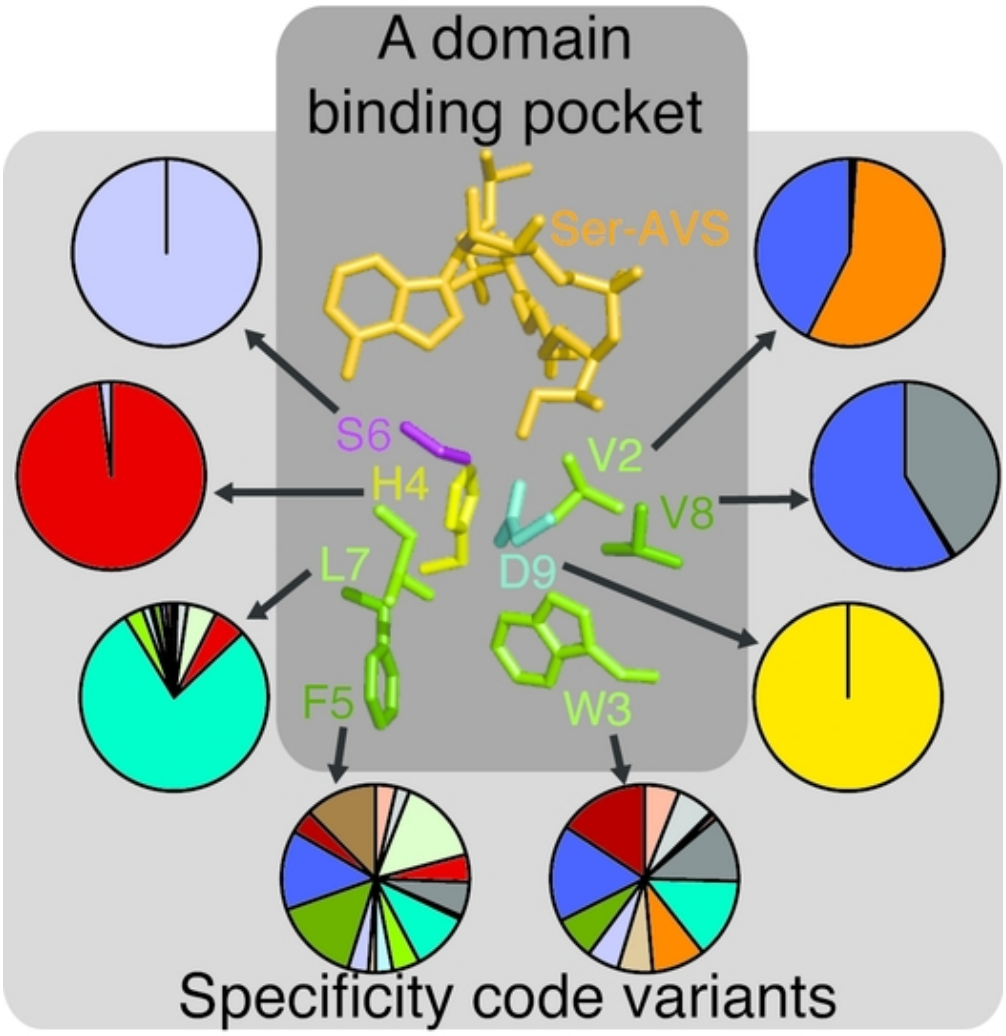

Table of Contents Figure

44x46mm (300 x 300 DPI)