Engineering antimicrobial peptides and enzymes: Realizing opportunities to combat antibiotic resistance through high-throughput, data-driven design and testing

A DISSERTATION
SUBMITTED TO THE FACULTY OF THE
UNIVERSITY OF MINNESOTA
BY

Seth Christopher Ritter

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

Benjamin J. Hackel

October 2019

Acknowledgements

The work presented herein would not have been possible without the challenge, encouragement, and guidance of my mentors, friends, and family. I am eternally grateful for the innumerable people's lives with which mine has had the pleasure of crossing, for however long or brief. The composition of this manuscript, and of myself, has been molded by the gifts of so many.

I must first thank those that came first, my family. To my first friend, Katie, being your older brother has been, and continues to be, one of the greatest parts of my life. To my parents, your sacrifices and dedication to my endeavors are what enabled me to accomplish this. And to my ever-expanding family, to Harmony, Natalie, James, and Ryan, and to my nieces and nephews, and to those you've brought together, thank you for being a part of my life. And lastly to my partner, Josh, you have taught me so much, and helped me become the person that I am today.

I must thank my friends. Too numerous to contain within these pages, I have had the fortune of being surrounded by friends of all types. During the composition of this thesis, I thank the past and present members of the Hackel and Kaznessis groups, it has been a pleasure to work alongside so many talented people. Of particular note, Sadie and Brittany, you have been and continue to be some of my closest friends.

I must thank my mentor, Dr. Benjamin Hackel. You are a leader unique in my life, presenting an ideal of not only knowledge and logic, but also of compassion and encouragement. Being your student has enhanced me both professionally and personally. Thank you for all the discussions, they were some of the best I've ever had.

Lastly, I must thank my committee. Drs. Wei-Shou Hu, Samira Azarin, and Will Pomerantz. You have each played important roles throughout my education, in courses, teaching-assistantships, and beyond.

Dedication

"If I have seen further it is by standing on the shoulders of Giants"

- Sir Isaac Newton

Proudly dedicated to Mom and Dad, my giants, who lifted me up so high that all I could see were clouds below and the stars above.

Abstract

The increase in occurrence of antibiotic resistant infections around the world as a result of overuse of broad-spectrum antibiotics in both agriculture and healthcare poses a significant threat to human health and societal productivity. When most antibiotics were discovered in the mid 20th century, biotechnology as a rigorous science was still far-off. Today, advances across a wide range of disciplines is finally permitting the detailed description and, more importantly, alteration of biological systems. Enabled by this rapidly progressing domain exist alternatives to traditional broad-spectrum antibiotics; of interest here are antimicrobial proteins, which are ribosomally synthesized. Being encoded for in DNA permits the sequences of these proteins to be mutagenized and their sequence-function landscapes rapidly explored. Such exploration was utilized herein to optimize specificity, activity, and stability of three antimicrobial proteins. During this exploration computational and experimental methods permitting high-throughput characterization were developed and applied.

First, a small-lasso antimicrobial protein, microcin J25, was engineered for improved specificity towards pathogenic *Salmonella* in contrast to commensal *Escherichia coli*, which were isolated from human patients. To accomplish this a plasmid containing a synthetic gene cluster encoding for the precursor peptide of microcin J25, under inducible expression, with three enzymes necessary for maturation and secretion was modified to facilitate mutation of the precursor of microcin J25. A collection of 207 point-mutants across 12 positions was evaluated for activity against a panel of pathogenic *Salmonella* and *E. coli* serotypes. Point-mutants demonstrating retention of activity and improvements in specificity were then integrated and screened as a combinatorial library. Multi-mutants demonstrated significant reduction in efficacy, with only 3.5% of sequences in the library having detectable activity against *Salmonella enterica* serovar Enteritidis. At the project's conclusion, a point mutant was identified which retained a high level of activity against the target *Salmonella* species, while reducing the off-target activity towards human commensal *E. coli* by an average of 81%.

Second, a large multi-domain antimicrobial protein which binds to and degrades the cell wall of pathogenic *Clostridium perfringens*, LysCP2, was validated and stabilized. Owing to its origin, thousands of homologous protein sequences, with structure and function similar to LysCP2, were readily identifiable from genomic databases. We hypothesized that this wealth of homologous information could be utilized to guide the design of combinatorial libraries of LysCP2 to improve its poor stability. Using coevolutionary models, which incorporate pairwise and sitewise information from the homologous sequences, a collection of ten multi-position libraries were designed and generated at different positions of LysCP2. From these libraries, nearly 10,000 variants of LysCP2 were experimentally assayed for stability. These data revealed that the fraction of stable variants for 8 out of 10 designed libraries was greater than fully-random libraries at the same positions in the protein. In addition, post-facto analysis incorporating a structural homology model of LysCP2 implied that structural features, such as contact number and secondary structure, may be reasonable filters to use for a priori selection of residues for which the predictions of the coevolutionary model are most accurate. Finally, due to the high functionality of the designed libraries, the experimental data was used directly to inform the generation of a five designs of LysCP2, with between five and six mutations, of which: all retained enzymatic activity; four demonstrated an increase in melting temperature; three demonstrated increased retention of activity after thermal shock. Of the designs, the highest performing improved the melting temperature of LysCP2 from approximately 38 °C to 42 °C.

Third, similar to LysCP2, another cell-wall-degrading enzyme with rich homologous sequence information was optimized for both activity and stability. This enzyme, LysEFm5, has activity against pathogenic and vancomycin-resistant *Enterococcus faecium*. Again, using coevolutionary models informed from homologous sequences, a collection of combinatorial libraries was generated focusing mutagenesis on sites supporting residues directly involved in interacting with the cell wall. These models were predictive of activity retention across 873 experimentally tested variants (AUC = 0.840 – 0.894). In addition, the accuracy of these models was assessed when systematically varying the type and amount of sequence information utilized, supporting the utility of using pairwise features and providing guidelines for the value of different types of sequences. Lastly, further exploration of random members of these libraries revealed an enhanced clone with 2x higher specific activity in addition to an 11 °C increase in melting temperature in comparison with wild-type LysEFm5.

This work provides evidence and methods which support the application of protein engineering to the optimization of antimicrobial proteins to improve their utility as next-generation antimicrobials.

Table of Contents

List of Tables

List of Figures

Chapter 1 Introduction

1.1 Discovery of and resistance to traditional broad-spectrum antibiotics

As a consequence of competing for limited resources many organisms have evolved pathways for the production and secretion of compounds which inhibit the growth of competing organisms. Within this classification, traditional antibiotics are small-molecule compounds evolved to inhibit the growth of bacteria. The most famous example of an antibiotic, penicillin, was discovered by Alexander Fleming in 1928 largely by accident. As the story goes, he observed the formation of a zone of growth inhibition surrounding a fungal colony on a plate which he had seeded with bacteria. Subsequent fermentation and purification of the growth liquid of this fungus, the supernatant, would reveal penicillin.[1]

Interestingly, the vast majority of antibiotics that have been discovered were found using a streamlined version of this crude methodology.[2,3] That is, a producing species, often soil-derived streptomycetes, were screened by plating with target organisms and the emergence of a zone of growth inhibition of the target was monitored for. Though crude, this method was for a time effective, and the pharmaceutical industry relied heavily on it from the 1940s-1960s to discover most antibiotic classes which are approved for clinical use today. This pattern of discovery is summarized in Figure 1.1.

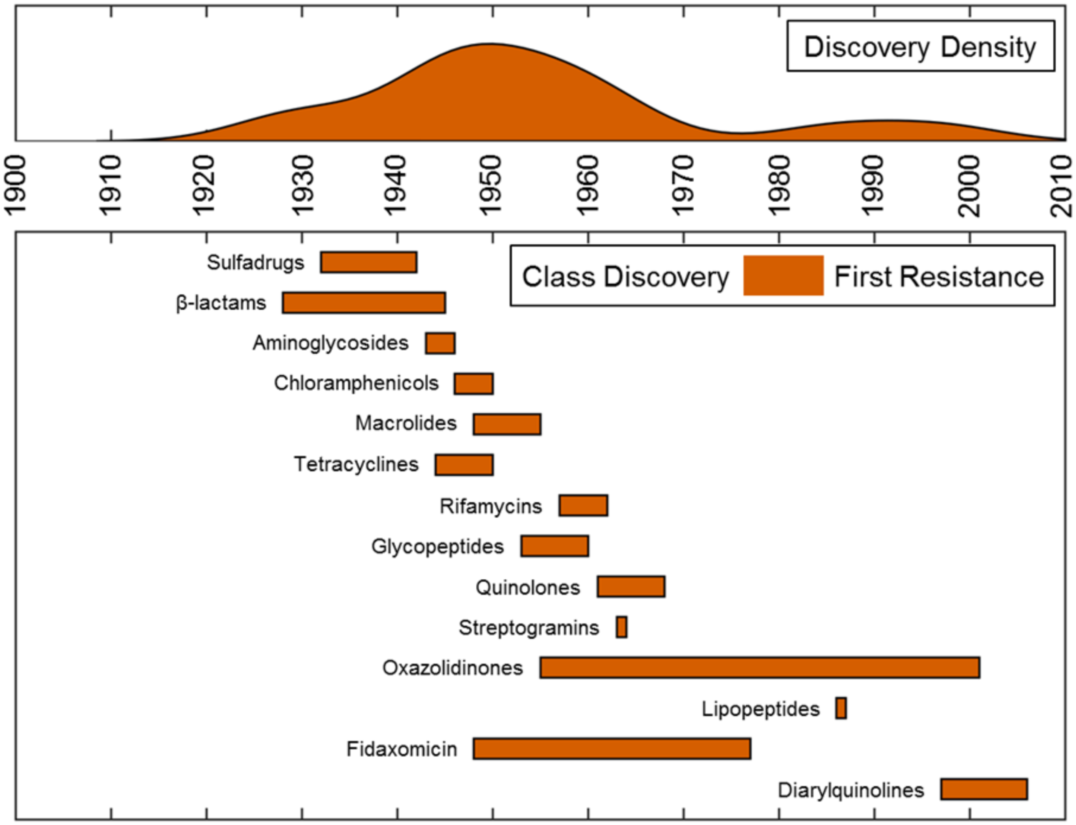


Figure 1.1. Discovery and resistance to antibiotic classes. (Top) The density of discovery of different classes of antibiotics. (Bottom) Major classes of antibiotics with their discovery dates and first record of clinical resistance indicated as the left and right side of the bars, respectively. Data adapted from [3].

Accompanying the rapid discovery of antibiotics was also the characterization of resistance to those antibiotics as it emerged in the clinic. Bacteria have adopted many different mechanisms to survive the application of antibiotics. These mechanisms can be categorized into three groups[4]: (1) decreasing intracellular concentration of antibiotic; (2) modifying the target of the antibiotic; (3) inactivating the antibiotic. In 2014, antibiotic resistant infections accounted for approximately 700,000 deaths annually and are estimated to grow to 10 million deaths per year by 2050[5].

Critical to the success of current and future antibiotics, it is necessary to discuss the rate with which different modes of antibiotic resistances emerge. Some mechanisms, such as mutations to a target protein which occur randomly, can cause resistance to an antibiotic to be rapidly developed within a target bacterial population. These mutations can occur directly within the target protein as has been observed in the context of streptomycin resistance which can be due to mutations in ribosomal protein S12[6]. Mutations can also occur within transcription factors altering the expression of proteins critical in the uptake of the antibiotic. It has also been observed that in instances where a significant fitness hit isn't taken, often due to redundant pathways, downregulation of transport proteins can result in insufficient antibiotic uptake into target bacteria, as has been observed for imipenem targeting *Enterobacter aerogenes*[7], and with many other *Enterobacteriaceae* targeted by carbapenems[8]. Indeed in laboratory settings this mutagenesis and downregulation can be observed over single days.[9] In contrast, antibiotics which target downstream metabolic products, such as vancomycin binding to and preventing cross-linking of bacterial cell-walls, require significantly more machinery for resistance[10]. In the case of vancomycin, stable resistance is conferred by the expression of a three-enzyme gene cluster which modifies the cell-wall stem peptide preventing the binding of vancomycin. This resistance was not observed in the clinic for more than 30 years, but today is spreading rapidly between bacteria via horizontal gene transfer[11]. Critically, it is not sufficient that a resistant mutation or gene acquisition be possible, but also that the alteration doesn't impose a fitness cost so significant that the bacterium cannot compete in the local environment.[12]

There have been recent advances which may offer a renewed interest in these brute force methods. All variations of these methods attempt to address the issue that most microorganisms in the environment cannot be cultured under laboratory conditions[13,14]. Further, many of the pathways responsible for antibiotic production are not active under culture conditions used. To address the first shortcoming single cells from an environment can be isolated and grown within chambers possessing semipermeable barriers permitting the isolated growth of cells *in situ*. These isolation chip, or iCHIP, growth techniques increase the number of culturable soil microbes from approximately 1% to nearly 50%. This method was successfully used to discover a novel antibiotic compound[15,16]. Addressing the latter shortcoming, metagenomic sequencing, where the collection of all DNA within an environment is sequenced and computationally assembled into larger sections, is actively being applied to discover possible gene clusters encoding for antibiotic compounds[17]. This latter approach affords the possibility of reducing the occurrence of repeated discovery of the same antibiotics while offering the possibility of focusing study on those pathways whose compounds are predicted to have efficacy towards target bacteria of interest.

Though these innovations will offer reprieve, they are still limited to the pool of naturally occurring antibiotics. Though antibiotics can be chemically modified to alter pharmacokinetic properties, efficacy, and specificity, these changes often offer only minor optimization at extremely low throughput.

1.2 Properties of next-generation antimicrobials desired

As alluded to in the previous section, it is desirable to explore methodologies permitting the transition from the discovery of antibiotics and more broadly antimicrobials to their development instead. This transition is desirable for several reasons.

First, the design space for most traditional broad-spectrum antibiotics is oftentimes confined to a narrow chemical space surrounding the original antibiotic. By logical extension, it would be anticipated that this design space would contain few modifications to optimize the antibiotic. This limited optimization potential diminishes efforts not only to enhance efficacy or specificity, but also reduces the opportunities for reducing side effects that accompany administration. For example fluoroquinolones can trigger headaches, dizziness, as well as seizures.[18] It is desirable, therefore, that a next-generation antimicrobial platform permit a design space with sufficient breadth that multiple properties can be optimized for.

Second, however exhaustive, mining the naturally occurring antimicrobials represent a fixed source of molecules. Though better practices can reduce the onset and spread of antibiotic resistance[19], it cannot be reduced to zero. As a result, naturally occurring antibiotics, or more importantly the subset with appropriate pharmacokinetic properties for clinical application, will gradually be overcome by antibiotic resistance. Indeed, the present scenario of antibiotic resistance is an example of this. It is desirable, therefore, that a next-generation antimicrobial platform permit expedited discovery and optimization so that resistance can be overcome as it emerges.

Third, there is growing evidence that modifying the composition of different microflora in humans, for example in the gastrointestinal tract, can affect many different aspects of human health. These effects can include blood-glucose response to food[20], the effectiveness of cancer therapy[21], and inflammation of the brain[22]. Modifications to flora composition can also lead to secondary infection, as is the case with *Clostridium difficile* which kills an estimated at 29,000 annually in the United States in 2011.[23] It is desirable therefore, that the candidate next-generation antimicrobial platform present new opportunities for improved specificity.

1.3 Ribosomally synthesized antimicrobial proteins

A candidate platform which presents opportunities as a next-generation antimicrobial are antimicrobial proteins (AMPs). Narrowing one's view in the domain of AMPs to those which are ribosomally synthesized affords other critical advantages. The most important of these advantages is that because these AMPs are ribosomally synthesized the instructions, or sequence, for their generation is directly encoded for by DNA. This attribute permits the rapid application of molecular biology enabling expedited design iteration through tailored gene synthesis and mutagenesis. In addition to changing the sequence of the AMP itself, being ribosomally synthesized also permits, under appropriate conditions, the sequence of the AMP to be extended to generation of fusion proteins. These fusions can add mechanisms for detection and isolation to the AMP which were not present naturally. Critical for some of the studies presented herein, detection mechanisms permit the translation of a molecular process, such as protein unfolding, to a readily measurable signal such as fluorescence. As will be explored, appropriate application of this methodology permits the efficient analysis of tens of thousands of proteins in single experiments, with the capacity to scale to millions.

Within the smaller scales, some AMPs are the result of a precursor peptide undergoing multiple processing steps from accompanying enzymes. For example, the lasso peptides constrain their geometries by generating one or more loops of residues and threading chains through them such that they become sterically locked[24]. These peptides are extremely resistant to thermal, chemical, and proteolytic degradation.

Characterized lasso AMPs have been found to bind to and inhibit a number of bacterial enzymes, including RNA polymerase[25], and prolyl endopeptidase[26]. A recent genomic analysis has identified approximately 1300 different gene clusters potentially able to generate lasso peptides[27].

Increasing size substantially, one can find the multi-domain endolysins. These enzymes bind to and degrade the cell wall of target bacteria[28]. Endolysins originate from the bacteriophage lytic cycle and are released through holes in the cell membrane formed from holin proteins. Upon release, these enzymes degrade the cell wall of the infected cell, eventually causing cell lysis and release of bacteriophage progeny. As a secondary consequence of their bacteriophage origin, sequences of putative endolysins have been readily identified in prophage domains of bacterial genomes[29], or from metagenomic studies of uncultured bacteriophage in the environment[30]. Beyond endolysins, the catalytic domains are present in endogenous proteins involved in bacterial cell wall construction, remodeling, and recycling.[31] Collectively these sources present tens of thousands of candidate molecules to learn from and explore.

1.4 Searching and optimizing in protein sequence space

To realize the potential of AMPs as an antimicrobial development platform, it is necessary to continue to mature the broader field of protein engineering. Upon initial characterization, it would seem that protein space is too vast for efficient exploration. Indeed, if there were no structure to the mapping between protein sequence and function this observation would largely be correct. Protein space scales exponentially with the length of the protein's primary sequence (L), which for proteins using the naturally occurring amino acids is 20^L. Within this space a protein of modest size, say length 70, would have more sequence possibilities than there are protons in the observable universe. Even if one begins protein design from a known starting point, for example to enhance the specificity of an AMP, the collection of variants to be tested can quickly outstrip comprehensive evaluation.

The first important observation, however, made in principle several hundred years ago by Charles Darwin, is that gradual changes can mature different features of biological systems[32]. Indeed, randomly mutating a protein in the laboratory and selecting for improvements to its physical properties was the basis for a portion of the Nobel Prize in Chemistry in 2018. This process of mutation and screening in the laboratory is termed broadly as directed evolution. This process demonstrates that there exists a degree of connectedness in protein sequence space, which can be explored to mature protein properties[33]. Going further than naïve exploration, it has been demonstrated for a number of proteins that the sequence space can be described with a parameter space significantly smaller than all possible proteins within the explored domain.[34–39] Critically, these sequence and function models are generated using experimental data pertaining to the desired or a closely related physical properties of the studied proteins.

Accurately inferring the mapping between sequence and function a priori is one of the central goals of protein engineering today. Broadly, these efforts can be categorized into several domains. Bottom-up approaches, such as molecular dynamics simulations[40], FoldX[41], and Rosetta[42], employ empirical and knowledge-based energy potentials in conjunction with sampling techniques in order to study the emergence of physical properties, such as stability[43], directly from the movement of atoms comprising the protein of interest. In contrast, top-down approaches aim to transfer the experimentally inferred models of studied proteins to those beyond the study. Examples of approaches which have found success incorporating sequence and structural properties include the prediction of: aggregation propensity[44,45]; stability changes upon mutation[46–49]; solubility[50]; immunogenicity[51].

Indeed, the future of protein engineering will likely employ a combination of increasingly performant protein library design using a priori information, followed by increasingly performant high-throughput protein characterization and subsequent modelling. This is because as our capacity to predict the properties of protein designs at increasing mutational distances from known functional sequences grows, it permits the subsequent computational exploration of significantly larger sequence space once experimental data has been gathered. As a hypothetical example, if sufficiently descriptive a priori models informed the design of a combinatorial mutagenic library wherein 20 positions were mutated to any of 5 amino acids simultaneously, only about $5x10^4$ proteins would need to be studied to provide a description of approximately 10^{14} possible sequences, using a maximum entropy model which incorporates pairwise interactions between each of the different residues at each of the different positions[52].

Historically, engineering of AMPs has focused on the characterization of a small number of rationally chosen mutants of promising wild-type candidates. Though this approach has been fruitful in limited contexts, it is often amenable only to limited library sizes and modification. As presented in this work the application of systems biology, yeast surface display, high-throughput screening, and next-generation sequencing, both translate and advance protein engineering to the context of AMP engineering. The methods and results presented herein are both immediately applicable and provide guidance for further advancement. Work such as what is presented has, with increasing probability, the opportunity to transform antimicrobials from discovery to development, permitting the efficient optimization of physical properties most relevant to real-world application.

Chapter 2 Multi-species activity screening of microcin J25 mutants yields antimicrobials with increased specificity towards pathogenic *Salmonella* species relative to human commensal *Escherichia coli*

Adapted from “Ritter, S.C., Yang, M.L., Kaznessis, Y.N., Hackel B.J., ‘Multispecies activity screening of microcin J25 mutants yields antimicrobials with increased specificity toward pathogenic *Salmonella* species relative to human commensal *Escherichia coli*.’ *Biotechnology and Bioengineering* 2018, 115: 2394– 2404.

Permission to reuse all figures and text contained in this chapter has been granted by John Wiley and Sons.

2.1 Synopsis

Modern large-scale agricultural practices that incorporate high density farming with sub-therapeutic antibiotic dosing are considered a major contributor to the rise of antibiotic resistant bacterial infections of humans with species of *Salmonella* being a leading agriculture-based bacterial infection. Microcin J25, a potent and highly stable antimicrobial protein active against Enterobacteriaceae is a candidate antimicrobial against multiple *Salmonella* species. Emerging evidence supports the hypothesis that the composition of the microbiota of the gastrointestinal tract prevents a variety of diseases by preventing infectious agents from proliferating. Reducing clearance of off-target bacteria may decrease susceptibility to secondary infection. Of the Enterobacteriaceae susceptible to microcin J25, *Escherichia coli* are the most abundant within the human gut. To explore the modulation of specificity, a collection of 207 mutants encompassing 12 positions in both the ring and loop of microcin J25 was built and tested for activity against *Salmonella* and *Escherichia coli* strains. As has been found previously, mutational tolerance of ring residues was lower than loop residues, with 22% and 51% of mutations, respectively, retaining activity towards at least one target within the target organism test panel. The multi-target screening elucidated increased mutational tolerance at position G2, G3, and G14 than previously identified in panels composed of single targets. Multiple mutations conferred differential response between the different targets. Examination of specificity differences between mutants found that 30% showed significant improvements to specificity towards any of the targets. Generation and testing of a combinatorial library designed from the point-mutant study revealed that microcin J25^{I13T} reduces off-target activity towards commensal human-derived *E. coli* isolates by 81% relative to *Salmonella enterica* serovar Enteritidis. These *in vitro* specificity improvements are likely to improve *in vivo* treatment efficacy by reducing clearance of commensal bacteria in the gastrointestinal tract of hosts.

2.2 Introduction

It is estimated that every year two million people in the United States suffer from antibiotic resistant infections, with at least 23,000 resulting in death[53]. Overuse and decreased development of antibiotics in multiple economic sectors in both developed and emerging countries is projected to result in increased numbers of infections and deaths over time[5]. One of the largest avenues of infection by antibiotic resistant bacteria is from agricultural products[54]. In the United States a leading foodborne infectious agent is *Salmonella*[55]. Due to both concerns over development and spread of antibiotic resistance amongst bacterial communities as well as the decreased antibiotic discovery rate across the pharmaceutical industry[3], there is growing desire for so-called next generation antimicrobials[56]. Among other features, these agents are designed to have improved efficacy and specificity towards target organisms. Improved specificity is important for at least the following two reasons: first, for reducing selective pressure on broader bacterial communities to develop mechanisms of resistance, which can be spread by horizontal gene transfer[57]; second for reducing clearance, or modification of the composition, of microbiota within bacterial communities such as in the gastrointestinal tract. These microfloral environments provide protection against primary and opportunistic infection and are a contributor to a wide variety of factors of human health[58].

One potential development platform is the use of antimicrobial proteins (AMPs), either as free proteins[59], or secreted via engineered probiotics at the site of infection[60–64]. Microcin J25 (MccJ25) is one such AMP with high activity against multiple species of *Salmonella* and other enterobacteriaceae such as *Escherichia coli*[24,65–70]. Mature MccJ25 is a 21 amino acid lasso protoknot with high thermal, chemical, and proteolytic stability[71]. Its complete maturation and export require the co-expression of two cytoplasmic enzymes and an export protein (Figure 2.1.A, proteins B, C, D)[72]. Its primary mode of action is the inhibition of RNA polymerase via competitive binding to the secondary channel[25,73].

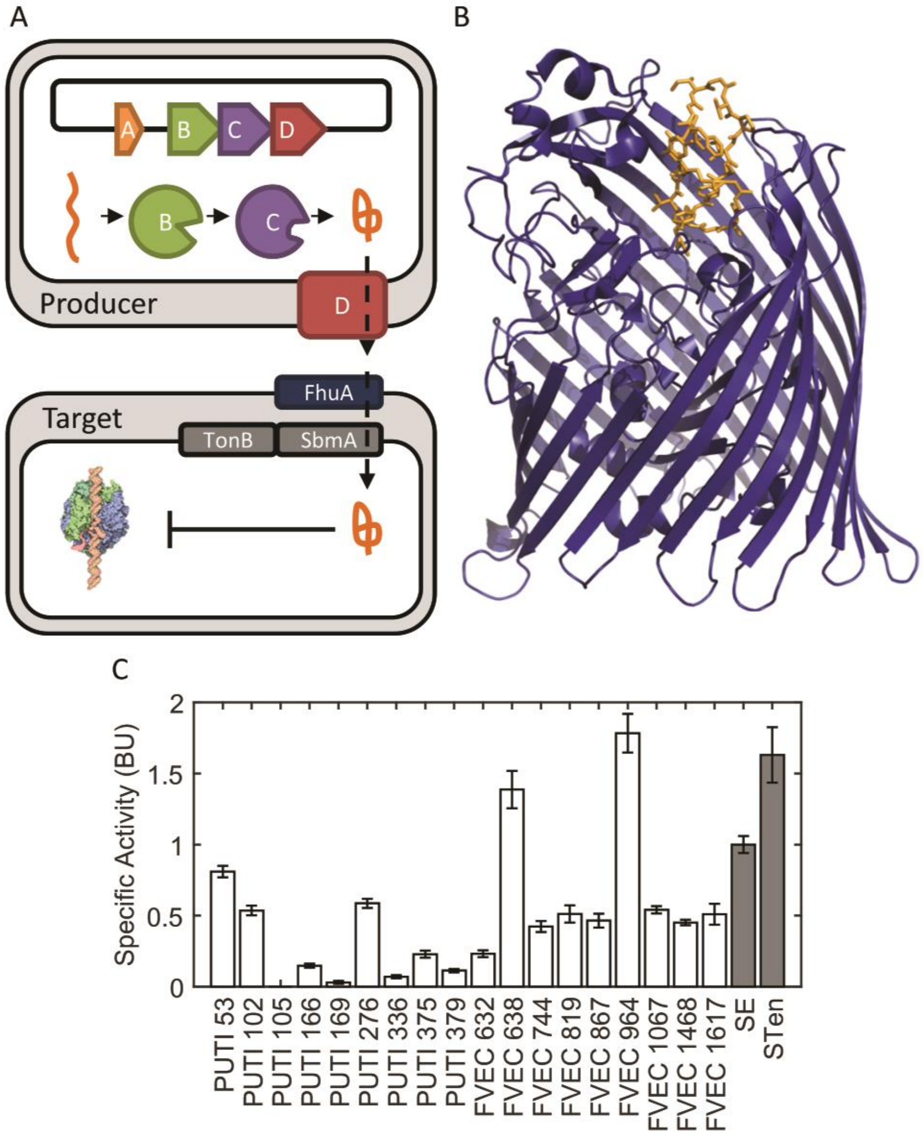


Figure 2.1. MccJ25 maturation, export, and activity (A) MccJ25 is the product of a 4-gene cluster composed of A, a 58 amino-acid precursor, B, an ATP-dependent cysteine protease, C, an amidotransferase, and D, an ATP-binding cassette transporter. Following export from a producing bacterium, mature MccJ25 is uptaken through homologs of the iron-siderophore receptor FhuA in a TonB-system mediated transport process. Once intracellular, MccJ25 exerts its more common mode of action, binding to the secondary channel of RNA polymerase thus inhibiting transcription. (B, PDB: 4CU4) Uptake through the FhuA receptor (blue) involves the specific interaction of MccJ25 (orange). (C) Wild-type MccJ25 shows strong activity against pathogenic *Salmonella* strains (grey) as well as non-pathogenic commensal *E. coli* (white) isolated from human patients.

Variants of MccJ25 have previously been studied in order to address questions about the flexibility of the maturation machinery, the export mechanism, as well as improved activity[74–76]. As it relates to this work Pavlova et al. previously evaluated nearly the complete set of single mutants of MccJ25 for production, *E. coli* RNA polymerase inhibition, as well as growth inhibition of *E. coli* strain DH5α. However, with the exception of one multi-target comparison for a single mutant (MccJ25^{T15G}) the study exclusively focused on a single target organism. Their work revealed that only 45% of MccJ25 point-mutants, which could be produced, exported, and inhibit RNA polymerase *in vitro*, retained antimicrobial activity. This discrepancy is due to the import process of MccJ25, whose fine details are not yet fully understood[68,69]. Differences among proteins involved in the import of MccJ25 from different bacterial targets could potentially enable the specificity of MccJ25 to be modulated. The lack of a detailed biophysical description of MccJ25's interaction with import machinery necessitates high-throughput activity assays to develop MccJ25 variants with improved target specificity. This modulation is critical, as *E. coli* isolates from humans show high susceptibility to MccJ25 (Figure 2.1.C).

Though there has been considerable effort to study AMP mutants for improved activity[77–85] there are few examples of studies focusing on the modulation of AMP specificity[86,87]. To our knowledge an evaluation of specificity modulation of MccJ25 has not been undertaken and herein we present specificity differences between single-mutant variants of MccJ25 between a collection of *Salmonella* and *E. coli* targets (Table 3). We also explore the ability of a combinatorial library incorporating the most promising mutants to generate functional variants of MccJ25.

2.3 Materials and Methods

2.3.1 Bacterial culture and strains

All bacterial growth was performed in either liquid or solid (1% agar) lysogeny broth (LB) with or without supplements. For propagation and maintenance of NEB I^q cells transformed with pJP4 and derivatives, 100 µg/mL ampicillin were used. For induction of MccJ25 variant production, in liquid or solid, isopropyl β-D-1-thiogalactopyranoside (IPTG) was added to final concentration of 0.5 mM. For a complete list of plasmids, *Escherichia coli* production strains, as well as bacterial indicator strains including human commensal *E. coli* isolates (generously provided by University of Minnesota students and Veterans Affairs Hospital patients and/or their families via Dr. Johnson of the VA Medical Center of Minneapolis), pathogenic *E. coli,* and *Salmonella* see Table 3.

2.3.2 Generation of pJP4 for MccJ25 variant expression

Plasmid containing MccJ25 and maturation operon (pJP3), provided by the Link lab from Princeton[88], was digested with XhoI and HindIII (New England Biolabs). Modified MccJ25 insert (Table 4) containing the sequence to move XhoI downstream as well as an optimized ribosomal binding site[89] were Gibson assembled (HiFi Assembly Kit, New England Biolabs) with digested plasmid to generate pJP4. This was done to both reduce the size of library inserts used by 30%, as well as remove high redundancy present in the promoter region, simplifying library assembly. This plasmid possesses the T5 promoter under regulation with LacO/LacI to enable inducible production of MccJ25.

2.3.3 Generation of single-site saturation mutagenic libraries of MccJ25

For single-site saturation mutagenesis, oligonucleotides (Table 4) with NNK degeneracy (all amino acids + stop) at positions 2-7 and 9-14 (each site independently randomized on an independent oligonucleotide) were synthesized and assembled via polymerase chain reaction (PCR, all PCRs conducted with NEB Q5 Polymerase) with additional oligonucleotides to construct MccJ25 variant inserts. Inserts and pJP4 were digested with XhoI and HindIII, and subsequently ligated together with T4 DNA ligase (New England Biolabs) overnight at 16°C. Subsequently, each of the 12 ligations was transformed separately into NEB Express I^q and plated on LB with ampicillin. Once grown, random colonies from each transformation were inoculated into 100 µL LB with ampicillin in 12 sterile 96-well plates and grown overnight. Once grown, 5 µL from each well on each plate were added to a new sterile 96-well plate and mixed to preserve well locations in the final mixed plate. Whole-cell PCR was then conducted to amplify each well separately with unique 4-base pair row and column indices adjacent to the coding region of mature MccJ25. PCR products were then mixed, purified (Qiagen spin column), and amplified with a subsequent PCR to append appropriate adapters for Illumina Sequencing.

Illumina MiSeq sequencing (1/8th lane) generated 1.3 million reads. Sequences lacking full-length MccJ25 with indices with E>33 at all positions were discarded. Unique sequences with >100 reads were isolated and analyzed by a custom MATLAB script. The identity of each mutant within the 96-well plates was processed excluding out-of-library mutants as well as wells with multiple entries identified. For each amino acid mutant, the optimal codon sampled was selected based on that codon's usage in *E. coli* [90].

2.3.4 Generation of multi-site saturation mutagenesis libraries of MccJ25

Oligonucleotides (Table 4) were synthesized with the introduction of degenerate codons at positions 2, 3, 6, 11, 13, and 14. Degenerate codons were selected that included the desired set of mutant amino acids as well as wild-type. Oligonucleotides were assembled via PCR, restriction digested with XhoI and HindIII, ligated into linearized pJP4 as described above and transformed into NEB Express I^{q} cells. Random colonies were selected from transformation plates and used directly for screening as described in the following section.

After screening, active clones from the multi-mutant library were indexed in a 96-well plate format using the same indexing primers described previously. To sequence all variants screened, screened plates were scraped for bacteria and plasmid DNA was purified. Plasmid DNA recovered was used for amplification in the same aforementioned indexed primer set. Illumina MiSeq sequencing (1/8th lane) was then used to determine sequence identities for active variants as well as the total pool of variants screened.

2.3.5 Agar-diffusion assay to measure growth inhibition of MccJ25 and variants

For single-mutant evaluation selected MccJ25 variants were inoculated from fresh colonies into 100 µL of LB in 96 well sterile polystyrene plates and incubated overnight shaking at 250 rpm at 37°C. Separately, pathogens to be evaluated were inoculated into 3 mL of LB and grown overnight at 250 rpm at 37°C. The following day, optical densities were determined, and wells were diluted to a final OD600 value of 1.0 (pathlength corrected to 1.0 cm). In addition, plates containing pathogenic target were prepared. This was done by first spreading 50 mL of LB agar supplemented with IPTG. Once solidified, 25 mL of LB agar (cooled to 42°C) supplemented with IPTG as well as 1000x dilution of pathogenic target grown overnight was spread as a thin layer on the plate. Once solidified, 2 µL of cell suspension was then deposited on target pathogen plates using a multi-channel pipette, allowed to dry, and placed at 37°C for overnight growth. After growth, the zone of inhibition was measured using a custom MATLAB script (see Figure 2.6 for example output).

For initial multi-mutant library panning, plates containing *Salmonella enterica* serovar Enteritidis were prepared as described above. Fresh, randomly selected colonies from the multi-mutant library were transferred onto the pathogen plate and grown overnight at 37°C. Following growth, colonies with any sized zone of growth inhibition surrounding them were classified as active.

For evaluation of active candidates from the multi-mutant screen, fresh colonies of mutants were inoculated into 1 mL of LB supplemented with IPTG in sterile deep 96-well plates and grown at 250 rpm at 37°C for 20 hours. Following this, cells were pelleted at 1000g for 15 minutes. The top 0.5 mL were removed from each well and stored in 1.5 mL sterilized microcentrifuge tubes (supernatant). Supernatant was then sterilized by heating to 98°C for 15 minutes. Plates containing *Salmonella* targets as well as the subset of commensal *E. coli* susceptible to MccJ25 were prepared as described above. 5 µL of sterilized supernatant of each variant were then applied using a multi-channel pipette. Once supernatants had dried, plates were grown overnight at 37°C.

For determination of activity of wild-type MccJ25 and MccJ25^{I13T} against *Salmonella* and commensal *E. coli*, supernatants from producers as prepared before were diluted with spent LB (prepared the same as supernatants described above but with a non-expressing producer cell line). 5 µL of dilutions were plated on *Salmonella* as well as commensal *E. coli* plates as described above, allowed to dry, and grown overnight at 37°C. Minimum inhibitory concentration was determined as a linear fit of the dilution number to diameter of zone of inhibited target growth.

2.3.6 Proteolytic stability assay

Supernatants of MccJ25 and variants, whose production was described previously, were incubated at 60°C for 10 minutes as a 1:1 mixture with phosphate buffered saline (PBS) and varying concentrations of proteinase K (New Englad Biolabs). The samples were then heated to 98°C for 20 minutes to inactive proteinase K. Residual activity of treated supernatants was then determined as described previously using indicator strain *Salmonella enterica* serovar Enteritidis.

2.3.7 Acid stability assay

Supernatants of MccJ25 and variants, whose production was described previously, had pH adjusted to 1.5 using 1 M HCl and were incubated at 37°C for 30 minutes. Following incubation pH was normalized to 7.0 using 1 M NaOH. Residual activity of treated supernatants was then determined as described previously using indicator strain *Salmonella enterica* serovar Enteritidis.

2.3.8 Bactericidal/bacteriostatic assay

Supernatants of MccJ25 and variants, whose production was described previously, were applied to indicator strains, in exponential phase growth at 10^6 colony forming units (cfu)/mL in LB, to a final volume fraction of 20% MccJ25 supernatant. Indicator strains were incubated for one hour at 37°C. Following incubation, cell suspensions were dilution-plated to determine cfu after treatment.

2.3.9 MccJ25 purification

Following [88], supernatants of MccJ25 and variants, whose production was described previously, were vigorously mixed with two volumes of n-butanol. The n-butanol phase was removed and dried under vacuum; dried solute was then resuspended in 200 µL of ultrapure (MilliQ) water. Samples were then analyzed by reverse-phase high-performance liquid chromatography (RP-HPLC) with a XBridge Peptide BEH C18 300 Å column with a 10-90% gradient of acetonitrile in water and 0.1% trifluoroacetic acid. All peaks were isolated and evaluated for activity against indicator *Salmonella* strain. Active peaks were freeze-dried under vacuum, resuspended in ultrapure water and analyzed for identity via matrix-assisted laser desorption/ionization time-of-flight mass spectrometry using an AB Sciex TOF/TOF 5800. Active peaks were re-assessed for activity in ultrapure water against indicator *Salmonella* strain.

2.3.10 Normalized activity analysis

Throughout this text, normalized activity ($N_{i,j}$) for a particular mutant (j) on a particular target organism (i) refers to the normalization of activity data to the activity data of wild-type MccJ25. In the case of the agar-plate assay presented, these wild-type data come from the same plate; thus, this is referred to as intra-plate normalization.

2.3.11 Specific normalized activity analysis

Throughout this text, specific normalized activity refers to the ratio of normalized activity between some organism and the global reference organism (*i.e. Salmonella enterica* serovar Enteritidis in this text). The principle for this normalization is to evaluate the relative effectiveness of a mutant of MccJ25 compared with wild-type MccJ25 in the context of the assays conducted.

2.3.12 Specificity analysis

In the context of the four-pathogen screening, a mutation (j) is deemed to be specific if it causes a statistically significant deviation in the specificity metric compared to wild type. This metric is defined to be:

Equation 1

$$S_j = \left(4 - \frac{\sum_i^4 N_{i,j}}{\max\left(N_{i,j}\right)}\right)\left(\frac{1}{3}\right)$$

There exists no universal metric to evaluate specificity of a process. To enable quantitative analysis of specificity the above metric was generated with the following property: range from zero (all equal normalized activity) to one (only one target having activity greater than zero). The significance of this metric was determined by bootstrap sampling wild-type measurements that were normalized to the intra-plate average activity of two wild-type MccJ25 measurements and performing the above calculation. Based upon this null distribution, values of $S_j > 0.3127$ were deemed significant (p<0.001).

Statistical analysis of specific normalized activity of MccJ25^{I13T}

For statistical analysis of specific normalized activity of MccJ25^{I13T} against commensal *E. coli* relative to *Salmonella enteritidis* a two-sample t-test with unequal variance with the Bonferri correction was utilized. Isolates were deemed significant with p<0.05.

2.4 Results

2.4.1 Library Construction and Validation

A collection of degenerate oligonucleotides was used to synthesize point-mutants of MccJ25 at 12 residues. Six of these sites were in the ring region (positions 2-7) which has been previously identified to be highly sensitive to mutation[74] and form critical interactions with the plug domain of FhuA (Figure 2.1.B)[91,92]. Six other sites from the loop region were selected to incorporate positions immediately following the lactam linked E8 as well as several sites previously identified[74,75] to have high mutational tolerance (positions 9-14). Randomly selected colonies from each sub-library were grown in 96-well plates. The following day, aliquots from each plate were pooled with maintained plate indices into a master 96-well plate. Whole-cell PCR was conducted on each of the wells in isolation (see Table 4 for primers). This process appended row and column indices to each construct which were then used to identify mutant identity via deep sequencing.

Following colony isolation and high-throughput sequencing it was determined that 207 of the 228 (91%) possible point mutations across MccJ25 positions 2-7 and 9-14 were generated via the site saturation mutagenesis. This compares reasonably with the 221 variants (97%) that would be expected by completely random sampling with three-fold coverage of the NNK diversity, as some sampling efficiency was lost due to wells containing multiple mutants.

2.4.2 Single Mutant Activity Analysis

Point mutants were evaluated for growth inhibition, upon recombinant production and secretion, via an agar diffusion assay against two strains of *Salmonella enterica* (serovar Enteritidis (SE) and serovar Tennessee (STen)) and two strains of *E. coli* (JJ1887 and O157:H7). These strains were selected due to their known susceptibility to MccJ25, prevalence in multiple domains as pathogens, and high level of laboratory characterization including available genomic data. Following overnight growth, the sizes of the zones of growth inhibition were measured and normalized to the sizes of wild-type MccJ25 zones on each plate (see Materials and Methods for analysis). These data, summarized in Figure 2.2.A, reveal a range of mutational tolerance. Against all targets, mutations at positions G4 and F10 had no detectable activity, while tolerance was strongly limited at H5 (only R mutant tolerated), P7 (only A mutant tolerated and with notably weaker activity), and F9 (only Y mutant tolerated and with notably weaker activity) (Figure 2.2.B). Conversely, sites 3 and 12-14 exhibit high tolerance, with moderate mutational tolerance observed for sites 2, 6, and 11. In the context of the assay, ring and loop positions showed an average per-residue mutational tolerance of 22% and 51%, respectively (Figure 2.7).

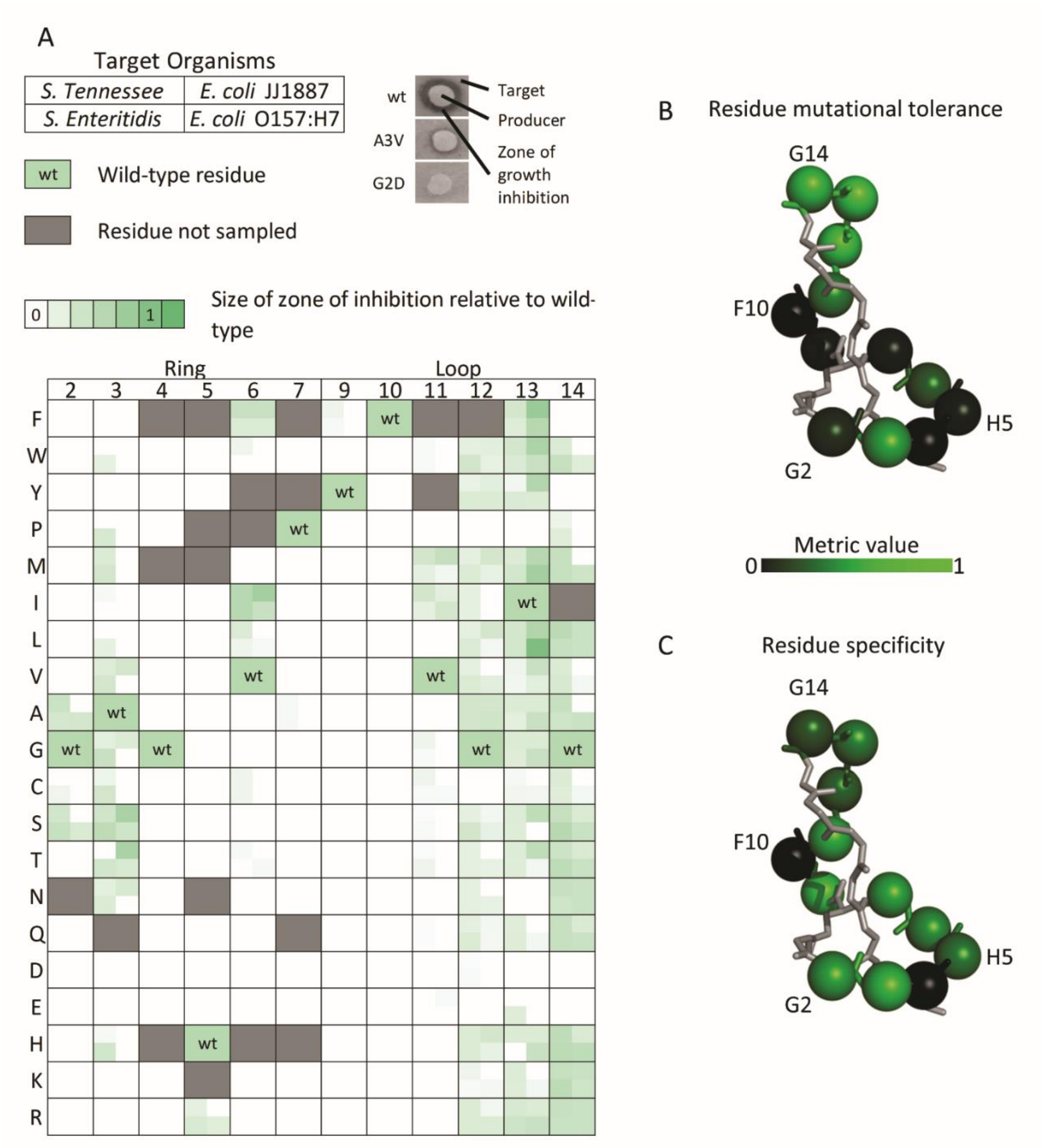


Figure 2.2. Normalized activity of MccJ25 single-site mutants against four targets. (A) The normalized activities of MccJ25 mutants against two strains of *Salmonella* as well as two pathogenic *E. coli* were determined via the size of zones of growth inhibition surrounding colonies of producer cells in solid culture. Each box corresponds to a MccJ25 site and particular amino acid variant (or wild-type as indicated) with four quadrants each that correspond to activity against the four indicated strains. (B-C) The main chain structure of mature MccJ25 with residues highlighted across the mutated position in the ring (2-7), and loop (9-14) as spheres. (B) Mutated positions are colored according to their mutational tolerance, defined as the fraction of mutations showing activity against at least one target.

(C) Mutated positions colored according to the average of the specificity metric for mutants at that position which had activity against at least one target.

Negatively charged mutations are broadly unsupported (active in only 4% of test pairs), whereas positively charged mutations show higher broad tolerance (31%), with near-complete tolerance in positions 12-14 (92%).

Chemical homology between wild-type residues and mutants was predictive of activity differences at positions G2, A3, V6, V11, and I13 ($p < 0.05$; Figure 2.8). Activities of mutants of G12 or G14 were not predictable by homology, despite having high mutational tolerance (84% and 78%, respectively). Tolerated mutations at G2 and A3 were mostly substitutions of small amino acids (A, G, C, S). Against *Salmonella* targets, partial activity was observed for several hydrophobic mutations (W, P, M, I, L, V) of A3.

This multi-target activity evaluation also enables the determination of how mutations impact the normalized specificity. Across the entirety of MccJ25, 63/207 (30%) mutations show statistically significant specificity modulation ($p < 0.001$; Figure 2.3.A). Specificity was negatively correlated with the maximum activity against the four targets ($p < 0.001$; Figure 2.3.B). Mutants in the ring region active against at least one target showed higher specificity than mutants in the loop ($p < 0.01$; Figure 2.2.C). Chemical homology was not found to be predictive of specificity (Figure 2.9).

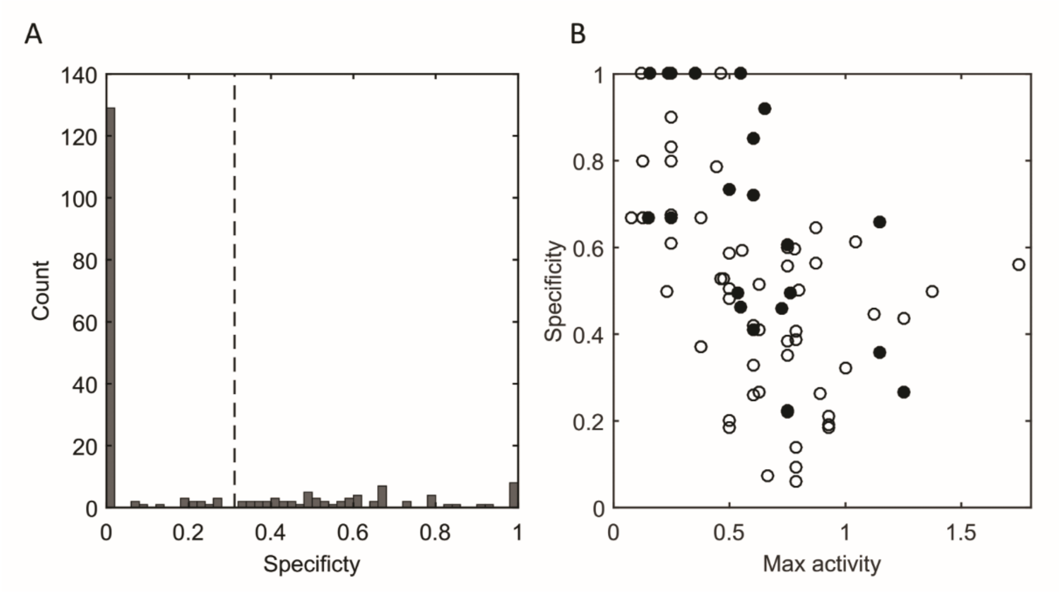


Figure 2.3. Mutant specificity and correlation with maximum activity. (A) The computed specificity metric across all mutants (including non-active, which have specificity of zero) is presented. The dashed line indicates the 99.999 percentile of the bootstrap sampled wild-type distribution which was defined as the null. Mutants with specificity greater than this were deemed statistically significant from 0. (B) The specificity of active mutants is plotted versus the maximum activity and specificity for mutations with a max activity above 0 for ring (closed circle), and loop (open circle) positions. Spearman's rho: -0.58 (p = 2.3x10-8). When comparing the specificities of the ring and loop positions, it is found that the mutations in the ring confer greater specificity than those of the loop when considering active mutants (p = 0.0016, Wilcoxon rank-sum right-tailed statistic, $\mu_{ring}>\mu_{loop}$).

2.4.3 Multiple Mutant Activity Analysis

Utilizing the single mutant activity data, a collection of mutations across both the ring and loop regions were combined in a multi-mutant library. To design the library, the set of mutants with an average activity against SE and STen of at least 0.5 and a significant specificity metric were identified. A3T was selected as an exception due to differential response between the two *Salmonella* strains. Three positions in each region were selected: residues 2, 3, and 6 from the ring and residues 11, 13, and 14 from the loop. Due to higher specificity metrics, ring positions were preferentially given more phenotypic variation (4, 5, and 3 options) relative to loop positions (2, 4, and 2 options).

Repertoire for each residue was selected with preference for SE and STen as targets with reduced activity against one or both pathogenic *E. coli* strains. G2T and A3D were included because of codon degeneracy and not selected for attributes of interest. G12 remained fixed in contrast to 13 and 14 to reduce library diversity and give preference to the top of the loop region (Table 1). In addition to aforementioned mutational options, library degeneracy included all wild-type residues.

Table 1. Multi-mutant saturation mutagenic library of MccJ25. The composition of the multi-mutant library. Each of the listed positions was encoded using a set of degenerate oligonucleotides to include all combinations of the wild-type and mutant amino acids. The genetic diversity as well as screening results for activity are included.

Position	Wild-Type	Mutants
2	G	T,S,A
3	A	M,N,D,T
6	V	F,L
11	V	M
13	I	Y,S,T
14	G	A

Theoretical Diversity	960
Sampled Colonies	2000
Unique In-library Sequences	919
Active Sequences	32

From this library of 960 possible unique sequences, 2000 randomly selected colonies were screened for activity against SE with 64 colonies showing some level of activity (Table 1). The relative rarity of active clones was surprising given the library was composed of combinations of active single-mutants (extended observation provided in *Discussion*). Following Illumina deep sequencing of active and inactive colonies, it was found that the screened set contained 919 unique in-library sequences covering 96% of the designed library sequence space. There was a total of 32 unique sequences (31 mutants and wild-type, Table 2) which showed functional activity against SE in the assay. Six mutants of the 12 sampled active single-mutants did not show activity in this assay whereas they did in the first (G2S, A3N, V6F, V6L, I13S, I14A). Codon usage is unable to explain these differences (p > 0.1; Table 5). Though unidentified as single mutants, each of these single mutants was represented in variants with 2 or more mutations.

Table 2. Active mutants from multi-mutant saturation mutagenesis library. Sequences of active variants from the multi-mutant initial screening with at least two mutations are listed. Blank cells are wild-type amino acids. The activities of these variants against *Salmonella enterica* serovar enteritidis (SE) are listed.

Position																					Activity against SE[†]
1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	
Wild-type MccJ25																					
G	G	A	G	H	V	P	E	Y	F	V	G	I	G	T	P	I	S	F	Y	G	+++++
Active multi-mutant																					
	A											Y									++
					F							T									+
					F							Y									++
					L							Y									+
												S	A								+
		N			F																++
		N								M											+
		N											A								+
		N										S									+
		T			F																+
		T											A								+++
		T										T									++++
		T										Y									+++
	S				F																++++
	S											T									+
	S	N																			+
	A	N										Y									+
	A	T			F																+
					F							T	A								+
		T								M		Y									+
	S									M		Y									+
	S	N										Y									+
	S	T										Y									+
	A	M			L							T									+

[†] Activity against SE was determined by zone-of-inhibition size for each mutant and indicated in one of five bins <20%, <40%, <60%, <80%, and 100% of wild-type MccJ25 indicated by +, ++, +++, ++++, and +++++, respectively.

The activity of MccJ25 and all active multi-mutants were tested against randomly selected human commensal *E. coli* isolates. These isolates are identified via genetic screening to be probable non-extraintestinal pathogenic *E. coli*[93], representing non-pathogenic commensal *E. coli* and are gathered from both urine and fecal samples. Of these 20 isolates, 18 were found to have susceptibility to wild-type MccJ25 in a liquid-culture growth inhibition assay (Figure 2.10).

Those isolates which showed susceptibility were tested against heat-sterilized supernatant derived from active variants from the multi-mutant MccJ25 library (examples from multi-mutant screening stages displayed in Figure 2.11). It was found that the I13T single mutation significantly reduces the specific normalized activity against 15 of 18 of the strains, susceptible to MccJ25 in liquid culture, relative to SE ($p<0.05$; Figure 2.4.A-C). Assessment of MccJ25 and I13T for bactericidal/bacteriostatic action demonstrated similar responses against three commensal *E. coli* strains with a range of MccJ25 responses as well as the indicator *Salmonella* strains (Figure 2.4.D). On average, this mutation reduced the specific normalized activity against the 18 evaluated commensals by 81%. This mutation also maintained >50% supernatant activity compared to wild-type MccJ25 against SE. This contrasted with the majority of multi-mutants which had considerable lower normalized activity against SE (Table 2).

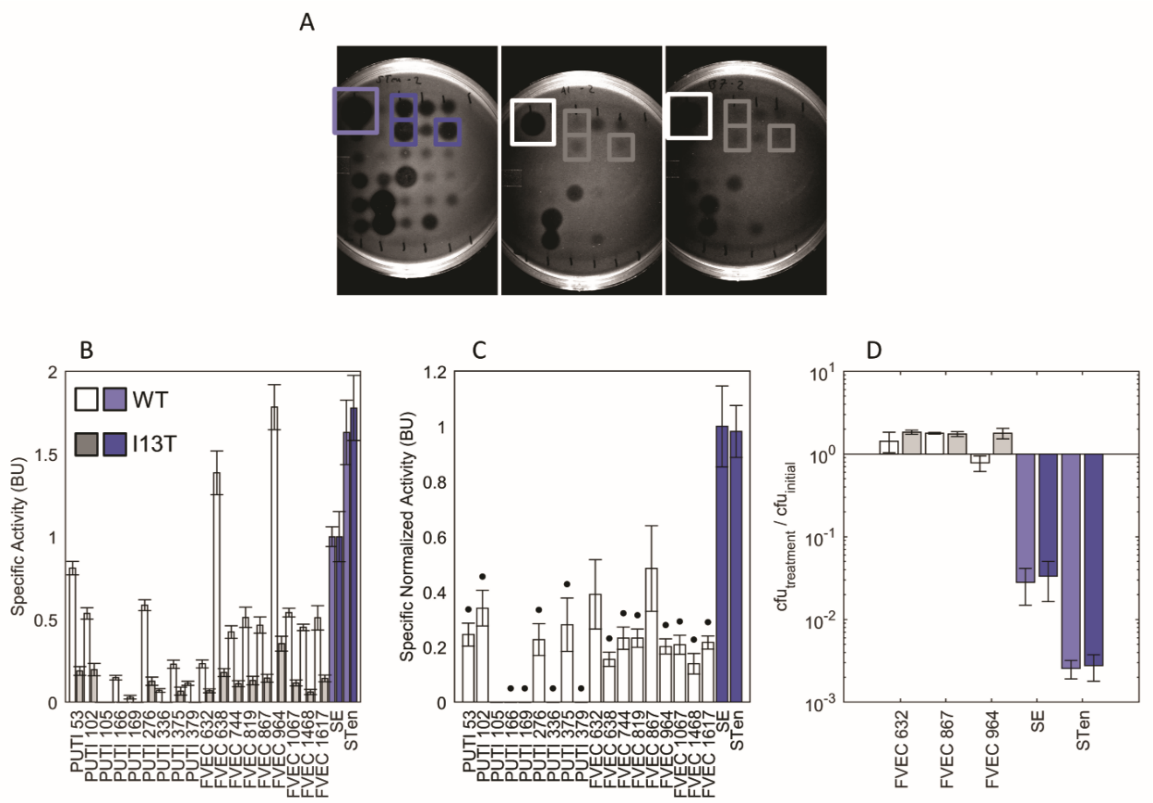


Figure 2.4. MccJ25^{I13T} demonstrates improved specificity. For pre-purified activity quantification MccJ25 mutants were first grown and induced in liquid culture, then the supernatants were isolated and sterilized by heating to 98 ºC for 15 min. and then plated on agar containing target organisms yielding a spectrum of responses (A, left: STen, middle: PUTI 53, right: FVEC 964). The specific activity, assessed via dilution plating, of wild-type MccJ25 (B, light bars) and MccJ25^{I13T} (B, dark bars), as well as the specific normalized activity (C) are shown against 18 commensal *E. coli* human isolates (grey bars) as well as well as two strains of *Salmonella* (blue bars). Data use *Salmonella enterica* serovar enteritidis (SE) as a reference point. These data demonstrate significant reduction of specific normalized activity for 15 of the 18 *E. coli* isolates ($p<0.05$). (D) To differentiate bactericidal from only bacteriostatic response, a subset of commensal *E. coli* and both *Salmonella* were incubated for one hour in exponential phase growth with MccJ25 or MccJ25^{I13T} and colony forming units (cfu) determined. A value below unity (10^0) is indicative of bactericidal activity.

MccJ25 and MccJ25^{I13T} exhibit good thermal stability as evidenced by their activity after heat sterilization (98ºC for 15 min) in the previous assay. Stabilities in the presence of protease and acidic conditions were also evaluated. MccJ25^{I13T} and MccJ25 are essentially equivalent in protease tolerance up to the highly stringent condition of 0.25

U/mL proteinase K at 60ºC for 10 minutes. At more extreme conditions, MccJ25^{I13T} does lose activity, dropping to zero with 1 U/mL protease whereas MccJ25 retained 50% residual activity up to 4 U/mL protease (Figure 2.5.A). To assess stability in acid conditions, supernatants of both MccJ25 and MccJ25^{I13T} were incubated at pH 1.5 for 30 min via HCl addition and returned to pH 7.0 via NaOH addition. Both peptides retained full activity (Figure 2.5.B). The activities of purified MccJ25 and MccJ25^{I13T}, extracted and purified from supernatant, were assessed against SE, demonstrating activity of the pure form of both (Figure 2.12).

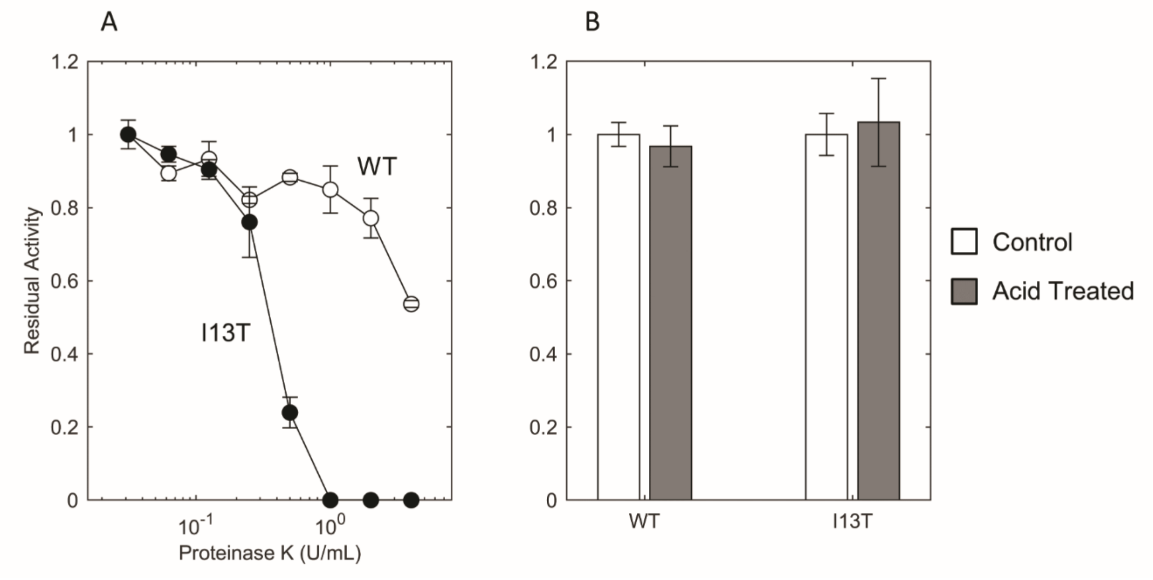


Figure 2.5. Proteolytic and acid stability of MccJ25 and MccJ25^{I13T}. (A) Supernatant of wild-type MccJ25 (WT, open circles) and MccJ25^{I13T} (I13T, filled circles) were incubated at 60 °C for 10 minutes in the presence of varying concentrations of Proteinase K followed by Proteinase K inactivation by incubating at 98°C for 20 minutes. These processed supernatants were then deposited on SE-seeded LB agar plates. Following growth, the size of the zones of growth inhibition were measured to determine residual activity. (B) Supernatant of WT and I13T were incubated at 37°C for 30 minutes at pH 1.5 (via HCl, grey bars) followed by normalization to pH 7.0 (via NaOH). Residual activity was then determined using SE as an indicator strain.

2.5 Discussion

Significant effort remains to develop platforms for next generation antimicrobials. The use of ribosomally synthesized AMPs offers the advantage to explore functional sequence space using straightforward genetic manipulation toolkits. Though many efforts have explored the functional sequence landscape of ribosomally synthesized AMPs, few studies have focused on the modulation of specificity of these proteins. Improved specificity for target pathogens can reduce the pressure for AMP resistance development against off-target bacteria as well as improve patient standard of care by reducing microfloral disruption.

Through the evaluation of a collection of single- and multi-mutants of MccJ25, an I13T mutation was identified which significantly reduces activity towards 15 of 18 randomly selected *E. coli* isolates from humans relative to activity against SE. Though wild-type MccJ25 specificity and efficacy is an improvement over tradition broad-spectrum antibiotics, its high efficacy towards commensal *E. coli* could pose an issue for human application. Of the *E. coli* isolates tested, 90% have susceptibility to MccJ25, 55% have susceptibility at least half that of SE and 10% show higher susceptibility. In contrast, 83% of MccJ25-susceptible isolates have reduced susceptibility to MccJ25^{I13T} relative to SE, with an average reduction of 81%.

All testing was done using assays that measure total activity, the combination of production rate and per-molecule activity. The high specificity modulation of ring-position mutants, possibly driven by interactions during binding to the plug domain of FhuA, lost out in the screening performed likely because of the significant losses to activity these mutations acquired. Future studies could explore promising ring-position mutations further to evaluate per-molecule activity.

In addition to the specificity modulation, it should be noted that though the mutational tolerance data provided by singletons is complementary of work done by Pavlova et al., there are several deviations. Several mutations at position G2 (A, C, S) were found to be active in the agar diffusion assay used here, however they were not detected for production, maturation, export, and stability by Pavlova et al. It is possible that the expression system in this work may produce larger quantities of MccJ25 and variants than the naturally-occurring gene cluster utilized in Pavlova et al.'s work. The gene cluster used in this work is derived from the cluster developed by Pan and co-workers[88].

As demonstrated in Figure 2.3.B, there is an inverse relationship between specificity and activity for single-mutants of MccJ25. Though unexplored in AMP research, this trade-off is well known in other protein classes, most notably enzymes[94]. This tradeoff is due to the large combinatorial landscape of proteins, resulting in multi-function optima being rare. A direct consequence of this property is the lack of multi-mutants retaining high levels of activity, further highlighting the necessity of high-throughput screening to discover mutants with both specificity as well as high activity.

MccJ25 and MccJ25^{I13T} exhibit high stability under thermal, acidic, and proteolytic stresses. Though MccJ25^{I13T} showed higher susceptibility to proteinase K, the conditions of the assay (60°C for 10 minutes, high concentration of nonspecific protease) were elevated in comparison to physiological conditions in order to stress the peptides.

In this work it was demonstrated that mutagenic libraries can be used to identify variants of MccJ25 with improved specificity towards pathogenic targets over commensal organisms. This study resonates with previous work regarding the flexibility of MccJ25's loop region to functional modification[75,95,96] suggesting a capacity of MccJ25 to be tailored to applications of interest. Though these methods are laborious, scaling linearly with target screening, they offer tremendous opportunity to tune AMPs for use in treating human infections. For many AMPs, in particular bacteriocins produced by gram-negative bacteria, the lack of information provided by homologous sequences, solved structures, or descriptions of uptake and mode of action, necessitates studies such as this to provide insight which can be utilized to design sequences with desired properties.

2.6 Acknowledgements

This work was supported by grants from the National Institutes of Health (GM121777 and GM111358) and a grant from the National Science Foundation (CBET-1412283). S.C.R was also supported by a predoctoral National Institutes of Health traineeship (T32GM008347).

We thank Dr. Timothy Johnson and Dr. Michael Sadowsky of the University of Minnesota for their donations of pathogenic *E. coli* and *Salmonella* strains used in this study. In addition, we thank Dr. James Johnson of the Veterans Affairs Hospital of Minneapolis for his donation of pathogenic and commensal *E. coli* strains used in this study. We thank Dr. James Link of Princeton for providing the pJP3 vector. Support from the University of Minnesota Genomics Center, the Digital Technology Center, and the Biotechnology Institute is gratefully acknowledged.

2.7 Competing financial interests

Y. Kaznessis is the founder and president of General Probiotics, Inc., a startup company that plans to commercialize technologies based on antimicrobial probiotics. This interest has been reviewed and managed by the University of Minnesota in accordance with its Conflict of Interest policies.

2.8 Supporting Information

2.8.1 Algorithm to measure zones of inhibition

Sizes of zones of inhibition were measured using a custom MATLAB script (see attached code and example). This script works according to the following pseudocode:

1. Import image, downscale and apply a median smoothing to reduce noise
2. User selects starting pixel within the colony of the producing bacteria
3. Algorithm steps out to adjacent pixels, adding them to the set of colony pixels if their intensity is above some fraction (user defined, default to 0.95) of the average of added colony pixels.
4. Pixels which are surrounded by colony pixels are automatically added to the colony pixels and assumed to be image errors or inconsistency in growth region.
5. Algorithm then steps out from boundary of colony pixels to add zone of inhibition pixels, adding pixels below some user fraction of the average (default to 1.0).
6. Surrounded pixels are added as well
7. Program outputs the size of each zone in pixels

Below is an example output:

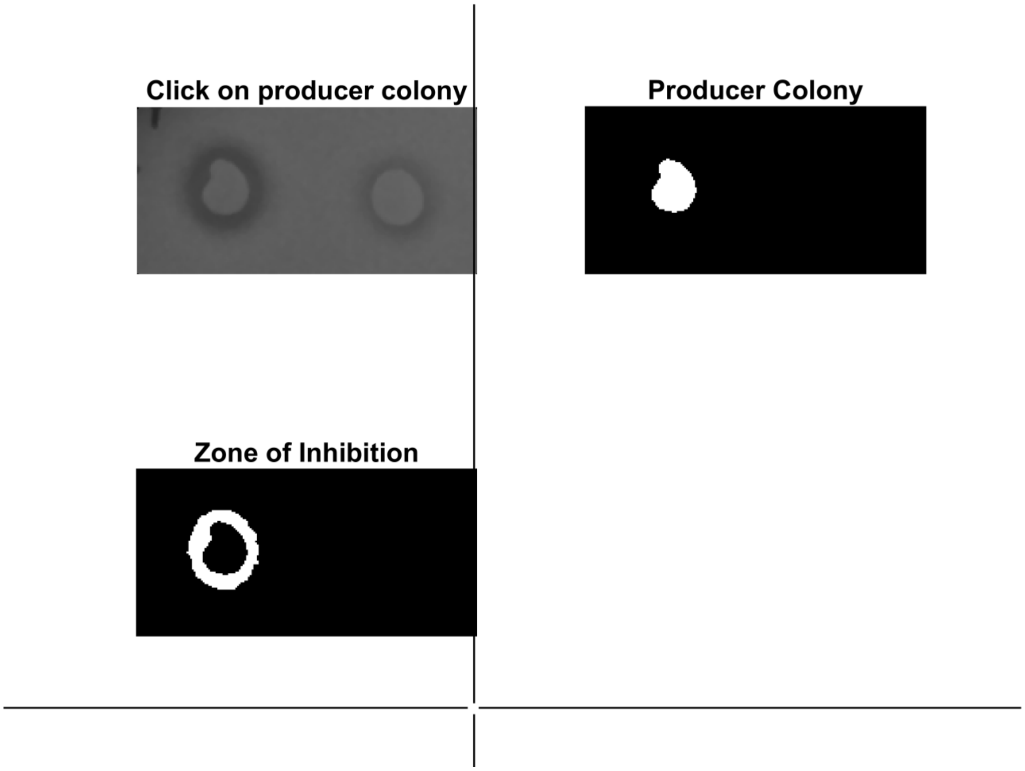


Figure 2.6. Automated measuring of zones of inhibition (Top-left) Imported and smoothed image taken by user showing two producer colonies surrounded by zones of inhibition as a result of antimicrobial production. (Top-right) Upon selection of the left producer colony, the algorithm generates the displayed producer colony (size = 531 pixels). (Bottom-left) The algorithm displays the zone of inhibition (size = 807 pixels).

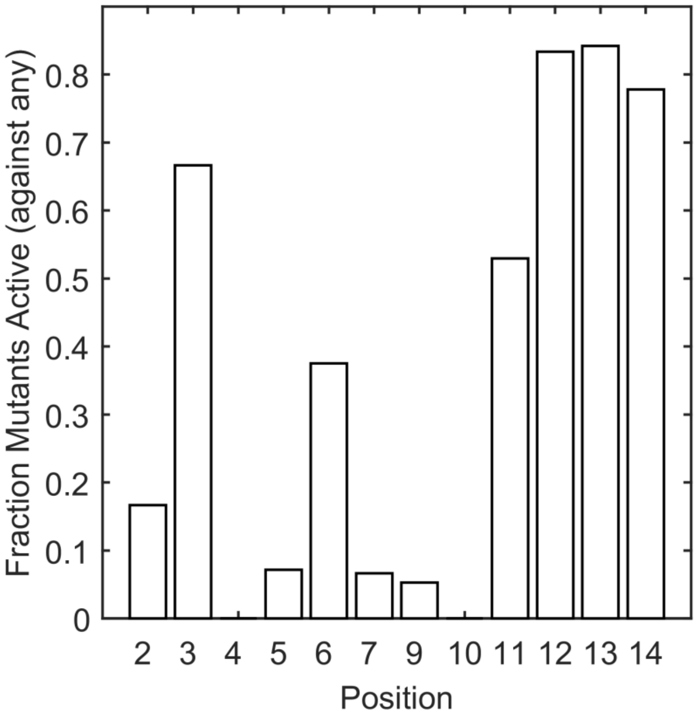


Figure 2.7. Per-residue mutational tolerance. Displayed are the fractions of sampled mutants (n = 14-18 per site) which retained detectable activity against any of the four targets are presented. The analysis finds that loop positions 11-14 form the continuous set of residues with the highest mutational tolerance. Positions 7 and 9 are known to have a low tolerance for mutation. Pavlova et al showed that mutating positions 7 and 9 reduced RNAP inhibition.

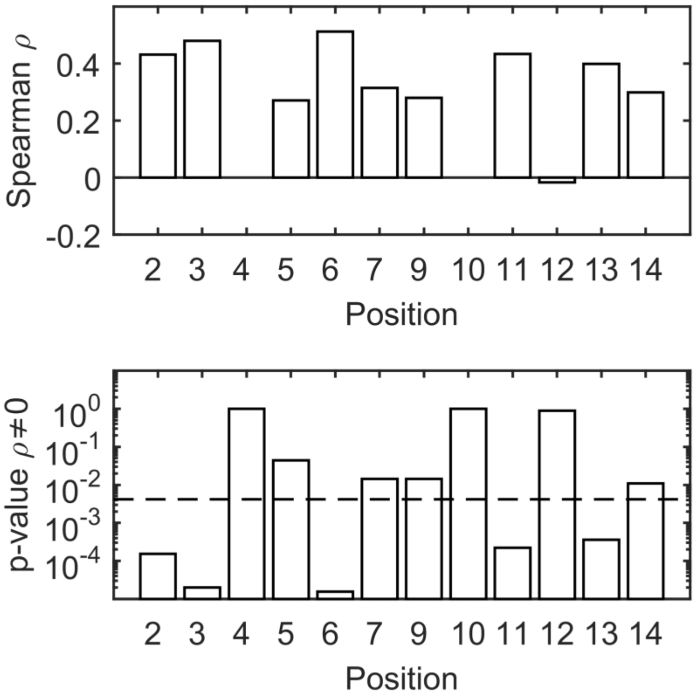


Figure 2.8. Relationship between chemical homology and MccJ25 mutant activity per residue.

Chemical homology, a global relationship of the tendencies of different amino acid substitutions at evolutionarily related sites, was assessed as a predictor of activity of MccJ25 single-mutants. For each position the correlation between activities, against all four pathogenic targets of mutants sampled, and the BLOSUM 90 score associated with the (wild-type, mutant) pair was analyzed.

It is found (Figure 2.8) that 5 of 12 positions tested show a significant Spearman's correlation ($p<0.05$, Bonferri corrected, dashed line bottom plot. Computed using exact permutation.). Position 2 shows a preference for small amino acids, G2[A,C,S]. Position 3 shows a preference for small amino acids, A3[G,C,S,T], but also supports many others. However, many mutations show differences between the different targets, implying that this interaction may serve to enable some specificity. Though position 5 showed conservation amongst mutations sampled, as well as supporting literature data[91] (requiring charged residue), only 1 other mutation was tolerated and sampled (H5R), which was insufficient evidence for testing. Position 6 shows a preference for hydrophobic residues, V6[L,I,W,F]. Position 11 shows a preference for hydrophobic residues, V11[I,M,W]. Position 13 shows a small preference for hydrophobic residues, I13[F,W,Y,M,L,V,A], but also tolerates many hydrophilic mutations. Surprisingly, though positions 12 and 14 show a high tolerance for mutation, they don't show a relationship between the BLOSUM90 scores for mutants in reference to wild-type (Glycine). These data could be explained by the lack of a meaningful interaction between residues 12 and 14 of MccJ25 and any proteins in the host involved in transport or activity.

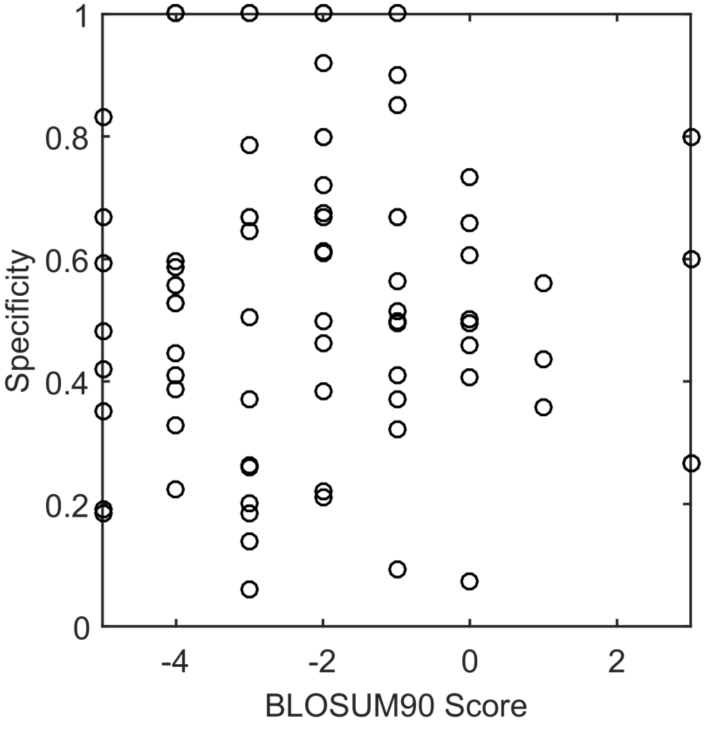


Figure 2.9. Correlation between specificity and chemical homology.

Similar to the analysis done in Figure 2.8, the relationship between chemical homology and specificity for each single-mutant was assessed (Figure 2.9). There is insufficient evidence that chemical homology between wild type and mutant predict the specificity (p > 0.1).

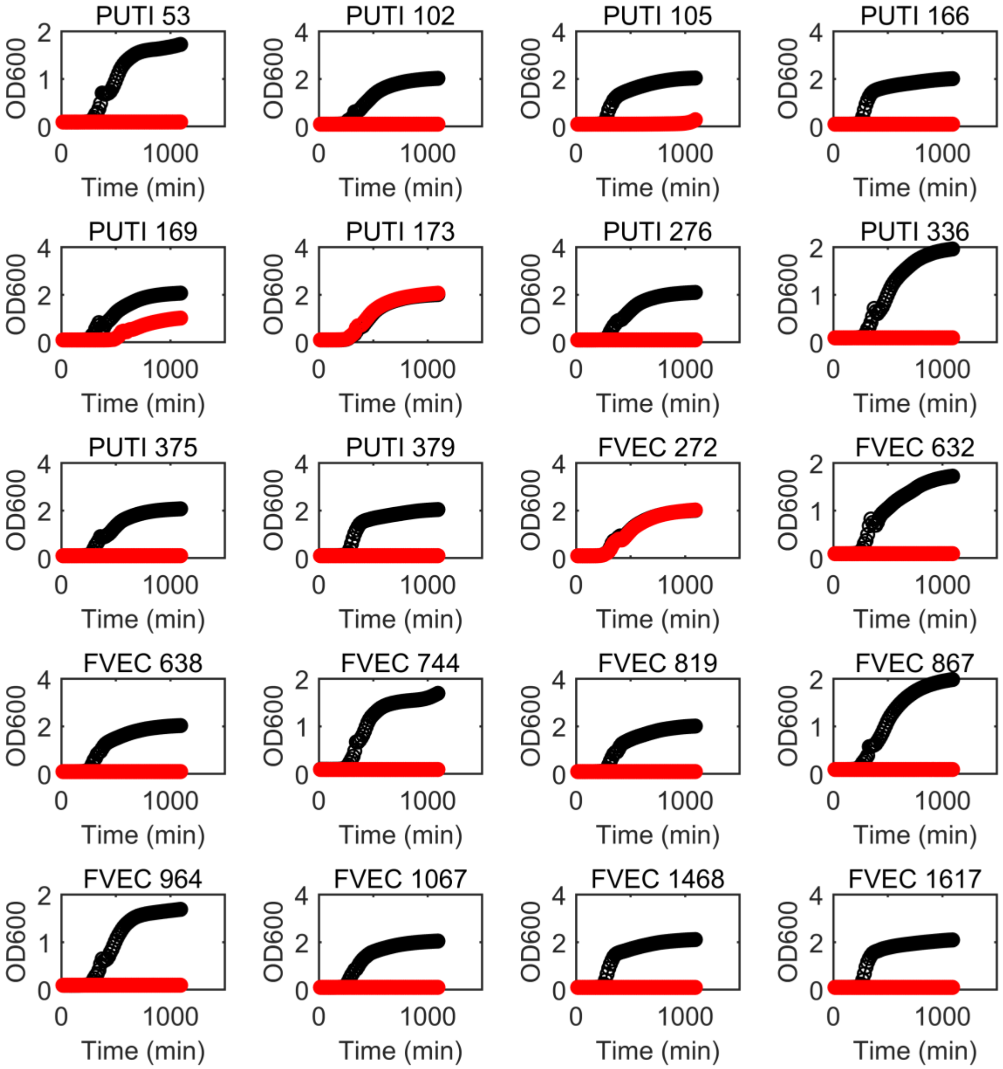


Figure 2.10. Evaluation of commensal E. coli susceptibility to wild-type MccJ25. Commensal *E. coli* isolates were grown overnight to stationary phase in LB. The following day they were diluted 1000x in fresh LB supplemented to 20 vol% sterilized supernatant from a MccJ25 -producing cell line. Their growth was then monitored at 37 °C via optical density readings to observe the impact of MccJ25 -containing supernatant on growth. Of the 20 isolates evaluated, 18 showed susceptibility to MccJ25. The above are averages

of 2 replicates per isolate exposed to both MccJ25 (red) and stop-codon mutated MccJ25 (black) supernatant. PUTI 173 and FVEC 272 do not show susceptibility to MccJ25.

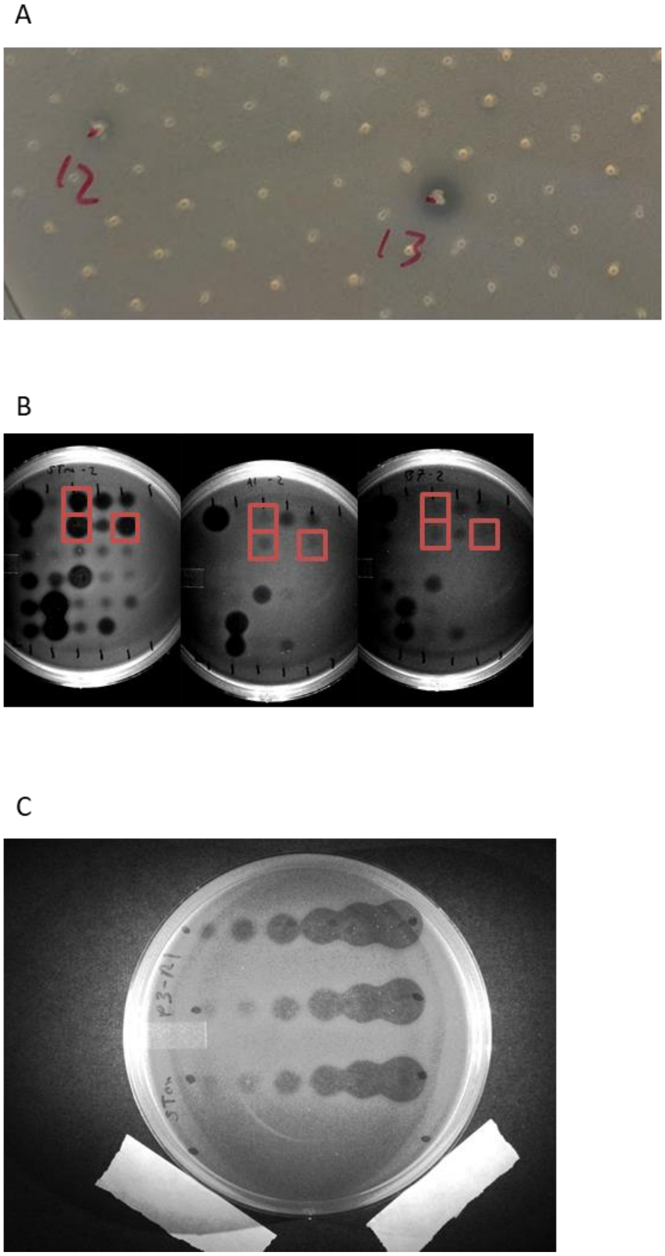


Figure 2.11. Screening strategy for multi-mutant MccJ25 library. (A) Colonies of randomly-selected producers of mutants of MccJ25 were transferred to LB-agar plates seeded with SE and IPTG. Following outgrowth, colonies showing zones of inhibition were gathered. (B) Heat-sterilized supernatants produced from the 64 active colonies from (A) were assayed for activity against pathogenic strains as well as human commensal *E. coli*. In the examples displayed, the orange boxes outline clones that were sequence-verified

to be MccJ25^{I13T}. The differential response in activity between STen (left) and two human commensal *E. coli* strains (middle, right) can be seen. (C) Activities of supernatants of candidate producers showing promising specificity modulation in (B) were quantified via dilution-series against all pathogens and human commensal *E. coli*.

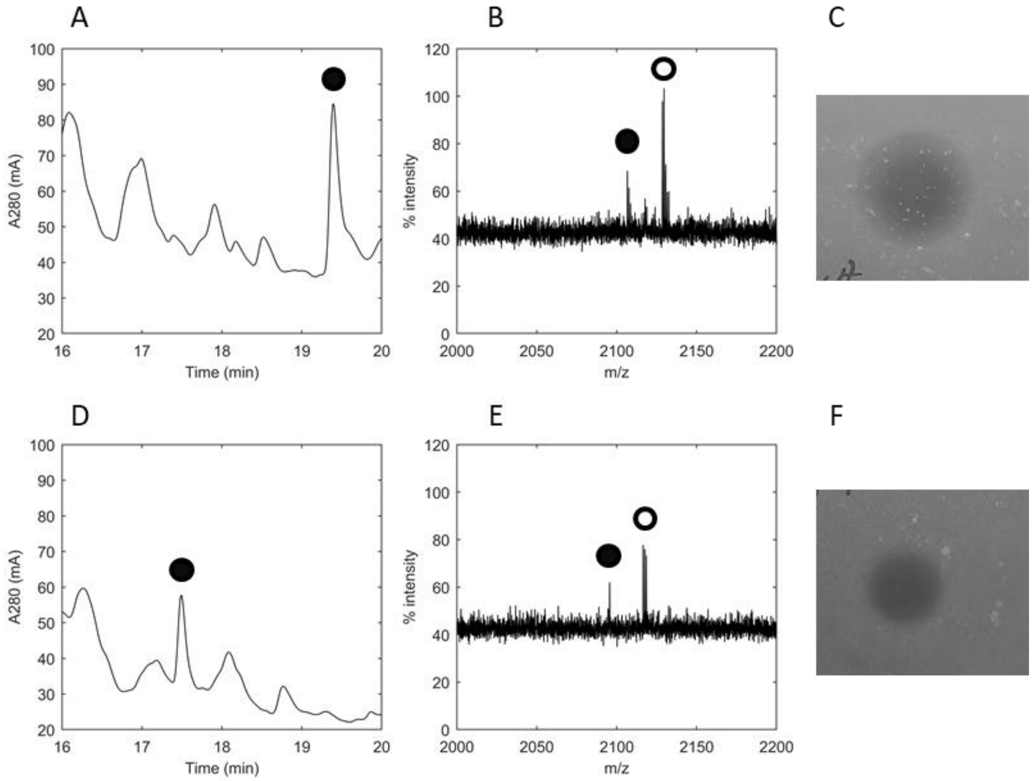


Figure 2.12. Assessment of activity of purified MccJ25 and MccJ25^{I13T}. MccJ25 (A-C) and MccJ25^{I13T} (D-F) were purified from supernatant of expressing cells. Following a 2-volume n-butanol extraction, samples were dried and run on RP-HPLC (A, D). All peaks absorbing at 280 nm were collected and evaluated for activity. Only the indicated peaks (closed circles) displayed activity against the *Salmonella* indicator strain. To assess identity, active peaks were freeze-dried under vacuum, resuspended in ultrapure water and evaluated via matrix-assisted laser-desorption ionization mass spectrometry (B, E). The processed form (closed circle) and linear form (open circle) of MccJ25 and MccJ25^{I13T} were identified. Previous efforts in the literature have demonstrated that the linear form is not active[97]. Water-solubilized purified peptides exhibited activity against *Salmonella* indicator strain in an agar plate assay analogous to Figure 2.4 (C, F).

Table 3. Bacteria and plasmids used.

Bacterial Strain or Plasmid	Description	Reference or Source
NEB Express I^q	BL21 derivative, no antimicrobial peptides or proteins, overexpresses LacI	New England Biolabs
Pathogenic Targets		
Salmonella enterica serovar Enteritidis	#MH9189	Timothy Johnson, Veterinary and Biological Science, University of Minnesota
Salmonella enterica serovar Tennessee	#ST101	Timothy Johnson, Veterinary and Biological Science, University of Minnesota
Escherichia coli JJ1887	Urinary tract infection causing	J. Johnson, Veterans Affairs Hospital, Minneapolis, MN
Escherichia coli O157:H7	#2026	Michael Sadowsky, Biotechnology Institute, University of Minnesota
Commensal Targets		
Escherichia coli PUTI 53	Human isolate. Sourced sample type: Urine	J. Johnson, Veterans Affairs Hospital, Minneapolis, MN
Escherichia coli PUTI 102	Human isolate. Sourced sample type: Urine	J. Johnson, Veterans Affairs Hospital, Minneapolis, MN
Escherichia coli PUTI 105	Human isolate. Sourced sample type: Urine	J. Johnson, Veterans Affairs Hospital, Minneapolis, MN
Escherichia coli PUTI 166	Human isolate. Sourced sample type: Urine	J. Johnson, Veterans Affairs Hospital, Minneapolis, MN
Escherichia coli PUTI 169	Human isolate. Sourced sample type: Urine	J. Johnson, Veterans Affairs Hospital, Minneapolis, MN
Escherichia coli PUTI 173	Human isolate. Sourced sample type: Urine	J. Johnson, Veterans Affairs Hospital, Minneapolis, MN
Escherichia coli PUTI 276	Human isolate. Sourced sample type: Urine	J. Johnson, Veterans Affairs Hospital, Minneapolis, MN

Escherichia coli PUTI 336	Human isolate. Sourced sample type: Fecal	J. Johnson, Veterans Affairs Hospital, Minneapolis, MN
Escherichia coli PUTI 375	Human isolate. Sourced sample type: Fecal	J. Johnson, Veterans Affairs Hospital, Minneapolis, MN
Escherichia coli PUTI 379	Human isolate. Sourced sample type: Fecal	J. Johnson, Veterans Affairs Hospital, Minneapolis, MN
Escherichia coli FVEC 272	Human isolate. Sourced sample type: Fecal	J. Johnson, Veterans Affairs Hospital, Minneapolis, MN
Escherichia coli FVEC 632	Human isolate. Sourced sample type: Fecal	J. Johnson, Veterans Affairs Hospital, Minneapolis, MN
Escherichia coli FVEC 638	Human isolate. Sourced sample type: Fecal	J. Johnson, Veterans Affairs Hospital, Minneapolis, MN
Escherichia coli FVEC 744	Human isolate. Sourced sample type: Fecal	J. Johnson, Veterans Affairs Hospital, Minneapolis, MN
Escherichia coli FVEC 819	Human isolate. Sourced sample type: Fecal	J. Johnson, Veterans Affairs Hospital, Minneapolis, MN
Escherichia coli FVEC 867	Human isolate. Sourced sample type: Fecal	J. Johnson, Veterans Affairs Hospital, Minneapolis, MN
Escherichia coli FVEC 964	Human isolate. Sourced sample type: Fecal	J. Johnson, Veterans Affairs Hospital, Minneapolis, MN
Escherichia coli FVEC 1067	Human isolate. Sourced sample type: Fecal	J. Johnson, Veterans Affairs Hospital, Minneapolis, MN
Escherichia coli FVEC 1468	Human isolate. Sourced sample type: Fecal	J. Johnson, Veterans Affairs Hospital, Minneapolis, MN
Escherichia coli FVEC 1617	Human isolate. Sourced sample type: Fecal	J. Johnson, Veterans Affairs Hospital, Minneapolis, MN
Plasmid		
pJP3	Microcin J25 expression cassette containing plamid. McjA under Laco/T5 for controlled induction.	Link group, Princeton [88]

pJP4	pJP3 variant with XhoI moved downstream and optimized ribosomal binding site	This work

Table 4. Sequences for genes and oligonucleotides used.

pJP4	
Modified mcjA expression cassette	CCCTTTCGTCTTCACCTCGATCGATCATAAAAAATTTATTTGCTTT GTGAGCGGATAACAATTATAATACTCGAGGCGCC AGTCTCCCCATAAGGAGGTTAACATACATGATCAAACATTTTCACT TCAACAAACTGTCAAGCGGTAAGAAGAATAATGT TCCGAGCCCAGCAAAGGGAGTGATTCAGATTAAGAAGAGCGCCT CGCAATTAACGAAGGGCGGTGCTGGTCATGTCCCTG AATATTTCGTGGGCATCGGGACCCCAATCTCCTTCTATGGGTAAA AGCTTAATTAGCTGAGCTTGGACTC
Single-site saturation mutagenesis	
mcjaFWD1	AATTCTCGAGGCGCCAGTCTCCCCATAAGGAGGTTAACATACATGATCAAACATTTTCAC
mcjaFWD2	TACATGATCAAACATTTTCACTTCAACAAACTGTCAAGCGGTAAGAAGAATAATGTTCCG
mcjaFWD3	GTAAGAAGAATAATGTTCCGAGCCCAGCAAAGGGAGTGATTCAGATTAAGAAGAGCGCCT
mcjaFWD4	TCAGATTAAGAAGAGCGCCTCGCAATTAACGAAG
mcjaFWD4_lib_3'_set1	GAATATTTCGTGGGCATCGGGACCCCAATCTCCTTCTATGGGTAAAAGCTTAGCCGACCG
mcjaFWD4_lib_5'_set2	GCGCCTCGCAATTAACGAAGGGCGGTGCTGGTCATGTCCC
mcjaFWD4_lib_3'_set2	ACCCCAATCTCCTTCTATGGGTAAAAGCTTAGCCGACCG
mcjaFWD4_lib_5'_set3	GGCGGTGCTGGTCATGTCCCTGAATATTTCGTGGGCATCG
mcjaFWD_lib_5' _lib2	GCGCCTCGCAATTAACGAAGGGCNNKGCTGGTCATGTCCCTGAATATTTCGTGGGCATCG
mcjaFWD_lib_5' _lib3	GCGCCTCGCAATTAACGAAGGGCGGTNNKGGTCATGTCCCTGAATATTTCGTGGGCATCG
mcjaFWD_lib_5' _lib4	GCGCCTCGCAATTAACGAAGGGCGGTGCTNNKCATGTCCCTGAATATTTCGTGGGCATCG
mcjaFWD_lib_5' _lib5	GCGCCTCGCAATTAACGAAGGGCGGTGCTGGTNNKGTCCCTGAATATTTCGTGGGCATCG
mcjaFWD_lib_5' _lib6	GCGCCTCGCAATTAACGAAGGGCGGTGCTGGTCATNNKCCTGAATATTTCGTGGGCATCG
mcjaFWD_lib_5' _lib7	GCGCCTCGCAATTAACGAAGGGCGGTGCTGGTCATGTCNNKGAATATTTCGTGGGCATCG
mcjaFWD_lib_5' _lib9	GGCGGTGCTGGTCATGTCCCTGAANNKTTCGTGGGCATCGGGACCCCAATCTCCTTCTAT
mcjaFWD_lib_5' _lib10	GGCGGTGCTGGTCATGTCCCTGAATATNNKGTGGGCATCGGGACCCCAATCTCCTTCTAT
mcjaFWD_lib_5' _lib11	GGCGGTGCTGGTCATGTCCCTGAATATTTCNNKGGCATCGGGACCCCAATCTCCTTCTAT

mcjaFWD_lib_5' _lib12	GGCGGTGCTGGTCATGTCCCTGAATATTTCGTGNNKATCGGGACCCCAATCTCCTTCTAT
mcjaFWD_lib_5' _lib13	GGCGGTGCTGGTCATGTCCCTGAATATTTCGTGGGCNNKGGGACCCCAATCTCCTTCTAT
mcjaFWD_lib_5' _lib14	GGCGGTGCTGGTCATGTCCCTGAATATTTCGTGGGCATCNNKACCCCAATCTCCTTCTAT
REV_amplify_10_03_16	CGGTCGGCTAAGCTTTTA
Multi-site saturation mutagenesis	
Lib3s1_Lib13s1	CCTCGCAATTAACGAAGGGCRSCRMCGGACACBTCCCAGAATACTTCRTGGGATMCGSAACCCCAATCTCCTTCTATGG
Lib3s2_Lib13s1	CCTCGCAATTAACGAAGGGCRSCATGGGACACBTCCCAGAATACTTCRTGGGATMCGSAACCCCAATCTCCTTCTATGG
Lib3s1_Lib13s2	CCTCGCAATTAACGAAGGGCRSCRMCGGACACBTCCCAGAATACTTCRTGGGAAYAGSAACCCCAATCTCCTTCTATGG
Lib3s2_Lib13s2	CCTCGCAATTAACGAAGGGCRSCATGGGACACBTCCCAGAATACTTCRTGGGAAYAGSAACCCCAATCTCCTTCTATGG
FWD1	TATAATACTCGAGGCGCCAGTCTCCCCATAAGGAGGTTAACATACATGATCAAACATTTT
FWD2	ACATACATGATCAAACATTTTCACTTCAACAAACTGTCAAGCGGTAAGAAGAATAATGTT
FWD3	GCGGTAAGAAGAATAATGTTCCGAGCCCAGCAAAGGGAGTGATTCAGATTAAGAAGAGCG
FWD4	GATTCAGATTAAGAAGAGCGCCTCGCAATTAACGAAGGGC
REV	AGCTAATTAAGCTTTTACCCATAGAAGGAGATTGGGGT
Next-generation Sequencing	
TSUA1	AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCT
BarcodeR1	ACACGACGCTCTTCCGATCTCTAGGCGCCTCGCAATTAACGAAG
BarcodeR2	ACACGACGCTCTTCCGATCTCATAGCGCCTCGCAATTAACGAAG
BarcodeR3	ACACGACGCTCTTCCGATCTATATGCGCCTCGCAATTAACGAAG
BarcodeR4	ACACGACGCTCTTCCGATCTTCTAGCGCCTCGCAATTAACGAAG
BarcodeR5	ACACGACGCTCTTCCGATCTTACGGCGCCTCGCAATTAACGAAG
BarcodeR6	ACACGACGCTCTTCCGATCTGCGTGCGCCTCGCAATTAACGAAG
BarcodeR7	ACACGACGCTCTTCCGATCTTGGCGCGCCTCGCAATTAACGAAG
BarcodeR8	ACACGACGCTCTTCCGATCTGACTGCGCCTCGCAATTAACGAAG
BarcodeC1	AGACGTGTGCTCTTCCGATCACCGCTCAGCTAATTAAGCTTTTA
BarcodeC2	AGACGTGTGCTCTTCCGATCTGGTCTCAGCTAATTAAGCTTTTA
BarcodeC3	AGACGTGTGCTCTTCCGATCTCTACTCAGCTAATTAAGCTTTTA
BarcodeC4	AGACGTGTGCTCTTCCGATCCAGTCTCAGCTAATTAAGCTTTTA
BarcodeC5	AGACGTGTGCTCTTCCGATCCTACCTCAGCTAATTAAGCTTTTA
BarcodeC6	AGACGTGTGCTCTTCCGATCGGCACTCAGCTAATTAAGCTTTTA
BarcodeC7	AGACGTGTGCTCTTCCGATCAGTGCTCAGCTAATTAAGCTTTTA
BarcodeC8	AGACGTGTGCTCTTCCGATCCAACCTCAGCTAATTAAGCTTTTA

BarcodeC9	AGACGTGTGCTCTTCCGATCCTGCCTCAGCTAATTAAGCTTTTA
BarcodeC10	AGACGTGTGCTCTTCCGATCGAGTCTCAGCTAATTAAGCTTTTA
BarcodeC11	AGACGTGTGCTCTTCCGATCCGTGCTCAGCTAATTAAGCTTTTA
BarcodeC12	AGACGTGTGCTCTTCCGATCTTAACTCAGCTAATTAAGCTTTTA
TSI3'	CAAGCAGAAGACGGCATACGAGATCGTGATGTGACTGGAGTTCA GACGTGTGCTCTTCCG

Table 5. Codon usage and multi-mutant library activity.

		Codon		Codon Usage[90],[a]		Activity vs SE	
Position	Amino Acid	Singletons	Multi-mutant	Singletons	Multi-mutant	Singletons	Multi-mutant
2	T	ACG	ACC	8.5	14.3	0	1
2	S	TCT	AGC	21.2	15.9	0.8	0
2	A	GCG	GCC	54.9	14.1	0.6	1
2	G	GGT	GGC	14.3	14.3	1	1
3	M	ATG	ATG	8.5	8	0.6	1
3	N	AAT	AAC	32.6	19.6	0.5	0
3	T	ACG	ACC	14.6	25.2	0.7	1
3	D	GAT	GAC	32.6	25.5	0	1
3	A	GCT	GCC	24.4	33.1	1	1
6	L	CTG	CTC	37.4	37.4	0.3	0
6	V	GTC	GTC	14.6	25.2	1	0
6	F	TTT	TTC	30	15.1	0.5	1
11	M	ATG	ATG	13.8	25.5	0.4	1
11	V	GTG	GTG	28.9	18.8	1	1
13	Y	TAT	TAC	37.4	37.4	0.5	1
13	S	TCT	TCC	33.9	33.9	0.5	0
13	T	ACG	ACA	18.6	8.5	0.5	1
13	I	ATC	ATC	14.6	6.1	1	1
14	A	GCG	GCA	37.4	37.4	0.7	0
14	G	GGG	GGA	14.3	8.2	1	1

[a]Codon usage estimated for *Escherichia coli*

It is found that only 50% of inactive singletons from the multi-mutants library can be explained by codon usage differences. The p-values for the grid of comparisons do not support codon usage predicting activities of singletons in the multi-mutant library ($p > 0.1$).

Chapter 3 Validation and stabilization of a prophage lysin of *Clostridium perfringens* by yeast surface display and co-evolutionary models

Adapted from "Ritter, S.C., Hackel B.J., '8 Validation and stabilization of a prophage lysin of *Clostridium perfringens* by yeast surface display and co-evolutionary models.' *Applied and Environmental Microbiology* 2019, 85(10): e00054-19.

Permission to reuse all figures and text contained in this chapter has been granted by the American Society for Microbiology.

3.1 Synopsis

Bacteriophage lysins are compelling antimicrobial proteins whose biotechnological utility and evolvability would be aided by elevated stability. Lysin catalytic domains, which evolved as modular entities distinct from cell wall binding domains, can be classified into one of several families with highly conserved structure and function, many of which contain thousands of annotated homologous sequences. Motivated by the quality of this evolutionary data, the performance of generative protein models incorporating co-evolutionary information was analyzed to predict the stability of variants in a collection of 9,749 multi-mutants across 10 libraries diversified at different regions of a putative lysin from a prophage region of a *Clostridium perfringens* genome. Protein stability was assessed via a yeast surface display assay with accompanying high-throughput sequencing. Statistical fitness of mutant sequences, derived from second-order Potts models inferred with different levels of sequence homolog information, was predictive of experimental stability with AUCs ranging from 0.78 to 0.85. To extract an experimentally derived model of stability, a logistic model with site-wise score contributions was regressed on the collection of multi-mutants. This achieved a cross-validated classification performance of 0.95. Using this experimentally derived model, 5 designs incorporating 5 or 6 mutations from multiple libraries were constructed. All designs retained enzymatic activity with 4 of 5 increasing melting temperature, with the highest performing design achieving an improvement of +4 °C.

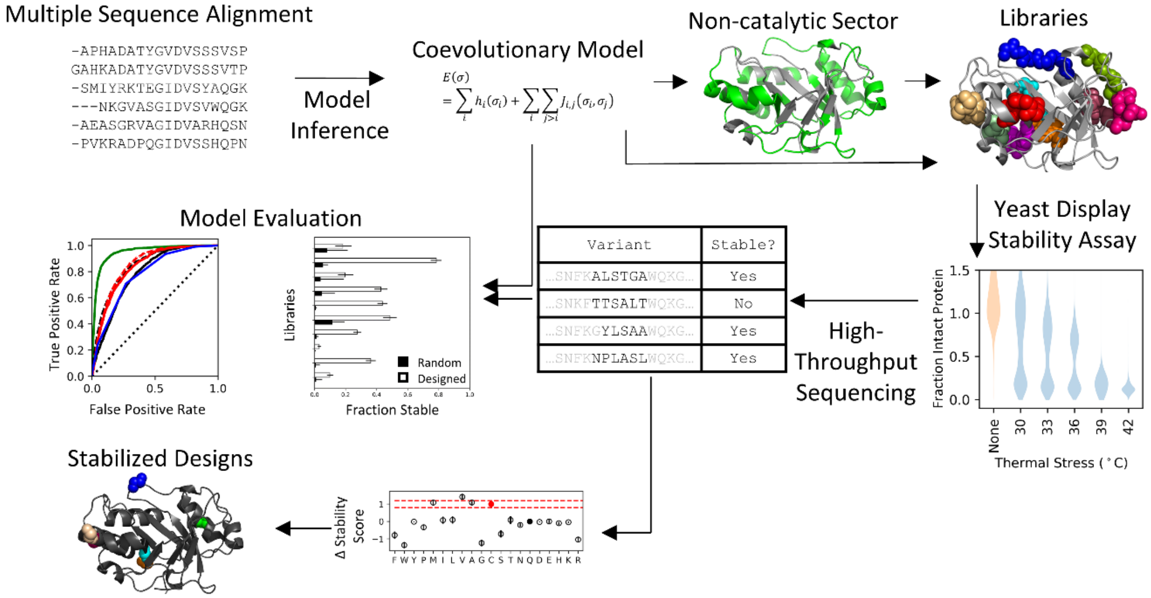
3.2 Introduction

Increased prevalence of antibiotic-resistant bacteria[5] is necessitating the development of new platforms for the discovery and optimization of antimicrobials. Lysins, enzymes that degrade peptidoglycan within the cell wall of bacteria[98], possess the potential to be one such platform. Important for bacteriophage lytic cycle as well as bacterial host cell wall remodeling, lysins have evolved to have high host specificity.[99] Most gram-positive lysins are multi-domain proteins possessing a domain which binds specifically to the cell wall of target bacteria, and a catalytic domain which hydrolyzes bonds in the peptidoglycan backbone.[100] Lysins, however, generally only have mild stabilities and there has been considerable effort to improve the stabilities of several lysins, traditionally by utilizing domain swapping[101,102]. Mutagenesis studies – either random[103] or rational when crystal structures are available to be used in conjunction with protein design frameworks such as Rosetta and FoldX[104] – have had mild success. Lysin engineering efforts have been extensively reviewed.[105] High-throughput engineering strategies employed for lysins include microfluidic encapsulation[106] and microtiter plates[103].

Beneficial mutations have been preferentially identified based on their presence in protein homologs, both natural[107,108] and synthetic[109,110]. The hybrid of homolog-guided design and combinatorial library screening – to design efficient sets of mutants with a high frequency of beneficial mutants – has proven effective.[111–113] Though the cell wall binding domains of lysins tend to be unique, the catalytic domains have high structural and sequence homology[114] with other members of the same catalytic families. Some catalytic domain families, due to the presence of prophage genomic elements as well as autolysins, have tens of thousands of homologous family members. The concordance of these two facts implies that homology-guided generative models incorporating higher-order interactions are promising tools for guiding design of mutagenic lysin libraries.

The promise of bacteriophage lysins as antimicrobials has been hindered by environmental as well as growth-phase-dependent cell wall modifications, leading to poor *in vivo* performance in many instances despite *in vitro* success.[28] Improving lysin catalytic performance through mutagenesis may be greatly enhanced by first stabilizing the catalytic domain, as seen in other enzymes.[115,116] This is because most mutations will be destabilizing[117], and improving the starting stability increases the fraction of folded variants, thereby increasing the available sequence space that can be explored for performance optimization.

Yeast surface display[118,119] has been utilized to engineer a wide variety of protein properties including thermostability[120], binding affinity[38,118,121], and enzyme activity[48,122]. This is accomplished, but also limited, by the nature of the assay, which tethers a protein of interest via a flexible linker to a displayed protein on the yeast surface. Herein, yeast surface display was used as a high-throughput protein stability assay to evaluate use of homology-guided protein generative models for the design of mutagenic libraries of the catalytic domain of a putative lysin[123] from a prophage region of the genome of *Clostridium perfringens* ATCC 13124. Further, a first-order Potts model was regressed over the stability information of the multi-mutants in order to extract single-mutant stability contributions. From these data, promising mutations from multiple libraries were combined into designs not seen individually in the assay to generate a small collection of designed proteins (Figure 3.1). Of the five designs tested with between five and six mutations: all degrade *C. perfringens* cell walls; four have a higher melting temperature than wild-type, with the largest increase being approximately 4 °C; and three retain more activity after incubation at 42 °C than wild-type.



Variant	Stable?
...SNFKALSTGAWQKG...	Yes
...SNKFTTSALTWQKG...	No
...SNFKGYLSAAWQKG...	Yes
...SNFKNPLASLWQKG...	Yes

Figure 3.1. Experimental and data analysis workflow.

3.3 Results

3.3.1 LysCP2 degrades C. perfringens cell walls and exhibits modest stability

LysCP2 was previously identified as a putative lysin in the genome of *C. perfringens* ATCC 13124[123]. Of the list of putative lysins, LysCP2 was selected for the current study for the following reasons: it belonged to the largest catalytic family (glycoside hydrolase 25, GH25) of identified lysin genes; it possessed an SH3 domain, common among almost all putative lysins identified in the study; and it originated from a prophage domain of the genome, indicating it is most likely of bacteriophage origin. To aid in visualization and identification of catalytic-domain residues the automated pipeline of SWISS-MODEL[124] was used to generate a homology model of LysCP2. This process identified an x-ray structure of the endolysin of *C. perfringens* phage phiSM101[125] as a template with 40% identity with LysCP2. The resulting template had a QMEAN score of -3.14, with the highest confidence within the catalytic domain. This structure comprises two functional domains, an N-terminal GH25 domain and a C-terminal SH3 cell wall binding domain (Figure 3.2.A and B). None of the cysteines, either in the catalytic domain or the cell-wall-binding-domain, are oriented such that one would expect them to form disulfides, as predicted in the homology model. *E. coli* effectively produced LysCP2 in the soluble lysate fraction with a genetically appended GSHHHHHH epitope at the C-terminus to facilitate purification by cobalt affinity chromatography (Figure 3.2.C).

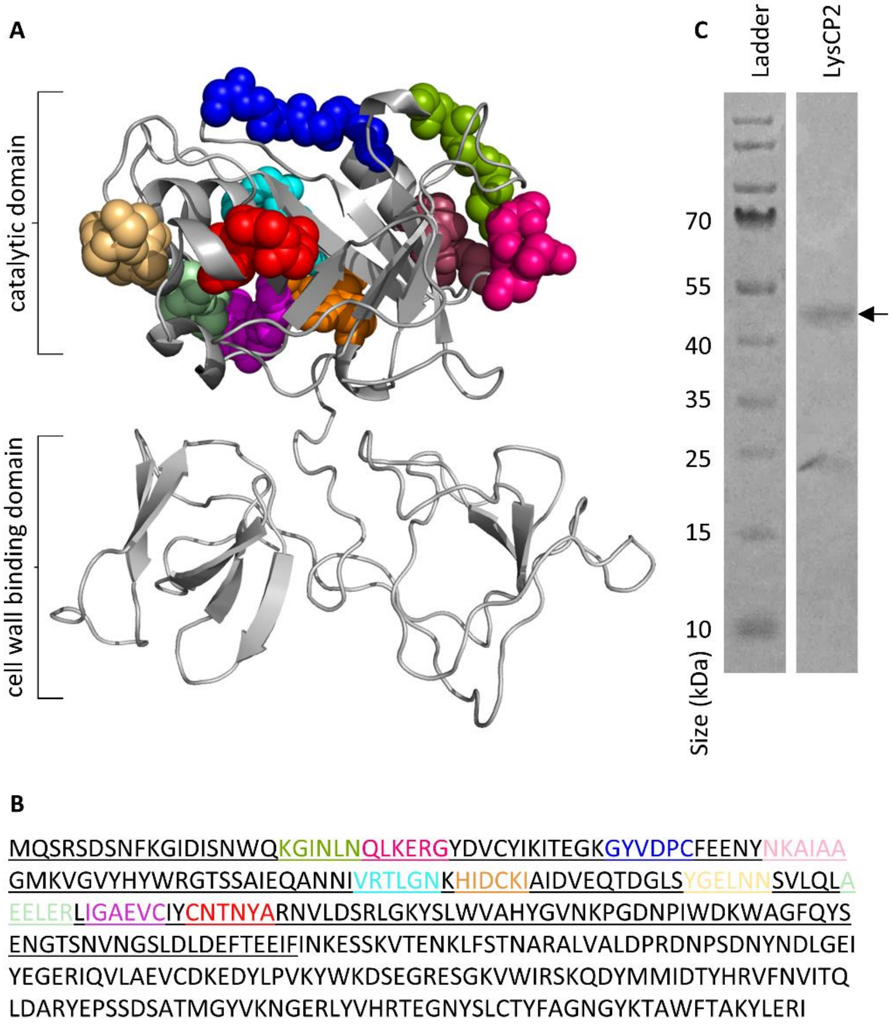

A
catalytic domain
cell wall binding domain
C
Ladder
LysCP2
70
55
40
35
25
15
10
Size (kDa)
B
MQSRSDSNFKGIDISNWQKGINLNQLKERGYDVCYIKITEGKGYVDPCFEENYNKAIAA
GMKVGVYHYWRGTSSAIEQANNIVRTLGNKHIDCKIAIDVEQTDGLSYGELNNSVLQLA
EELERLIGAEVCIYCNTNYARNVLDSRLGKYSLWVAHYGVNKPGDNPIWDKWAGFQYS
ENGTSNVNGSLDLDEFTEEIFINKESSKVTENKLFSTNARALVALDPRDNPSDNYNDLGEI
YEGERIQVLAEVCDKEDYLPVKYWKDSEGRESGKVWIRSKQDYMMIDTYHRVFNVITQ
LDARYEPSSDSATMGYVKNGERLYVHRTEGNYSLCTYFAGNGYKTAWFTAKYLERI

Figure 3.2. Sequence and purification of LysCP2. (A) Homology model of LysCP2, with the locations of designed mutagenic libraries highlighted as colored spheres. (B) Primary sequence of LysCP2 with underlined catalytic domain and mutagenic libraries with colors the same as in the structure. (C) SDS-PAGE analysis of purified LysCP2 resulting from the soluble fraction of *E. coli* lysate purified by Co-NTA chromatography. Expected molecular weight: 41.1 kDa.

The activity of purified LysCP2 was assessed against crude cell wall extract of *C. perfringens* ATCC 12916. 200 nM LysCP2 effectively degrades *C. perfringens* cell wall (Figure 3.3.A) consistent with its hypothesized function. The thermal stability of LysCP2 was determined by Sypro Orange thermal assay, which indicates a modest T_m of 38.3 ± 0.7 °C (Figure 3.3.B). To evaluate the influence of the catalytic domain on stability, the catalytic domain with addition of the flexible linker region (INKESSKVT) was expressed in isolation and evaluated for thermal stability (Figure 3.3.B). Though the T_m cannot be reliably measured, it can be estimated to be <30 °C from the data. Thus, LysCP2 exhibits the expected cell wall degradation function though it has modest thermal stability, including within its catalytic domain, which motivates evolution of elevated stability.

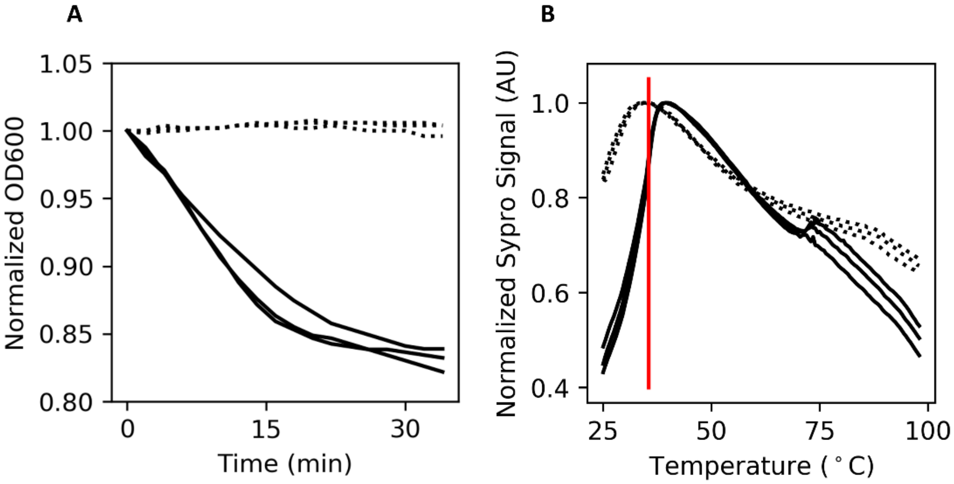


Figure 3.3. Purified LysCP2 degrades C. perfringens cell walls with modest stability. (A) *C. perfringens* ATCC 12916 cell walls were exposed to either 0 nM (dotted) or 200 nM (solid) LysCP2. Cell wall integrity was monitored via optical density at 600 nm. Three replicates of each condition are shown. (B) Full LysCP2 (solid) or the catalytic domain of LysCP2 (dotted) were heated from 25 °C to 98 °C, and Sypro Orange signal was monitored. The T_m of LysCP2 is indicated by a vertical red line.

3.3.2 Proteinase K degradation of yeast-displayed LysCP2 demonstrates thermal transition consistent with stability and inactivation properties

Yeast surface display can be used for high-throughput evaluation of protein stability based on resistance to proteinase K digestion under thermal stress (Figure 3.4.A)[126]. Yeast-displayed LysCP2 demonstrates resistance to proteinase K cutting until unfolding at 40 °C (Figure 3.4.B). This transition is consistent with stability as determined by purified protein Sypro Orange assay (Figure 3.3.B). This transition provided evidence that the proteinase K assay could be used to evaluate the stabilities of LysCP2 mutants in the display context (Figure 3.4.C; details follow).

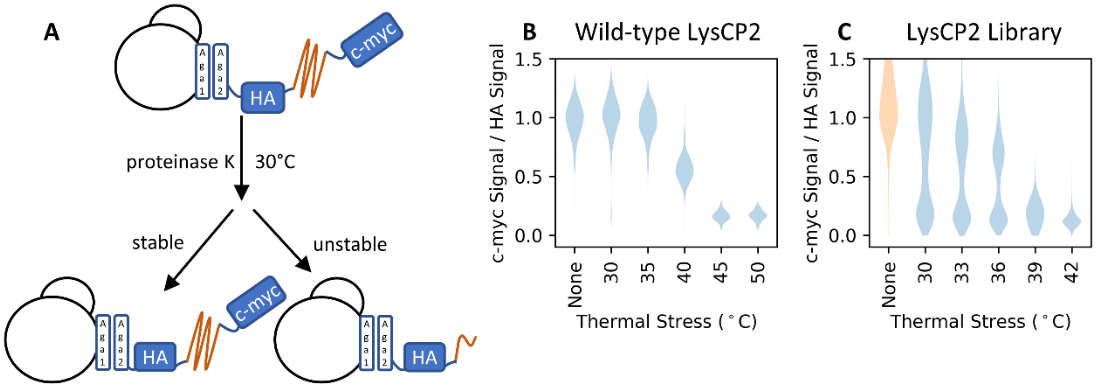


Figure 3.4. Proteinase K degradation of yeast-displayed LysCP2 with different thermal stresses. (A) Proteinase K preferentially cleaves unfolded displayed proteins. The fraction of protein cleaved is analyzed by dual-color flow cytometry labeling the N-terminal HA and C-terminal c-myc epitopes. (B, C) Clonal LysCP2 (B) or the full-length mutagenic library (C) was displayed, incubated in proteinase K at the indicated temperature (*None*: without proteinase K), and evaluated by flow cytometry. Note: *None* condition for (C) was assessed on an instrument different than accompanying thermal conditions; its signals have been linearly scaled so that the mode of the distribution is 1.0. The clonal sample exhibits unimodal distribution whereas the library is bimodal with a population that exhibits proteolytic cutting even at 30 °C. Note the difference in temperatures between B and C.

In addition to the goal of engineering enhanced stability in LysCP2, the capacity of co-evolutionary statistical models to aid in this process was assessed. In particular, a second-order Potts model was selected as the generative model of interest. The parameters of this model, inferred from thousands of homologous sequences (Table 6), decompose a protein sequence into a set of site-wise (h_i) and pairwise ($J_{i,j}$) contributions which sum to give a statistical fitness ($E(\sigma)$).

Equation 2

$$E(\sigma) = \sum_i h_i(\sigma_i) + \sum_i \sum_{j>i} J_{i,j}(\sigma_i, \sigma_j)$$

Given the importance of lysin folding to its function, it was hypothesized that the statistical fitness of lysins would correlate with their stabilities, as has been observed for other protein families[127].

Table 6. Generative model characteristics.

				Number of Sequences	
Model	Sequence Set	Positions	Coupling Regularization	Total	Effective[a]
Firm_L1	Firmicutes	197	L_1	10174	3607
Firm_L2	Firmicutes	197	L_2	10174	3607
GH25_L1	Glycoside hydrolase 25	169	L_1	25738	5523
GH25_L2	Glycoside hydrolase 25	169	L_2	25738	5523

a. Effective number of sequences is the sum of weights of sequences where the weight for each sequence is the inverse of the number of sequences with percent identity greater than 80%, including the sequence of interest.

The theoretical relevance of the model chosen can be derived from random energy machines[127], with the energy of a state (sequence) being equal to its statistical fitness and the distribution of protein sequences following a Boltzmann distribution.

Equation 3

$$P(\sigma) = \frac{e^{E(\sigma)}}{Z}$$

The model parameters were inferred – via maximization of a pseudolikelihood approximation of the intractable partition coefficient Z and previously described regularization coefficients[128] – from a multiple sequence alignment generated with homologous sequences of the catalytic domain of LysCP2. These sequences were identified with an iterative homology search in JackHmmer[129] using the catalytic domain of LysCP2 as the seed sequence and homologous sequences restricted to those annotated from the Firmicute phylum in the UniprotKB database[130]. Homology was restricted to this selection because there has been evidence that some lysin catalytic domains maintain similar specificity as their full-molecule counterparts[131–133], implying that some amount of specificity can be encoded for in the catalytic domain alone. These data provided the observations for inference of a Potts model using PLMC[128] and L_1 regularization. With the fitness model in place, combinatorial libraries were designed to both identify stabilized LysCP2 variants and evaluate the efficacy of the generative model. Each library diversified six contiguous residues restricted to sites without significant evolutionary couplings to the primarily connected catalytic residues (designated PLMC). Residues were determined to be evolutionarily coupled if their coupling scores[128] were greater than 0.2 (Figure 3.11.A). Residues that were connected to putative catalytic residues (Y66, D98, E100, Y156, Q174)[125] were defined to belong to the catalytic network. Thus these, and any sites evolutionarily coupled to these sites, were excluded from library diversification as we aimed to study mutations primarily impacting stability. The design algorithm (Materials and Methods) selects amino acid degeneracy assuming additive benefits only from site-wise mutations, prioritizing inclusion of the highest-performing variants and, when possible, selecting degenerate codons that avoid lowest-performing mutations. From this collection of possible libraries, ten were selected (as detailed in Section 3.7.1 and Figure 3.11). The designed diversity of the final libraries, and comparisons with the experimentally observed library diversity, are presented in Figure 3.12. At the positions of all designed libraries (Figure 3.5), a second library was also tested with full degeneracy (NNK codons). The use of designed and random libraries assesses the utility of the generative model and samples of a broad range of statistical fitnesses.

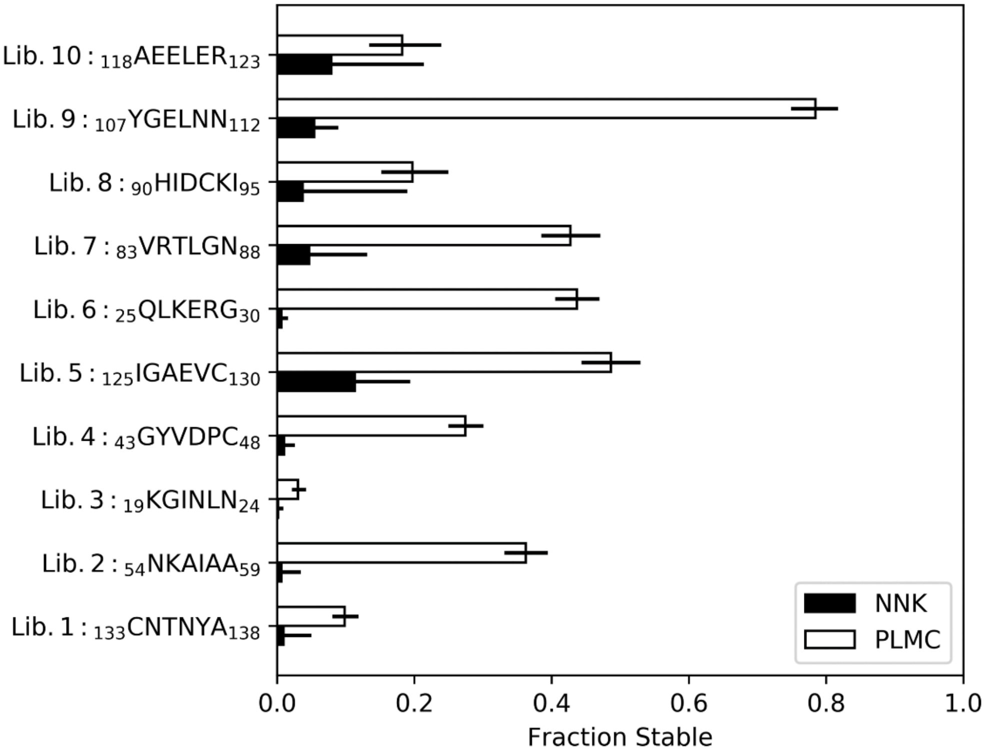


Figure 3.5. Fraction of stable variants observed across all libraries for different library design strategies. (Vertical) Each library number is followed by the indices of the first residue number and the identity of the six wild-type residues. Error bars presented are 95% confidence intervals computed with Clopper-Pearson interval.

Genes were synthesized via polymerase-chain reaction with degenerate oligonucleotides (for full procedure see Methods) and transferred into a yeast surface display system as fusions to the C-terminus of Aga2p with a PAS40-(Gly$_4$Ser)$_3$ linker[134]. 1.2 million yeast transformants were obtained. A collection of variants with full-length gene products was isolated by sorting 200 thousand yeast displaying C-terminal c-myc epitope signal. The collection of mutant libraries was then subjected to protease stability screening. A substantial fraction of the population exhibits reduced stability as evidenced by protease susceptibility at low temperature (30 °C; Figure 3.4.C). The bimodality of this population is not observed in the absence of protease treatment, providing evidence that even at low thermal stress many of the displayed mutants are unstable. Additional variants exhibit protease susceptibility in the mid-30s °C. Yet, similar to wild-type, a substantial fraction of the mutagenic library of LysCP2 does not display a transition until more elevated thermal treatment. Notably, very few variants exhibit stability superior to wild-type.

LysCP2 variant genes from stable and unstable populations at 30 °C (isolated by flow cytometry of yeast with high or low, respectively, c-myc/HA signals) were extracted and analyzed via Illumina MiSeq sequencing. This process generated 1.5 million sequences for analysis. Following preprocessing to remove noise and sequences that did not match library designs, sequences were classified as stable or unstable if at least five more sequence counts were observed in either the stable or unstable populations. This cutoff served to remove sequences with intermediate stability as well as those that were noisy observations due to erroneous sequencing. Increasing this threshold progressively removes sequences while improving performance as assessed by AUC, until too many sequences are removed, which degrades model quality (Figure 3.13). As a result, 9,749 multi-mutants were classified as stable or unstable, and summary statistics computed for each (Figure 3.5, Table 8). For 8 of 10 libraries, the designed library yielded a greater proportion of stable variants than the library of randomized sequences (Figure 3.5). The stable fractions of the remaining designed libraries, 8 and 10, were not statistically different from their random counterparts. Given the nature of the classification, these summary statistics do not directly represent the magnitude of stabilities of variants within the libraries.

3.3.3 High-throughput analysis of variant stabilities

Using the extensive set of stable and unstable sequences observed from the assay, model predictive performance was evaluated for different homolog datasets and regularization methods. Potts models inferred using PLMC on homologous sequences were used to compute the difference in statistical fitness relative to wild-type LysCP2 (Δ statistical fitness).

In addition to these homology-driven methods, a first-order Potts model was inferred using logistic regression using the collection of stable and unstable sequences from the yeast display stability assay. A logistic regression was selected as it is analytically equivalent to two-state thermal equilibrium and experiences less bias than for naïve Bayes[135]. To assess this experimental model's predictive capability on sequences outside of its training set, 90% of sequences were used to train the model, and the stabilities of the remaining 10% were predicted. This process was then restarted and repeated to predict the stability of all sequences (10-fold cross-validation). Further, bootstrap regression enabled error estimation for each library to elucidate mutations with statistically significant effect.

The predictive performance of each model was computed as the area under the curve (AUC) from a receiver operating characteristic (ROC) curve with weighting applied to equalize classes (stable and unstable) as well as library representation, so that each library had equal representation regardless of the number of sequences (Figure 3.6). All models were substantially superior to random. The weakest performance (AUC = 0.78) was observed for the model derived from the firmicute homologs with L_1 regularization. Notably, a model that merely considered the Hamming distance between mutant and wild-type yielded an equivalent AUC. Expansion of the homolog sequences to the entire glycoside hydrolase 25 family, maintaining L_1 regularization, elevated the AUC to 0.83. The use of L_2 regularization further improved performance to 0.85 for both homolog sets. The logistic regression from experimental data achieves the highest performance of any strategy (AUC = 0.95).

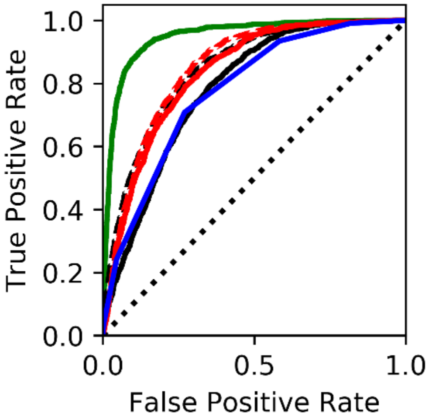


Figure 3.6. Performance of homology and experimentally driven models to predict stability of LysCP2 variants with up to six mutations. Receiver operating characteristic for: Hamming distance to LysCP2 (blue), or with generative models Firm_L1 (black-solid), Firm_L2 (black-dashed), GH25_L1 (red-solid), GH25_L2 (red-dashed), or with cross-validated experimental model (green).

There is a positive and significant correlation between site-wise Δ statistical fitness computed with the GH25_L2 model and the experimental logistic model across all mutated positions (Figure 3.7.A-C, slope 95% confidence interval: 0.085 ± 0.019 Δ stability score / Δ statistical fitness). The non-unity of the slope is not unexpected given that these values have not been transformed into common units. Under the conditions of the assay, a variant remaining 50% uncleaved would see that rise to 60% after the addition of a beneficial mutation with Δ statistical fitness of +5.0.

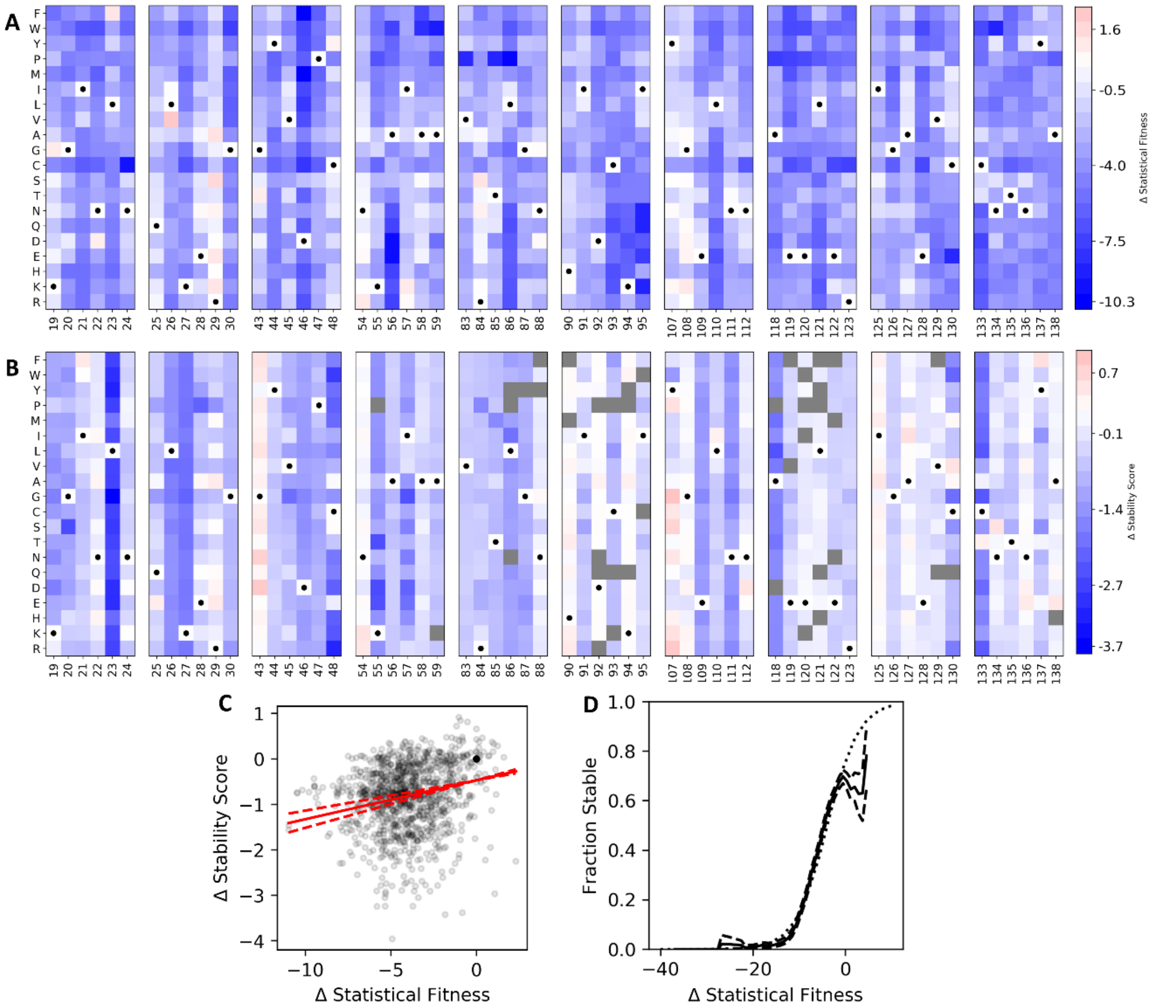


Figure 3.7. Connection between statistical fitness and stability. (A) Δ Statistical fitness values of all possible single mutations (rows), relative to wild-type residues (dotted squares), across each library position (columns). (B) Δ Stability score derived separately for each library at indicated positions for indicated single-mutants. (C) Δ Statistical fitness vs Δ Stability score exhibits significant correlation. Red lines show best fit as well as upper and lower bounds of 95% confidence interval (slope = 0.085 ± 0.019) with constant intercept. (D) The fraction of mutants that are stable at indicated levels of Δ statistical fitness with median (solid), 95% confidence interval (dashed), and a fitted two-state model (dotted).

When approximating the stable fraction of sampled protein variants as a function of Δ statistical fitness, a transition can be seen (Figure 3.7.D). The close agreement between the estimation and a fitted two-state model is consistent with models whereby the fitness penalty is proportional to the unfolded fraction of molecules of a protein.[127]

Designed sub-libraries exhibit a more equal representation of both stable and unstable sequences, in contrast to the completely random library (Figure 3.5). These sub-libraries were regressed the same as the larger libraries (cross-validated AUC between 0.9 and 0.97 for all sub-libraries) and stability scores computed and compared with statistical fitnesses (Figure 3.8.A). Focusing on this subset of sequences, libraries 1, 2, 7, 8, and 10 exhibit strong correlation between Δ stability score and Δ statistical fitness. Of these libraries: 1, 2, 7, and 10 are predicted to have dominant alpha-helical character; 8 is predicted to be buried and have low secondary structure. Of the non-correlated libraries: 6 and 9 are on opposite edges of the catalytic domain, and are possibly close to the peptide cross-bridges of the peptidoglycan[125]; 3 and 4 are on the upper face of the catalytic pocket (Figure 3.8.B).

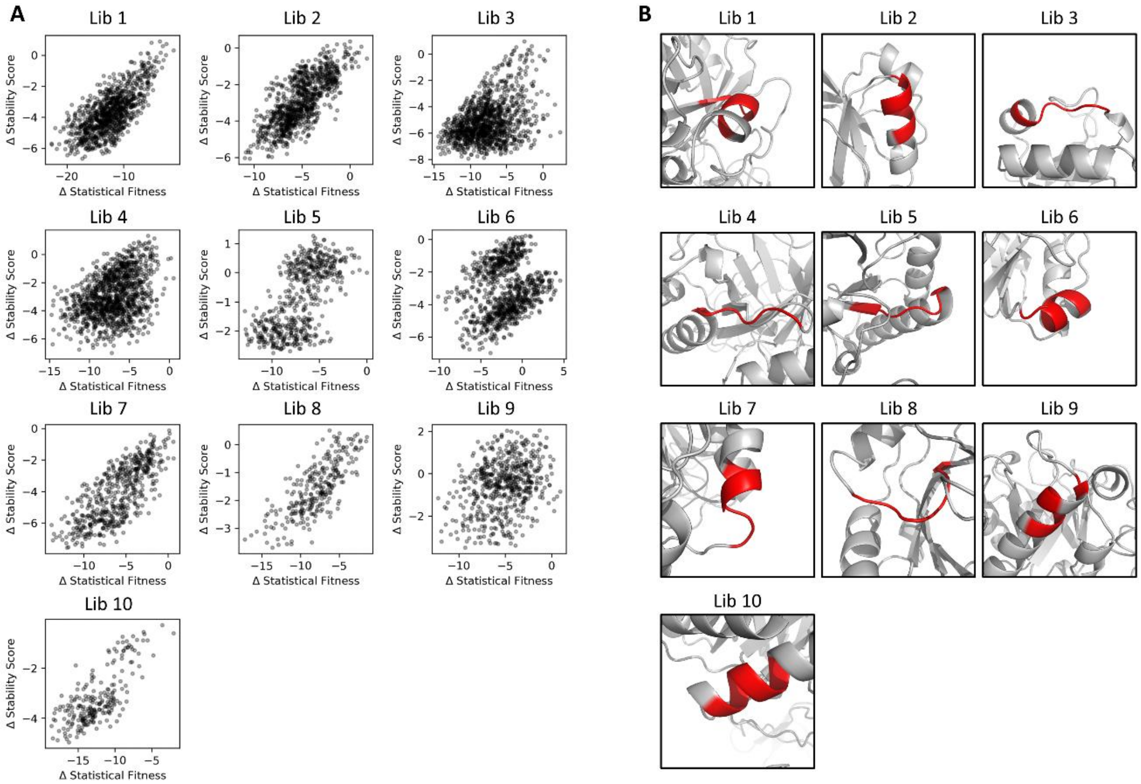


Figure 3.8. Structural insights provide context of where co-evolutionary models are most predictive. (A) Full-sequence Δ statistical fitness and Δ stability score of mutants belonging to libraries designed with Algorithm 1. The homology-model positions are highlighted in red (B).

Examining further, library 6 can be seen to contain two parallel sub-populations (Figure 3.8.A). These two sub-populations are distinguished by a highly inaccurate prediction at position 26. The designed sub-library allows only two mutations, wild-type L and mutant V. Though L26V is highly beneficial by statistical fitness, it is highly detrimental by stability score (Figure 3.7.A and B) thereby creating the bimodality. Library 5 displays two populations which are independently uncorrelated, however together are positively correlated. This appears to be due to C130, which is the only position with strong correlation.

These observations imply that statistical fitness offers the highest confidence in predicting stability changes of the catalytic domain of LysCP2 when considering positions with high secondary structure, low solvent accessible surface area, or are more distal from the active site.

3.3.4 A site-wise model of stability identified stabilizing mutants that maintain activity against C. perfringens cell walls

To estimate the error of model parameters, the coefficients for the experimental model were inferred across 1000 bootstraps to approximate a 90% confidence interval for each parameter. Using this information, mutations with significantly beneficial stability over wild-type were identified (Figure 3.9). From this collection: N54R was selected instead of N54K to preserve possible hydrogen bonding; Y107R and Y107G were selected as representatives of the two most beneficial sets of mutations at positions 107 (positive and small, respectively); G43E was not predicted to be significantly different than wild-type, however, it was added due to its similarity to G43D and G43N for comparison; N112D, A127I, and C130V are the only beneficial mutations at their respective positions. These were combined into a collection of 5 design sequences incorporating between 5 and 6 mutations (Figure 3.10.A and B).

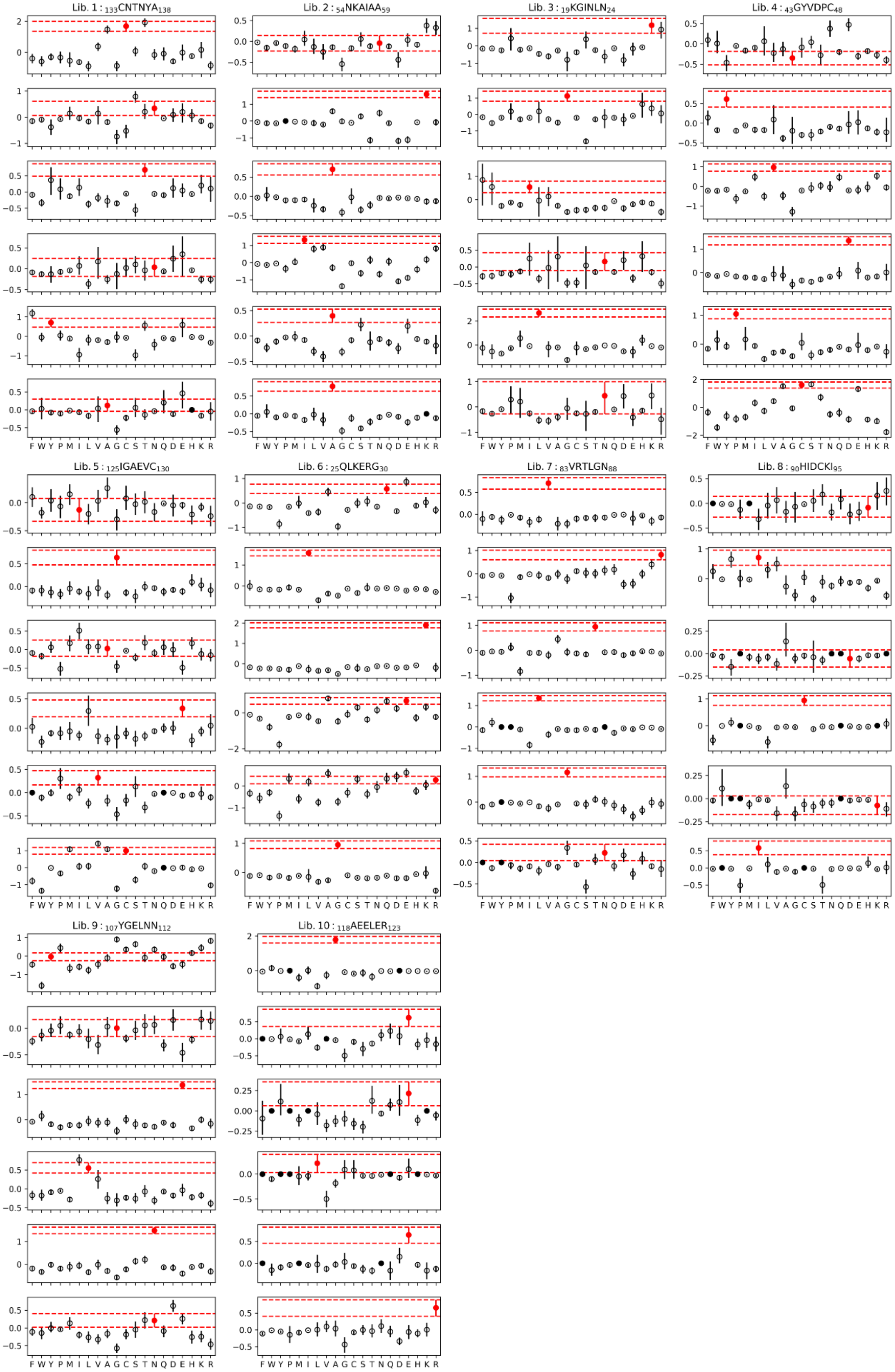
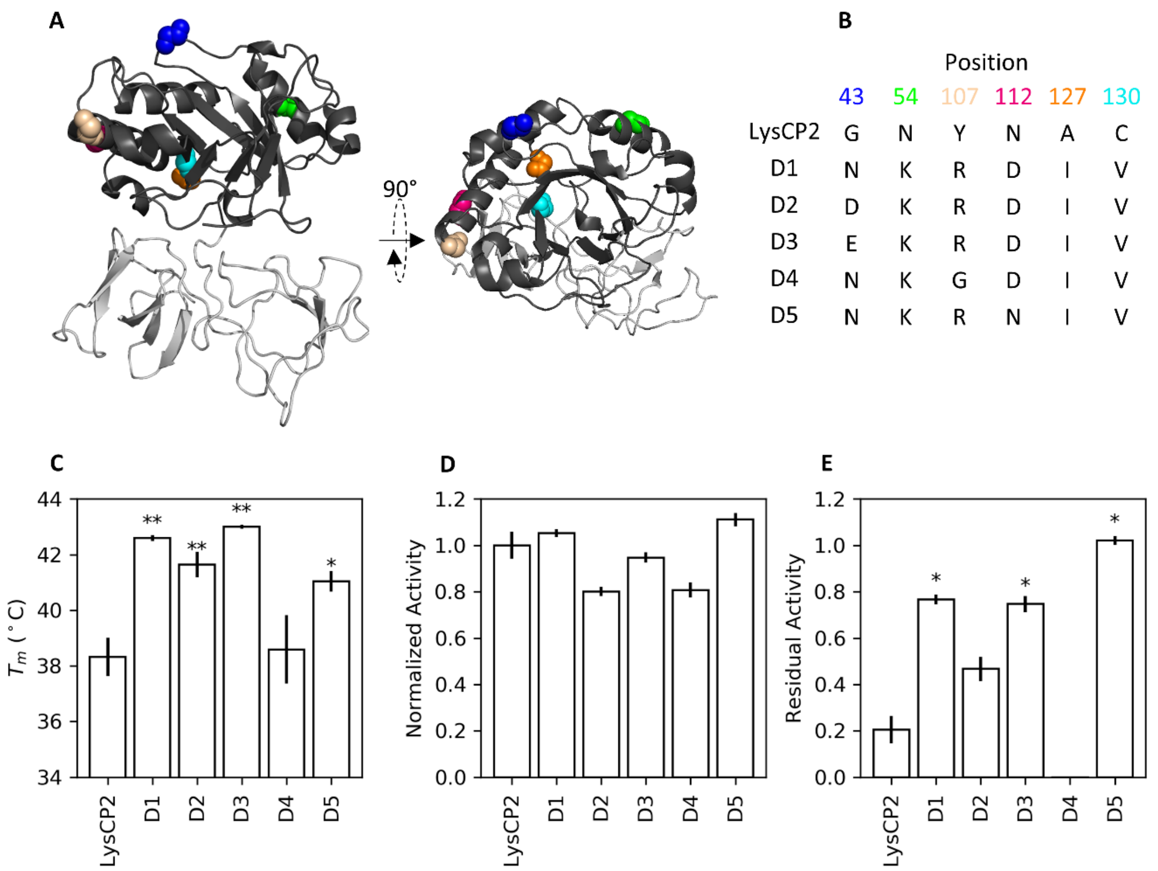

Lib. 1 : 133CNTNYA138
Lib. 2 : 54NKAIAA59
Lib. 3 : 19KGINLN24
Lib. 4 : 43GYVDPC48
Lib. 5 : 125IGAEVC130
Lib. 6 : 25QLKERG30
Lib. 7 : 83VRTLGN88
Lib. 8 : 90HIDCKI95
Lib. 9 : 107YGELNN112
Lib. 10 : 118AEELER123
F W Y P M I L V A G C S T N Q D E H K R

Figure 3.9. Bootstrap error estimates of residue stability scores. One thousand bootstraps drawn from library sequences were used to estimate site-wise stability score contributions for wild-type (red) and mutant (black) amino acids. Error-bars are 90th-percentile confidence intervals. Closed black circles indicate those mutants for which both the mean and error were zero, indicating insufficient information for parameter estimation.

The designs were produced in *E. coli* and evaluated for thermal stability and activity. Of the designs, 4 of the 5 display higher thermal stability than wild-type with improvements by up to 4 °C as assessed by Sypro Orange thermal denaturation assay (Figure 3.10.C). All tested designs degrade crude cell wall preps of *C. perfringens* ATCC 12916 at 37°C in PBS (Figure 3.10.D). The combination of stability and activity was tested by treating *C. perfringens* cell walls with 200 nM lysin after incubation at elevated temperature (42 °C) for 30 minutes. Variants D1, D3, and D5 exhibit 4- to 5-fold greater activity than wild-type following thermal treatment, which only retains 20% activity relative to unheated lysin (Figure 3.10.E). D2 is nominally more active than wild-type whereas D4 loses all activity despite thermal stability similar to wild-type.



	Position					
	43	54	107	112	127	130
LysCP2	G	N	Y	N	A	C
D1	N	K	R	D	I	V
D2	D	K	R	D	I	V
D3	E	K	R	D	I	V
D4	N	K	G	D	I	V
D5	N	K	R	N	I	V

Figure 3.10. Site-wise model of stability used to identify potentially stabilizing mutants yields improved thermal stability while maintaining activity against Clostridium perfringens cell walls. (A) Residues of LysCP2 identified for mutagenesis to improve stability. (B) Composition of different tested constructs. (C) Melting temperature (T_m) for each construct as determined via Sypro Orange thermal denaturation assay. n = 11 (D,E) Normalized or residual activity of constructs at 200 nM on crude purifications of *Clostridium perfringens* strain ATCC 12916 cell walls after 30 minutes at 4 °C (D), or 42 °C (E), respectively. n = 3. For C-E, error-bars are standard errors, statistical tests are two-sided unequal variance t-tests with Bonferroni correction, ** is $p < 0.01$, and * is $p < 0.05$.

3.4 Discussion

Generative models have been shown to predict protein fitness changes [128]. Yet, it is still being studied how statistical fitness, however accurate, can be interpreted to engineer physical properties, such as stability, especially in the context of multiple mutations. A protein’s contribution to organism fitness is a nonlinear function of many attributes including, for example, stability, solubility, and activity. In some cases, these attributes are the result of different spatially-connected regions of a protein.[136] Many proteins are only marginally stable at physiological conditions and destabilizing mutations therefore incur a fitness cost on the organism due to decreased effective concentration as well as increased aggregation and protein degradation stresses. Focusing mutations in regions of LysCP2 that were evolutionarily decoupled from the catalytic residues enabled the study of mutations primarily affecting folding properties including stability. The results shown here for the glycoside hydrolase 25 catalytic domain of a phage lytic enzyme isolated from *Clostridium perfringens* find that a second-order Potts co-evolutionary model is predictive in its classification of variants at these positions as stable or unstable. The near-wild-type activity of all designs (Figure 3.10.D) is consistent with the hypothesis that the selected positions were removed from catalytic activity thereby improving critique of the co-evolutionary model to infer stability changes.

Though libraries designed using a homology model did not display a high proportion of stabilizing effects in comparison with wild-type (Figure 3.3), it was desired to infer potentially stabilizing mutations from the data and combine those to generate an improved molecule. A site-wise global model was fit over the experimental data of each library and obtained high performance in predicting mutational outcome on assay stability (cross-validated ROC AUC = 0.95). Comparing predicted single-mutant statistical finesses of the homology model to stability scores of the experimental model demonstrated a correlation across all mutations. When examining each sub-library separately (Figure 3.8), five libraries (sites 54-59, 83-88, 90-95, 118-123, and 133-138) exhibit strong linear correlation between statistical fitness and stability score.

The disagreements of four libraries could come from several possibilities (while a fifth library's discordance is explained by a single strongly inaccurate prediction (L26V)). These possible inconsistencies arise from systematic biases in the experimental setup or in assumptions made between the connection of statistical fitness and stability. First, the assay itself is a measure of protease stability in the context of yeast surface display and may not be as strongly correlated with soluble form stability as assumed. Second, it is possible that the cell wall binding domain improves overall stability of the molecule, thereby decreasing the fitness burden of a less stable catalytic domain. Indeed, for LysCP2, when produced in isolation the catalytic domain has significantly reduced stability than the full form. Third, the *in vivo* function of phage lysins requires that they pass through holins to degrade the cell wall at the end of the lytic cycle. This process may be optimal for particular surface-displayed residues at the expense of other properties such as stability. Nonetheless, the highly correlated results of some libraries implies that selecting positions that have high secondary structure, are buried, or are distal from the catalytic pocket or possible substrate interactions can improve the predictive performance of statistical fitness.

Not all designs predicted to be more stable by the experimentally derived model were. The Y107G substitution was predicted to be nearly as stabilizing as Y107R, however it resulted in a significant decrease in both melting temperature as well as kinetic stability (Figure 3.10.C and E). This could potentially be due to biases of proteinase K's cutting activity which could result in Y107G being more resistant to protease degradation on average across the mutations presented in that sub-library. In another example, though less pronounced, these types of biases may also be seen in comparisons of G43D and G43E. The experimentally derived model predicted G43E to be less stabilizing than G43D, whereas in the pure form those results appear to be reversed. Protease biases can be mitigated by performing this stability assay with multiple proteases, as seen in the work of Rocklin et al.[126]

Though relatively small, the improvements to stability seen in the final designs which maintain wild-type activity levels support the use of co-evolutionary models to augment library design in the pursuit of stabilizing mutations to lysins. Within the context of methods presented here, experimental outcomes could be further improved in multiple ways: (1) Increased sampling under a wider range of environmental conditions could reduce noise and enable pairwise parameter estimation from experimental data; (2) Creation of second-generation libraries from these experimental results would enable direct evaluation of stabilities of many more designs (i.e. the current results were accomplished with a single iteration of evolution); (3) Selection of library positions based on structural arguments without requirement of continuity in primary sequence. Beyond these, co-evolutionary models could be used to augment existing stabilizing techniques, such as iterative-saturation-mutagenesis.

The designed libraries were limited to a maximum possible genetic diversity of 16,000 to demonstrate the capacity of a designed restrained library to outperform a random library at those same positions (the maximum diversity of the random libraries at each set of positions is approximately 1 billion members). The first arm of the study was meant to provide a large number of multi-mutant observations to assess the utility of the generative protein model, and though it could be expanded, the roughly 10,000 variants whose stabilities could be categorized by the assay was sufficient to explore this question. This work also presented support for the translatability of the protease stability assay, with thermal shifts, on the surface of yeast to a particular lysin. Moving beyond the work presented here, future studies attempting to stabilize lysin catalytic domains will benefit greatly by using co-evolutionary models to augment the design of libraries larger than those presented here (at many more positions) and be scalable to use the full capacity of fluorescent activated cell sorting and yeast display. Of particular note, given the principles applied here, one could choose to titrate the temperature used to assay these libraries and utilize multiple rounds of sorting in order to enrich directly for the rare multi-mutants which provide large benefits in stability.

3.5 Materials and Methods

3.5.1 Bacterial and yeast culture

Escherichia coli were grown in lysogeny broth (LB) in liquid, or solid with 1.5% agar, supplemented with either 100 µg/mL ampicillin or 50 µg/mL kanamycin when noted. All cultures were grown at 37 °C, with liquid cultures shaken at 250 rpm, unless otherwise noted.

Clostridium perfringens strains (generously provided by Dr. Swift in the Donovan Lab at the US Department of Agriculture) were grown on solid BHI (37 g/L brain heart infusion, 1.5% agar), or liquid BYC (BHI with 5 g/L yeast extract and 0.5 g/L L-cysteine). Solid culture plates were prepared fresh for each use and incubated overnight anaerobically using the AnaeroGen Compact system (ThermoFisher). Liquid cultures were sealed immediately after autoclaving to reduce dissolved gases. After equilibration to 37 °C, liquid culture was inoculated with fresh bacterial colonies.

Saccharomyces cerevisiae strain EBY100 was grown non-selectively in liquid YPD (10 g/L yeast extract, 20 g/L bacto peptone, and 20 g/L D-glucose) or solid with 1.5% agar. For selective growth, yeast were grown in SD-CAA (16.8 g/L sodium citrate dihydrate, 3.9

g/L citric acid, 20 g/L D-glucose, 6.7 g/L yeast nitrogen base, 5 g/L casamino acids). For induction, yeast was grown in SG-CAA (10.2 g/L Na_2HPO_4*$7H_2O$, 8.6 g/L NaH_2PO_4*H_2O, 19 g/L D-galactose, 1 g/L D-glucose, 6.7 g/L yeast nitrogen base, 5 g/L casamino acids). All growth was at 30 °C, 250 rpm where applicable for liquid culture.

3.5.2 Generation of LysCP2 expression plasmid

The pET-24 expression plasmid, previously modified to incorporate an C-terminal six-histidine tag (pETh) [39], was digested with NdeI and BamHI-HF (New England Biolabs) and isolated by gel electrophoresis. Digested plasmid was then assembled via Gibson assembly (HiFi, New England Biolabs) with the codon-optimized gBlock (Integrated DNA Technologies) for LysCP2 (WP_003469445), transformed into MC1061 F- *E. coli*. (Lucigen), plated on solid LB supplemented with 50 µg/mL kanamycin, and sequence verified.

3.5.3 Protein production and purification

Assembled plasmids were transformed into NEB T7 Express LysY/Iq (New England Biolabs), plated on solid LB with 50 µg/mL kanamycin, and grown overnight at 37 °C. Fresh colonies were then used to inoculate 3 mL LB with 50 µg/mL kanamycin, and grown at 37 °C, 250 rpm overnight. Culture was then diluted into 100 mL LB, grown to an optical density at 600 nm (OD600) between 0.5 and 0.8, and induced with isopropyl β-D-1-thiogalactopyranoside (IPTG) to a final concentration of 0.5 mM. Culture was then induced for 3 hours, chilled, and pelleted at 6000x*g* at 4 °C in 3 minute increments until all culture pelleted. Cell pellets were then resuspended in 0.6 mL of lysis buffer (137 mM NaCl, 2.7 mM KCl, 8 mM Na_2HPO_4, and 2mM K_2PO_4 (PBS), 5% glycerol, 3.1 g/L CHAPS, 1.7 g/L imidazole). Suspension was then freeze-thawed four times at -80 °C.

To each sample, 0.6 mL of wash buffer (20 mM imidazole in 1x PBS) was then added and mixed by inversion several times. Suspension was then spun at 17,000x*g* for 10 minutes at 4 °C. Following centrifugation, the insoluble fraction was removed by pipetting, and the soluble fraction was sterilized through a 0.22 µm syringe filter (GE Healthcare). Purification was performed with 0.2 mL HisPur Co-NTA spin columns according to manufacturer instructions (Thermo Fisher). Proteins were then desalted into PBS with Zeba desalting columns following manufacturer instructions (Thermo Fisher).

Protein purity was assessed by SDS-PAGE (sodium dodecyl sulfate polyacrylamide, NuPAGE Bis-Tris 4-12%) gel and ranged from 70%-88% for full-length lysins and 91% for catalytic only domain (Figure 3.14).

3.5.4 Sypro Orange thermal denaturation assay

Purified proteins were diluted in PBS to stock concentration of 5 µM, and 45 µL was aliquoted into optically clear PCR tubes. Stock solution of Sypro Orange (Thermo Fisher) was diluted to 200x in PBS, 5 µL of which was added to each tube and gently mixed. Samples were then loaded into a CFX Connect Real-Time PCR Detection System, and temperature was ramped from 25 °C to 98 °C in 0.5 °C increment, with 30 s each step to allow for equilibrium. Sypro Orange signal was monitored via the FRET channel (450-490 nm excitation with 560-580 nm emission). The temperature derivative of these data were determined by smoothing with local second-degree polynomials with window width of 2.5 °C through the use of a Savitzky-Golay filter (*sklearn*). The T_m of a sample is determined as the temperature with maximum derivative.

3.5.5 Clostridium perfringens cell wall extraction and crude purification

C. perfringens was cultured as described to mid-exponential phase in liquid. Liquid culture was then cooled on wet ice for 15 minutes, transferred to sterile 50 mL conicals, and centrifuged at 6000x*g* and 4 °C for 5 minutes. Cell pellet was then resuspended in 1 mL 50 mM Tris-HCl and added drop-wise to 20 mL boiling 5% SDS. Cells were then boiled for 15 minutes, cooled to room temperature, and centrifuged at 10,000x*g* for 10 minutes. Pellet was then washed twice with 1 mL 1 M NaCl, then 7 times with 1 mL deionized water, and then 2 times with 1 mL PBS, all washes pelleted at 17,000x*g* for 10 minutes. Crude cell walls were then stored at 4°C in PBS until use.

3.5.6 Cell wall degradation assay

Crude cell wall extract was diluted in PBS to OD600 = 1.0. In wells of a 384-well plate, 40 µL crude cell wall was added to 10 µL 1 µM test proteins or buffer. OD600 was then monitored via a spectrophotometer at 37 °C, collecting data at 2 minute increments. Activity was calculated as the slope of the linear region of OD600 vs. time. For thermal inactivation assays, proteins were first incubated at 42 °C for 30 minutes followed by cooling on wet ice before use in assay as described above.

3.5.7 Homology-guided generative model of LysCP2

The statistical fitness of lysin variants was predicted using a Potts model:

Equation 2

$$E(\sigma) = \sum_{i} h_i(\sigma_i) + \sum_{i} \sum_{j>i} J_{i,j}(\sigma_i, \sigma_j)$$

Equation 3

$$P(\sigma) = \frac{e^{E(\sigma)}}{Z}$$

where E(σ), the statistical fitness of a sequence σ, is the sum of site-wise compositional contributions, h_i, and pairwise contributions, $J_{i,j}$. The probability of observing a sequence in a dataset given a set of random energy machines sampling space, $P(\sigma)$, is a Boltzmann distribution. The parameters of this model can be inferred in a number of different ways, each approximating the intractable partition coefficient, *Z.* Inference was performed using PLMC (https://github.com/debbiemarkslab/plmc) with recommended regularization coefficients as described previously.[128] Those parameters were a neighborhood cutoff of 80% identity for each sequence, sitewise regularization coefficient of λ_h=0.01, and pairwise regularization coefficients of λ_J=39.2 and λ_J=33.6 for the Firmicute and Glycoside Hydrolase 25 family models, respectively.

Sequence datasets were gathered from two sources:

(1) The catalytic domain of LysCP2 was used as the seed sequence for an iterative JackHmmer search[129]. The UniprotKB database[130] was used with taxonomic restriction set to Firmicutes. Five iterations were performed with hit threshold set to an E-value of $1.0\text{x}10^{-20}$. Sequences were then aligned with PROMALS3D[137] with default settings.(2) The Pfam[138] family alignment of the catalytic domain of LysCP2 (glycoside hydrolase 25, PF01183), utilizing the NCBI database source.

3.5.8 Design of libraries using statistical fitness

The algorithm for library design according to statistical fitness (fully detailed in the Supplement) adds library diversity up to a fixed limit based upon single-mutation Δ statistical fitness information from parental sequence. The user identifies: (1) maximum library size L_{max} based on experimental synthesis and screening limits; (2) residues to be mutated $\{r_i\}$; (3) the parental sequence σ; (4) a scoring function to compute the fitness of a given sequence $f(\sigma)$; and (5) a codon table c. The script then sets the initial diversity of each codon to null and expands the library iteratively.

The fitness f_{ij} of all single-mutation substitutions at each r_i is computed and sorted in descending order. Moving progressively through the list, each mutation is proposed as an addition to the growing library. Given the collection of mutations already desired at that position, the set of degenerate codons which include all previously accepted mutations as well as the proposed additional mutation is generated. From this collection of degenerate codons, one is selected which maximizes the minimum f_{ij} while maintaining the smallest combinatorial genetic size. If the inclusion of this degenerate codon maintains the library size below L_{max}, then it replaces the current degenerate codon at that residue location. The algorithm then progresses with the next-best proposed mutation until L_{max} is reached.

3.5.9 Design of mutant libraries

The set of evolutionary coupling scores of LysCP2's catalytic domain, computed using the Potts parameters inferred from the Firmicute dataset using L_1 regularization, was computed as discussed previously[139]. The set of residues that were statistically coupled to the catalytic triad, as well as the residues connected to that set, were removed from the list of residues. For all primary sequence contiguous sets of residues of length six, the algorithm for library design as outlined above was performed with maximum rational library sizes of 16,000. Two metrics were calculated for each designed library: the maximum single mutation score and the 90th-percentile statistical fitness score as determined by random sampling from the library. A set of ten non-overlapping libraries were selected to maximize point-mutant scores with 90th-percentile scores used for tiebreaking (Supplement Figure 1 and Algorithm 2).

3.5.10 Generation of yeast surface display plasmid for LysCP2 and mutants

Sequence-verified LysCP2 was amplified from the pETh plasmid using primers LYS001 and LYS005 (for all indicated oligonucleotides see Table 7), it was then assembled with pCT40 plasmid[134] previously digested and isolated with NheI-HF and BamHI-HF (New England Biolabs), and transformed into MC1061 F- *E. coli* (Lucigen) and plated on solid LB supplemented with 100 µg/mL ampicillin. The resulting construct encodes for Aga2p—linker with HA epitope—LysCP2—Myc epitope.

Table 7. List of oligonucleotides.

PrimerID	Sequence
Lib.all.FWD	GTGGGGGCGGATCTGCTAGCATGCAGAGTAGAAGCGACAG
Lib.all.REV	ATAAGCTTTTGTTCGGATCCAATTCTTTCCAAGTATTTGG
Lib.1.PLMC	GGGGCTGAGGTGTGCATTTACNNSDRCDCAWMCWHCGCACGAAATGTTTTAGATTCACGA
Lib.2.PLMC	CCTTGTTTTGAGGAAAATTACRRSRVSGCAVNWKMAGCAGGGATGAAAGTAGGCGTGTAC
Lib.3.PLMC	ATCGATATCAGCAATTGGCAARRSRRCATARACKDDRRSCAGTTAAAAGAGCGAGGCTAC
Lib.4.PLMC	ATAAAAATCACAGAGGGGAAGDVCTWCVNAGACYCANNSTTTGAGGAAAATTACAACAAG
Lib.5.PLMC	GCCGAAGAATTAGAGCGTTTANNCGGAVNRGAAGTANBSATTTACTGCAATACGAATTAT
Lib.6.PLMC	CAAAAAGGCATTAACTTGAACNVASTAAAANMSNNSGGATACGATGTATGTTATATAAAA
Lib.7.PLMC	ATAGAGCAAGCCAACAACATCGYAVVSRYGWTAVRSRRCAAACATATCGACTGCAAGATT
Lib.8.PLMC	GTACGAACGTTGGGAAACAAAVNWNNWGACTKWAAAMYAGCCATAGACGTAGAGCAGACC
Lib.9.PLMC	CAGACCGACGGCCTTTCTNNSRVSGAAWTAAVCVASAGCGTACTTCAACTTGCCGAA
Lib.10.PLMC	AATAATAGCGTACTTCAACTTDYRVRSGAASTAVAAVVSTTAATAGGGGCTGAGGTGTGC
Lib.1.NNK	GGGGCTGAGGTGTGCATTTACNNKNNKNNKNNKNNKNNKCGAAATGTTTTAGATTCACGA
Lib.2.NNK	CCTTGTTTTGAGGAAAATTACNNKNNKNNKNNKNNKNNKGGGATGAAAGTAGGCGTGTAC
Lib.3.NNK	ATCGATATCAGCAATTGGCAANNKNNKNNKNNKNNKNNKCAGTTAAAAGAGCGAGGCTAC
Lib.4.NNK	ATAAAAATCACAGAGGGGAAGNNKNNKNNKNNKNNKNNKTTTGAGGAAAATTACAACAAG
Lib.5.NNK	GCCGAAGAATTAGAGCGTTTANNKNNKNNKNNKNNKNNKATTTACTGCAATACGAATTAT
Lib.6.NNK	CAAAAAGGCATTAACTTGAACNNKNNKNNKNNKNNKNNKTACGATGTATGTTATATAAAA
Lib.7.NNK	ATAGAGCAAGCCAACAACATCNNKNNKNNKNNKNNKNNKAAACATATCGACTGCAAGATT
Lib.8.NNK	GTACGAACGTTGGGAAACAAANNKNNKNNKNNKNNKNNKGCCATAGACGTAGAGCAGACC
Lib.9.NNK	CAGACCGACGGCCTTTCTNNKNNKNNKNNKNNKNNKAGCGTACTTCAACTTGCCGAA

Lib.10.NNK	AATAATAGCGTACTTCAACTTNNKNNKNNKNNKNNKNNKTTAATAGGGGCTGAGGTGTGC
Ni5N501	AATGATACGGCGACCACCGAGATCTACACTAGATCGCTCGTCGGCAGCGTC
Ni5N502	AATGATACGGCGACCACCGAGATCTACACCTCTCTATTCGTCGGCAGCGTC
Ni7N701	CAAGCAGAAGACGGCATACGAGATTCGCCTTAGTCTCGTGGGCTCGG
Ni7N702	CAAGCAGAAGACGGCATACGAGATCTAGTACGGTCTCGTGGGCTCGG
SetFWD_N	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGNGGGATCGATATCAGCAATTGGC
SetFWD_NN	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGNNGGGATCGATATCAGCAATTGGC
SetFWD_NNN	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGNNNGGGATCGATATCAGCAATTGGC
SetREV_N	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGNCCCAATCGTGAATCTAAAACATT
SetREV_NN	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGNNCCCAATCGTGAATCTAAAACATT
SetREV_NNN	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGNNNCCCAATCGTGAATCTAAAACATT

Multi-mutant libraries of LysCP2 were generated in two steps: (1) the generation of the mutagenic region with the downstream sequence; (2) the use of those fragments as megaprimers to complete the gene by appending the upstream sequence. First, pCT40:LysCP2 was amplified with library-specific mutagenic forward primers, Lib.#.PLMC and Lib.#.NNK for each, with a universal reverse primer Lib.all.REV via polymerase chain reaction with Q5 Polymerase (New England Biolabs), each in separate reactions. These fragments were then isolated by gel electrophoresis and amplified in the same manner but with a new universal forward primer Lib.all.FWD and the pCT40:LysCP2 plasmid in a two-cycle reaction. This product for all libraries was then mixed and transformed with pCT40:LysCP2 vector that had been linearized with NheI-HF and BamHI-HF as outlined in Woldring et al[38].

3.5.11 Stability assay via yeast surface display

Yeast transformed with pCT40 plasmids were induced by first growing to an OD600 of ~1.0 in SD-CAA. Cells were then pelleted for 1 minute at 6000x*g* and resuspended in SG-CAA and induced overnight at 20 °C, 250 rpm. The following day, yeast were washed twice with PBSA (PBS with 0.1% bovine serum albumin). Cell density was determined by dilution and measurement at OD600. 10 million cells were diluted into 200 µL PBSA per sample. Proteinase K (New England Biolabs) was diluted in PBSA to a concentration of 2.0×10^{-4} U/µL. The tubes of dilute proteinase K and yeast were separately incubated at 30-45 °C for 5 minutes. Then 200 µL proteinase K dilution was added to each yeast sample, mixed briefly by pipetting, and incubated for 10 minutes. 600 µL of ice-cold PBSA was then added, and samples were placed on ice. All future steps were performed on ice or at 4 °C in centrifuges.

Yeast were washed twice with 500 µL PBSA, labeled with 5 µg/mL anti-c-myc (clone 9e10, BioLegend, 626802) and 1 µg/mL anti-HA (chicken anti-HA, Abcam, ab9111) for 1 hour, washed twice with 500 µL PBSA, labeled with 2 µg/mL goat anti-mouse Alexa Fluor 647 (Thermo Fisher Scientific, A-21235) and 2 µg/mL goat anti-chicken Alexa Fluor 488 (Thermo Fisher Scientific, A-11039) for 30 minutes. Yeast were then washed twice in 500 µL PBSA and resuspended in 300 µL PBSA for sorting (fluorescent activated cell sorting, FACS) using a FACSAria II (Becton Dickinson Bioscience). Two sorting gates, gathering HA positive cells with high and low ratios of HA to c-myc signal, were used.

3.5.12 High-throughput sequencing and preprocessing

Following FACS, yeast populations were outgrown in SD-CAA overnight. The following day, to isolate plasmids for sequencing, approximately 10 million cells were centrifuged and resuspended in 200 µL of Solution 1 (0.13 g/100 mL NaH_2PO_4*H_2O, 1.09 g/100 mL Na_2HPO_4*$7H_2O$, 1 M sorbitol, 10 mM 2-mercaptoethanol), to which 10 µL of Longlife Zymolase (G-Biosciences) was then added, after which the mixture was incubated at 37 °C for 60 minutes. Following incubation, plasmid DNA was purified with GenCatch (Epoch Life Science) according to manufacturer instructions, substituting 200 µL of MX2 and 400 µL of MX3. Samples for Illumina sequencing were prepared by amplification with Ni5N501-502, Ni7N701-702, SetFWD_[N,NN,NNN], and SetREV_[N,NN,NNN] via Q5 PCR (New England Biolabs), and appropriate product isolated via gel electrophoresis. DNA populations were mixed to a final concentration of 5 ng/µL with relative proportions of each population matching the relative proportions of yeast collected in each. Sequencing using version 3 chemistry on an Illumina MiSeq generated 1.5 million reads.

Sequences were processed to remove those with more than one expected error using USearch v11[140](https://drive5.com/usearch/manual/whatsnewv11.html). Nucleotide sequences were then compared with wild-type LysCP2 and all library designs. Unique nucleotide sequences with greater than 3 reads were assigned to libraries with zero tolerance for mutations outside of degenerate regions. Singletons and doublets with at most one deviation outside degenerate regions were also assigned to most probable origin libraries.

Nucleotide sequences were then translated and counts pooled based on amino acid sequence, hereafter sequence. A sequence was then designated as stable or unstable if it occurred in predominantly the stable or unstable pool, respectively.

3.5.13 Site-wise logistic modeling of yeast-displayed protein stability and stability score

Similar to the homology model presented previously, *E*, now considered a stability score, can be modeled using a first-order Potts model ($J_{i,j}$ = 0). To infer the parameters of the model, logistic regression can be implemented over the dataset with model form:

Equation 4

$$P(stable|\sigma) = \frac{1}{1 + e^{-E(\sigma)}}$$

This model was regressed on the data using the *sklearn* package in Python (https://scikit-learn.org/stable/). Due to overfitting concerns, L_2 regularization with λ = 10. This value was chosen as it had maximum cross-validated accuracy while maintaining linear correlation between original-dataset parameter inference and means of those inferred from bootstrapped resampling (Figure 3.15). This bootstrapped error enabled the design of mutants which combined those features (parameters) that were consistently beneficial across the different bootstrap samplings.

The stability score of a sequence is defined as *E(σ)* using parameters inferred from the library's stability observations. The predictive performance of this experimentally-derived model was assessed via 10-fold cross-validation. To simulate predicting unseen observations from experimental data, the set of unique observations for each library was subdivided randomly into 10 folds with equal size. For each fold the prediction accuracy was assessed via a model regressed on the collection of observations in the other 9 folds.

3.5.14 Design of multi-mutant designs

First, the set of mutations for each library whose stability score contributions were significantly greater than wild-type (Figure 3.9) was identified. Within this set – sites 112, 127, and 130 – each had a single significant beneficial mutational choice: N112D, A127I, and C130V. Where multiple choices were possible selection was done to explore interesting options or preserve characteristics of the wild-type residue. N54R was selected instead of N54K to preserve possible hydrogen bonding. Though G43E was not predicted to be significantly different than wild-type it was selected due to its similarity to G43D and G43N. Y107 had multiple options, of those Y107R and Y107G were selected as representatives of the two most beneficial sets of mutations (positive and small, respectively).

Genes encoding the catalytic domains of designs containing a subset of combinations of these selected mutations were synthesized (Twist Bioscience), and assembled (HiFi, New England Biolabs) with the cell-wall binding domain of LysCP2 in the pETh vector.

3.5.15 Data availability

Sequences of mutant constructs as generated by Illumina sequencing are available as a repository on Zenodo (https://doi.org/10.5281/zenodo.2600306). Code required to reproduce analysis is available upon request.

3.6 Acknowledgements

This work was supported by a grant from the National Institutes of Health (R01 GM121777). Note that the authors have submitted a patent application pertaining to some of the engineered lysin molecules in this work.

3.7 Supporting Information

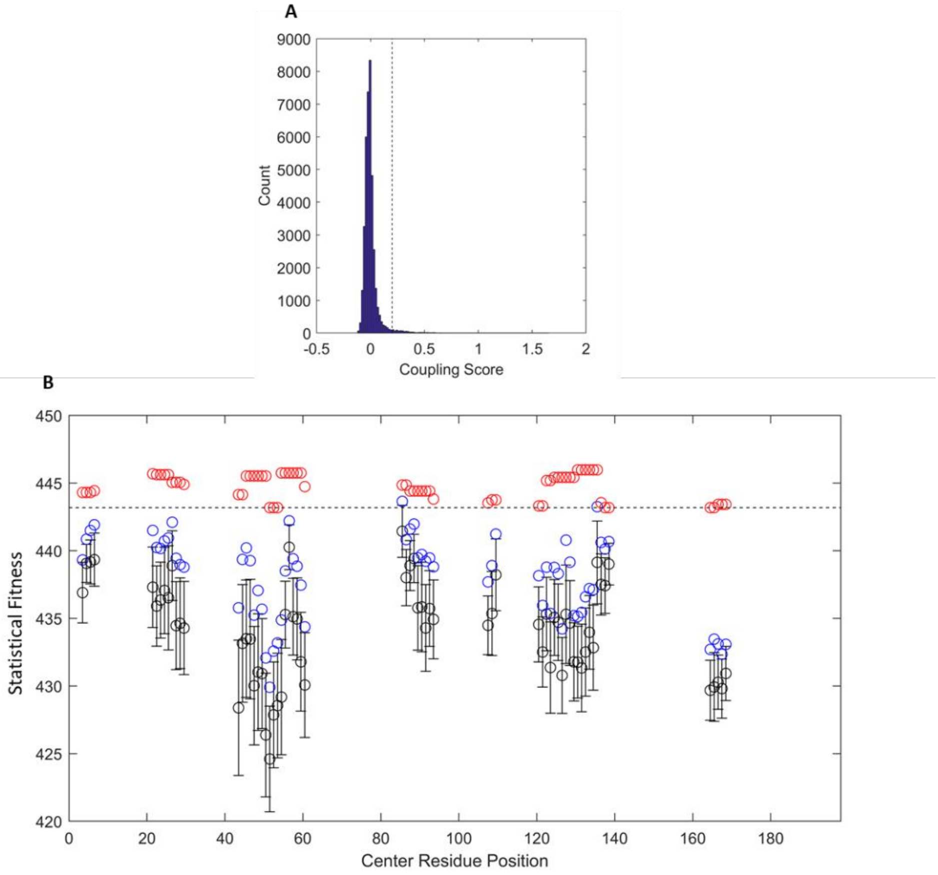


Figure 3.11.Selection of non-catalytic residues and designed library fitnesses. (A) Coupling scores can be classified as a mixture of a skewed normal distribution (left) representing noisy couplings with an extended tail representing significant couplings. (B) Statistical fitness of libraries designed following Algorithm 1 in 6-residue windows. Libraries were designed with a maximum diversity of 16,000. For each library, displayed are the maximum single-mutation score (red) and summaries of random sampling from designed libraries including the mean ± standard deviation (black) as well as the 90th-percentile (blue).

3.7.1 Supplemental Algorithm 1. Designed library sub-selection

1. Provide: a set of designed libraries with their maximum single-mutation statistical fitness and their 90th-percentile statistical fitness from random sampling.
2. Repeat the following until no possible libraries remain:

3. Select the set of libraries with maximum single-mutation statistical fitnesses
4. If the size of set (a) is greater than one:
    a. Select from set (a) the library with the maximum 90th-percentile statistical fitness
5. Add the selected library to list of final libraries
6. Remove all libraries with residues that overlap with the selected library

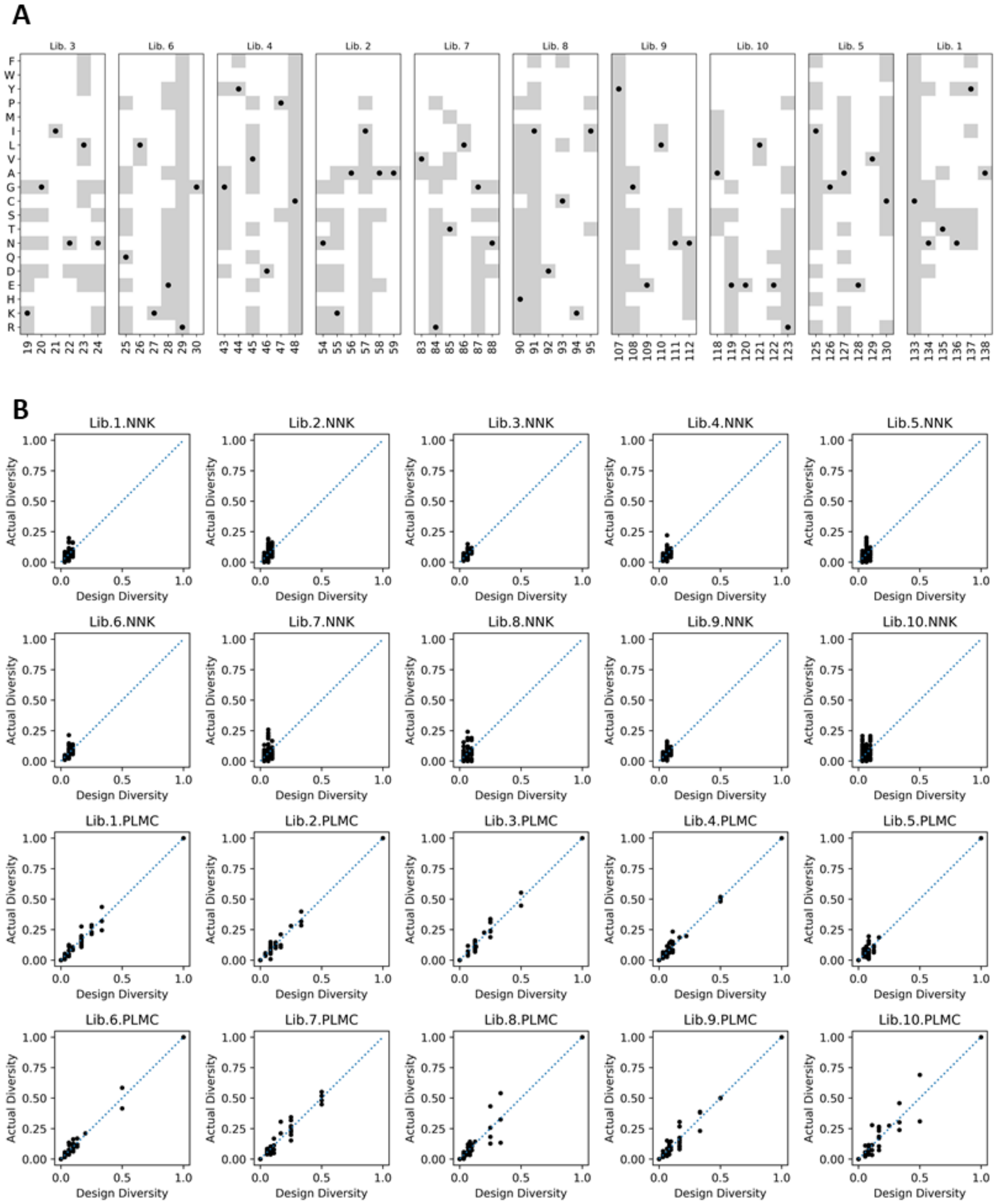


Figure 3.12. Diversity at each position of the 10 designed libraries. (A) The wild-type (dot) and designed diversity (grey) are indicated for each library. Amino acids are listed as rows, with positions listed as columns. (B) The designed and experimentally observed library diversities for each library.

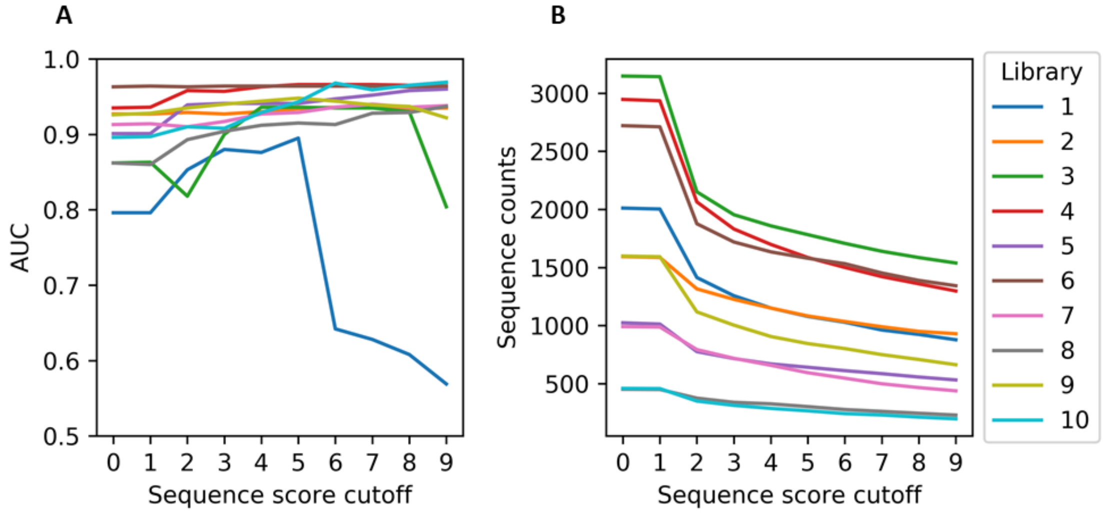


Figure 3.13. Performance of logistic regression as assessed by AUC at different levels of sequence score cutoff. (A) Area-under-curve of a receiver operating characteristic for different libraries given different levels of sequence score cutoff. (B) Number of sequences for each library after applying sequence score cutoff.

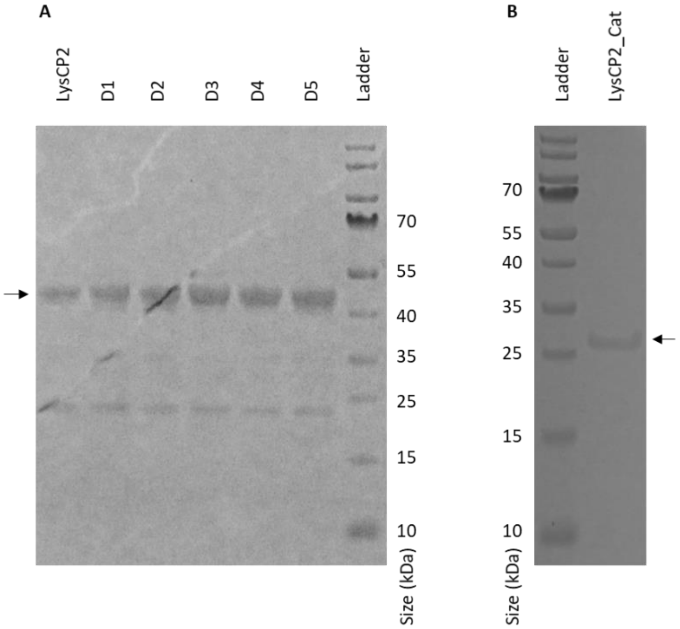


Figure 3.14. Purity assessment of lysin proteins by SDS-PAGE gel. (A) Full-length lysin constructs. Expected molecular weight of LysCP2: 41.1 kDa. (B) Isolated catalytic domain of LysCP2, expected molecular weight: 24.2 kDa.

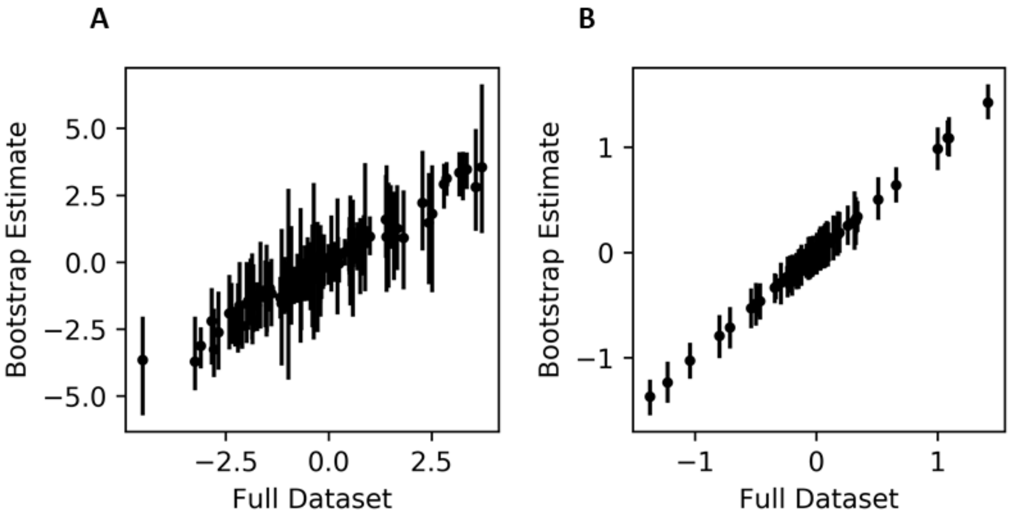


Figure 3.15. Selection of regularization coefficient. (A) λ = 0.1 shows high variance due to overfitting of bootstrap samples and produces poorly correlated agreement between full-dataset and bootstrap samples. (B) λ = 10.0 shows reduced variance and maintains strong correlation.

Table 8. Stable and unstable sequences above thresholds across the collection of libraries.

	Positions		NNK		PLMC	
Library	Start	End	Unstable	Stable	Unstable	Stable
1	133	138	109	1	873	95
2	54	59	160	1	588	334
3	19	24	626	1	1120	35
4	43	48	390	4	865	327
5	125	130	86	11	279	264
6	25	30	646	4	523	406
7	83	88	61	3	303	226
8	90	95	26	1	220	54
9	107	112	246	14	126	458
10	118	123	35	3	184	41

Chapter 4 Computationally-aided discovery of LysEFm5 variants with improved catalytic activity and stability

Tsvetelina H. Baryakova, Seth C. Ritter, and Benjamin J. Hackel

The work contained in this chapter, including computational analysis, experimental design and implementation, data interpretation, and composition of text was conducted by T.H.B, an undergraduate student working under the mentorship of S.C.R. and B.J.H., and S.C.R. In particular, S.C.R. was responsible for generation of the initial idea for the target and proposal to focus on secondary residues for optimization; assembly of sequences and generation of PLMC-fitted models; analysis of NGS data; sequencing, production, purification, and testing of isolated variants. The chapter content has been submitted for publication.

4.1 Synopsis

Bacteriophage-derived lysin proteins are potentially effective antimicrobials that would benefit from engineered improvements to their bioavailability and specific activity. Here, the catalytic domain of LysEFm5, a lysin with activity against vancomycin-resistant *Enterococcus faecium* (VRE), was subjected to site-saturation mutagenesis at positions whose selection was guided by sequence and structural information from homologous proteins. A second-order Potts model with parameters inferred from large sets of homologous sequence information was used to predict the average change in the statistical fitness for mutant libraries with diversity at pairs of sites within the secondary catalytic shell. Guided by the statistical fitness, nine double mutant saturation libraries were created and plated on agar containing autoclaved VRE to quickly identify and segregate catalytically active (halo-forming) and inactive (non-halo-forming) variants. High-throughput DNA sequencing of 873 unique variants showed that the statistical fitness was predictive of the retention or loss of catalytic activity (AUC = 0.840 – 0.894), with the

inclusion of more diverse sequences in the starting multiple sequence alignment improving the classification accuracy when pairwise amino acid couplings (epistasis) were considered. Of eight random halo-forming variants selected for more sensitive testing, one showed a 2.1 ± 0.2 – fold improvement in specific activity and an 11.5 ± 0.8 °C increase in melting temperature as compared to the wild-type. Our results demonstrate that a computationally-informed approach employing homologous protein information coupled with a mid-throughput screening assay allows for the expedited discovery of lysin variants with improved properties.

4.2 Introduction

4.2.1 Antimicrobial lysin proteins

The misuse of antibiotics is a growing problem in the twenty-first century[141]. In addition to the development of antibacterial resistance and subsequent loss of treatment efficacy, the use of broad-spectrum antibiotics can reduce the diversity of a patient's commensal flora[142]. This reduction in diversity has been correlated with the onset of multiple health issues, including several inflammatory and autoimmune diseases[143]. The development of alternative antimicrobial strategies that offer improved specificity could help mitigate both of these issues.

Native bacteriophage-derived lysin proteins are released during the last stage of the virus's lytic cycle to degrade the cell wall of the Gram-positive bacteria host[144]. These antimicrobial proteins have the potential to be used as effective alternatives that ameliorate many of the negative side effects of conventional antibiotics. The mechanism of action of lysins generally involves the cleavage of an essential and highly-conserved peptidoglycan bond in the bacterial cell wall; as such, the development of resistance to these antimicrobial proteins is expected to occur less easily[105,145–147]. Additionally, many lysins are specific, enabling them to kill pathogens without exhibiting significant activity against commensal bacteria species[148,149]. However, engineering lysins to have more desirable properties that contribute to improvements in functional bioavailability, such as higher rates of catalytic activity, heightened solubility, or increased thermal stability, is almost always necessary before a sufficient therapeutic response in the infected host can be achieved[105,150]. Improvements to catalytic activity in particular can reduce the necessary concentration, both in formulation and physiologically, thus decreasing the required solubility.

Lysins generally possess both a catalytic domain and cell wall-binding domain (CWBD) connected together via a flexible linker[100]. Catalytic domains can be categorized into five groups depending on their substrate: N-acetyl-β-D-muramidases (lysozymes), lytic transglycosylases, N-acetyl-β-D-glucosaminidases, N-acetylmuramoyl-L-alanine amidases, and endopeptidases[147]. N-acetylmuramoyl-L-alanine amidases hydrolyze the amide bond between N-acetylmuramic acid, a constituent in the repeating disaccharide of the glycan chain in the cell wall of Gram-positive bacteria, and L-alanine, the first amino acid residue of the stem peptide responsible for cross-linking neighboring glycan chains[151]. The structure of each major type of catalytic domain is well-conserved between lysins derived from different phage species[152], as is generally observed for functional sites in enzymes[153]. The CWBD of a lysin, in contrast, is responsible for co-localizing the catalytic domain with its substrate and usually possesses specific affinity for a particular species or subgroup of bacteria[105]. The two domains, although connected, are often thought of as capable of carrying out mechanistically distinct functions. This has allowed for the creation of chimeric lysins with altered activity and specificity via domain swapping[154,155].

LysEFm5 is a lysin with an N-acetylmuramoyl-L-alanine amidase as its catalytic domain. LysEFm5 was previously isolated and described as having killing activity against vancomycin-resistant *Enterococcus faecium* (VRE)[156]. E. faecium (EF) is found in the gastrointestinal tract of healthy individuals but can pose a serious threat if spread to the bloodstream, urinary tract, or wound of an immunocompromised patient, most often from a nosocomial infection. Vancomycin is typically only used as a “last resort” to treat infections of Gram-positive bacteria that are unresponsive to other antibiotics. As such, vancomycin resistance in patient-derived EF isolates has been correlated with poor patient outcome and even death[157–159].

LysEFm5 was shown to have a broader antibacterial range than IME-EFm5, its parent phage. LysEFm5 was able to lyse 19 out of 23 strains of EF, 7 of them VRE (as compared to 1 out of 23 strains of EF lysed by IME-EFm5) but possessed no apparent killing activity against the other Gram-positive or Gram-negative bacteria tested. The homology-based structure of the catalytic domain of LysEFm5 has also been reported[156]. E90 and T138 have been identified as putative catalytic residues and H27, H132, and C140 as putative zinc-coordinating residues. These two sets of residues are generally well-conserved in the ligand-binding groove of zinc-dependent peptidoglycan hydrolases[160,161].

LysEFm5 was chosen for further study based on the clinical relevance of its target, availability of homology-based structural information, and specificity towards EF (in contrast to other broadly-active, anti-EF lysins[162]).

Nine site-saturation mutagenic libraries were created to study the effectiveness of using structure and sequence information to direct lysin engineering efforts. To determine which residue positions in LysEFm5 to diversify, it was desired to find sites in the catalytic domain that were not critical for the catalytic activity of the protein but played a role in stabilizing other key, functional residues. In addition to identifying these residues using the crystal structure of a close homolog to LysEFm5, the choice of positions used in double mutant libraries was refined further using a computationally-informed approach. The overall methodology is given in Figure 4.1.

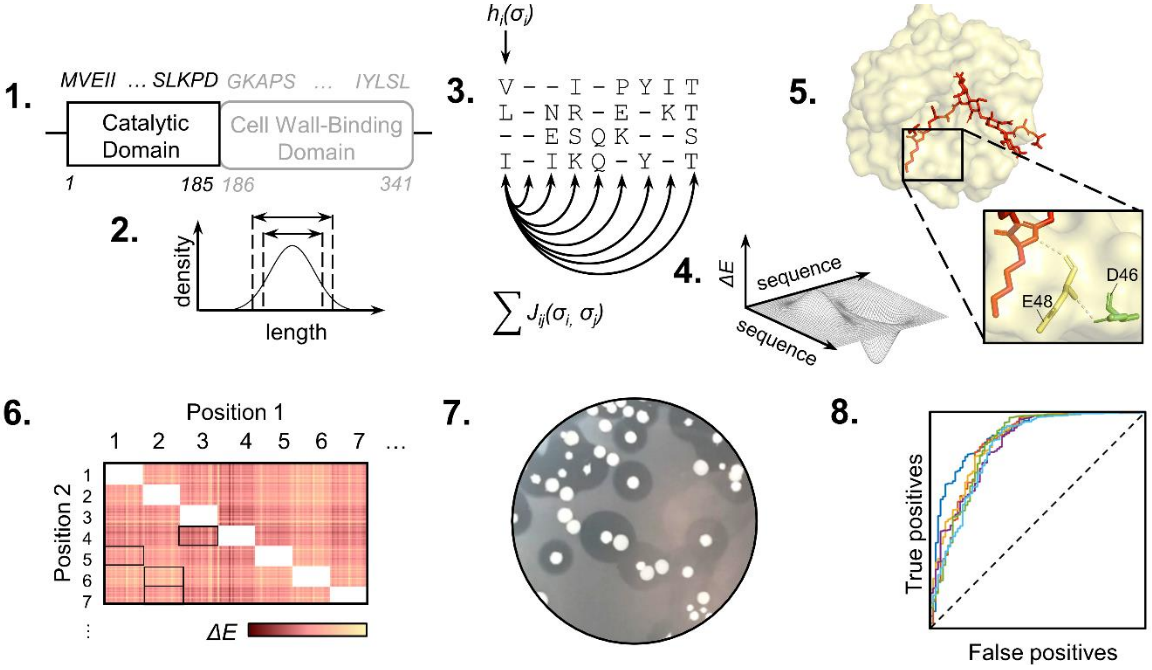


Figure 4.1. Research methodology. (1) LysEFm5 catalytic domain is used in an iterative homology search. (2) Resulting homologous sequences are subject to length cut-offs. (3-4) A structure-based MSA is created for each for each group of sequences. PLMC is used to infer site-dependent and pairwise coupling parameters and create a generative model for predicting the change in statistical fitness, Δ*E*, of mutants. (5) Residues in the putative secondary interaction shell of LysEFm5 are identified using the ligand-binding crystal structure of a homologous protein. (6) A matrix of predicted double mutation outcomes is created using PLMC. This is used to guide position selection for combinatorial library design. (7) Halo-forming and non-halo-forming variants from each library are observed, binned, and deep-sequenced. (8) The experimental retention of function is compared to the predicted statistical fitness for mutants.

4.2.2 Homolog-guided library design

Homologous protein sequences contain information about the structural and functional constraints imposed on a protein over the course of its evolution, which can be of value when directing engineering efforts[152,163,164]. The natural sequence record is assumed to contain mutations that allow for the retention of a protein's biological function. Sequences of protein homologs are often highly variable despite marked similarities in their structure and function. This suggests that the site-specific, or independent, trends in amino acid conservation alone may be insufficient to model sequence constraints experienced by proteins over evolutionary time[165]. Recently, statistical methods that consider the interactions between pairs of residues in an attempt to capture the nature of non-independent, or epistatic, mutations have emerged[128,165–168]. Models such as these that take epistatic interactions into account have been shown to more accurately predict the effects of mutations on a protein's function as compared to independent models that neglect pair couplings[128,169].

It has been shown that if the mutation of a protein is assumed to be a reversible Markov process, the resulting maximum-entropy ensemble that represents the distribution of natural sequences at equilibrium (functionally, long evolutionary times from the shared ancestral protein) obeys a Boltzmann distribution[127]. Thus, the probability P(σ) of observing any full-length amino acid sequence, σ, in the system can be computed.

Equation 3

$$P(\sigma) = \frac{e^{E(\sigma)}}{Z}$$

It is further assumed that the energy function E(σ) in Eq. 1 takes the form of a second-order Potts model with parameters that are fitted to reproduce the empirically-observed sitewise and pairwise statistics in the multiple sequence alignment (MSA)[170].

Equation 2

$$E(\sigma) = \sum_{i} h_i(\sigma_i) + \sum_{i} \sum_{j>i} J_{i,j}(\sigma_i, \sigma_j)$$

Where E(σ) is the statistical fitness, hi are the site-dependent constraints, and $J_{i,j}$ are the pairwise coupling constraints at positions i and j in the full-length amino acid sequence, σ.

The exact calculation of the parameters in the Potts model requires determination of the partition function, Z, in the Boltzmann distribution equation - a sum over all possible 20L protein sequences. The pseudolikelihood maximization inference method can be used to simplify this generally intractable calculation, requiring instead the calculation of L individual sums over 20 amino acids[165]. A Potts model with parameters inferred using pseudolikelihood maximization has been shown to accurately identify strongly-coupled pairs of amino acids, making pseudolikelihood maximization a useful inference method that is less computationally intensive than alternative, more precise methods[165,166,170,171].

Although not linked to any one molecular phenotype, the statistical fitness is more likely to correlate with a phenotype directly related to an organism's survival that would be selected for throughout its evolutionary history[128]. In this manner, the effect of mutation(s) on a protein can be predicted by calculating $\Delta E = E(\sigma_{mutant}) - E(\sigma_{wild\text{-}type})$, in so far as predicting whether the mutation(s) increase ($\Delta E > 0$) or decrease ($\Delta E < 0$) the probability of observing the new sequence in the protein family described originally by the MSA.

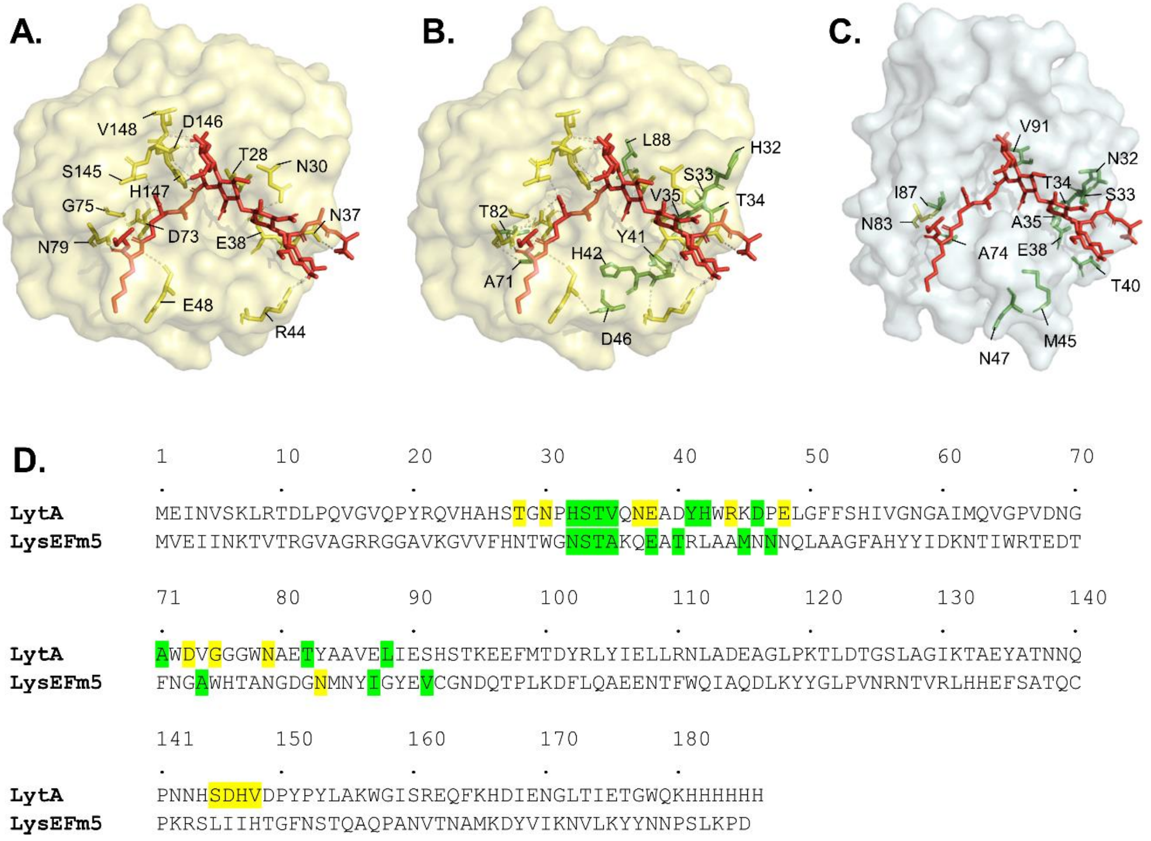
The framework of this methodology has been previously developed and released as an open-source code by the Marks lab at Harvard under the name pseudolikelihood maximization coupling inference (PLMC)[128]. PLMC was used to build a predictive model of mutational outcomes in the LysEFm5 catalytic domain and direct the selection of amino acid sites for site-saturation mutagenesis.

4.3 Results

4.3.1 Statistically-guided design and construction of a mutant lysin library

Only the catalytic domain of LysEFm5 was chosen for alteration; the CWBD was not edited in order to maintain the desired specificity of the protein. Within the catalytic domain of LysEFm5, a network of residues interact with the peptidoglycan substrate to hydrolyze the amide bond between N-acetylmuramic acid and L-alanine at the first position of the stem peptide. Within this network, there are residues that directly interact with the substrate (primary shell) and residues which position and stabilize primary residues without directly interacting with the substrate (secondary shell). We hypothesized that mutating these so-called secondary residues could optimize the catalytic performance of the enzyme, as has been seen before in other enzymes[172], and possibly improve antimicrobial activity.

The catalytic domain of the major pneumococcal autolysin LytA, initially evaluated due to the availability of the solved crystal structure of the domain bound to a synthetic peptidoglycan ligand[173], was identified as a homolog of the catalytic domain of LysEFm5 via sequence alignment (with a sequence similarity of 0.23). SWISS-Model provided a QMEAN score of -7.58, sequence identity of 8.33, and coverage of 1.00 when the catalytic domain of LytA (PDB code: 5CTV) was used as a template to model the first 180 amino acids in the catalytic domain of LysEFm5. Thirteen primary residues and ten putative secondary residues were identified in the structure of the LytA amidase (Figure 4.2.A, B). Eleven structurally-analogous secondary residues (N32, S33, T34, A35, E38, T40, M45, N47, A74, I87, and V91) were selected as candidates for mutation (Figure 4.2.C). One primary residue (N83) that occupied the same space in the structural homolog model of the LysEFm5 molecule as in the aligned structure of the LytA molecule was also included for comparison.



```
           1         10        20        30        40        50        60        70
           .         .         .         .         .         .         .         .
LytA       MEINVSKLRTDLPQVGVQPYRQVHAHSTGNPHSTVQNEADYHWRKDPELGFFSHIVGNGAIMQVGPVDNG
LysEFm5    MVEIINKTVTRGVAGRRGGAVKGVVFHNTWGNSTAKQEATRLAAMNNNQLAAGFAHYYIDKNTIWRTEDT

           71        80        90        100       110       120       130       140
           .         .         .         .         .         .         .         .
LytA       AWDVGGGWNAETYAAVELIESHSTKEEFMTDYRLYIELLRNLADEAGLPKTLDTGSLAGIKTAEYATNNQ
LysEFm5    FNGAWHTANGDGNMNYIGYEVCGNDQTPLKDFLQAEENTFWQIAQDLKYYGLPVNRNTVRLHHEFSATQC

           141       150       160       170       180
           .         .         .         .         .
LytA       PNNHSDHVDPYPYLAKWGISREQFKHDIENGLTIETGWQKHHHHHH
LysEFm5    PKRSLIIHTGFNSTQAQPANVTNAMKDYVIKNVLKYYNNPSLKPD
```

Figure 4.2. Primary and secondary amino acids in LytA and their putative structural analogs in LysEFm5. Molecules are aligned using the *align* command in PyMOL. (A) Surface representation of the LytA molecule, showing primary residues (yellow) interacting with the synthetic peptidoglycan ligand (red). (B) Putative secondary residues (green) that interact with primary residues. (C) Structural analogs of the eleven secondary residues and single primary residue in LysEFm5. (The ligand was superimposed following the structural alignment of LytA and LysEFm5 in PyMOL, and is not part of the reported

structure of LysEFm5.) (D) Map of the location of the relevant putative secondary and primary residues in the amino acid sequences of LytA and LysEFm5. Note that although the hydroxyl group, -OH, of E38 in LytA was predicted to bind to the O in the CH_2OH group of *N*-acetylmuramic acid in the peptidoglycan ligand, the analog E38 in LysEFm5 was still selected as a secondary residue (most primary residues were found to bind the ligand twice or more).

A computational model of sitewise and pairwise interactions, based on the sequence alignment of homologous sequences, was used to determine which pairs of sites to simultaneously mutate in the experimental libraries. To identify homologs to the LysEFm5 amidase domain sequence, a search of the UniProtKB protein database was performed via JackHmmer[129]. Sequence searches were constrained to three levels of evolutionary depth by restricting the acceptable taxonomy of the host organism to all organisms (no restrictions), bacteria only, or firmicutes only. Sequences that were either extremely short or long were excluded from further consideration by applying either a lax or stringent cut-off criterion for outlier detection (Materials and Methods). This generated a total of six sets of starting homologous sequences (Table 9). Each set was independently input into PROMALS3D, an alignment tool that incorporates both sequence and structure information[137], to create an MSA.

Table 9. Predictive performance of the statistical fitness when different groups of homologous sequences are used in the starting MSA.

Designation	Length[a]	No. of Sequences	Effective No. of Sequences[b] (±SD)	AUC for Epistatic Model (± SE)	AUC for Independent Model (± SE)
$All_{lax\text{-}29k}$	141 – 199	29,498	6,352	0.894	0.807
$All_{stringent\text{-}23k}$	155 – 186	23,176	4,809	0.856	0.857
$All_{lax\text{-}3k}$		3,037	1,420 ± 33	0.815 ± 0.013	0.744 ± 0.026

Bacteria$_{\textit{lax-27k}}$	137 – 200	26,950	5,565	0.868	0.888
Bacteria$_{\textit{stringent-23k}}$	149 – 188	23,194	4,595	0.851	0.871
Bacteria$_{\textit{lax-3k}}$		3,037	1,309 ± 39	0.839 ± 0.006	0.782 ± 0.013
Firmicutes$_{\textit{lax-3k}}$	133 – 201	3,037	940	0.852	0.795
Firmicutes$_{\textit{stringent-2k}}$	163 – 192	2,007	600	0.840	0.830
Firmicutes+All$_{\textit{lax-3k}}$		3,037	1,344 ± 34	0.856 ± 0.005	0.818 ± 0.014
Firmicutes+Bacteria$_{\textit{lax-3k}}$		3,037	1,317 ± 30	0.851 ± 0.004	0.814 ± 0.014
Firmicutes+Non-Bacteria$_{\textit{lax-2k}}$		5,585		0.865	0.799

[a]The acceptable amino acid lengths across the six initial groupings.

[b]The effective number is the sum of the inverse of the neighborhood size of each sequence, where the neighborhood is defined as the number of sequences within 80% identity.

[c]Results are presented as mean values for twenty sub-samplings from the parent group(s)

PLMC was then used to infer the parameters of a second-order Potts model for each MSA. Without knowledge a priori regarding the effect of the sequence diversity in the starting MSA on prediction accuracy, it was hypothesized that the most constrained and least diverse set of data would enable the most accurate prediction of activity. Thus, the MSA containing the least diverse set of sequences (Firmicutes$_{\text{stringent-2k}}$) was used to predict the change in statistical fitness, ΔE, as compared to the wild-type (WT) for all possible double mutants across the twelve sites of interest (Figure 4.3). Simultaneous mutation of S33 and T40 yielded the highest ΔE values; as such, these sites were randomized in Library 1. To evaluate the predictive performance of the statistical fitness parameter, seven additional combinations of positions were selected with a range of average ΔE values upon mutation, from the highest value of - 4 ± 2 observed for Library 1 to the lowest value of -12 ± 3 observed for Library 8. (Figure 4.3, Figure 4.4). Library 8 contained the putative primary residue, N83.

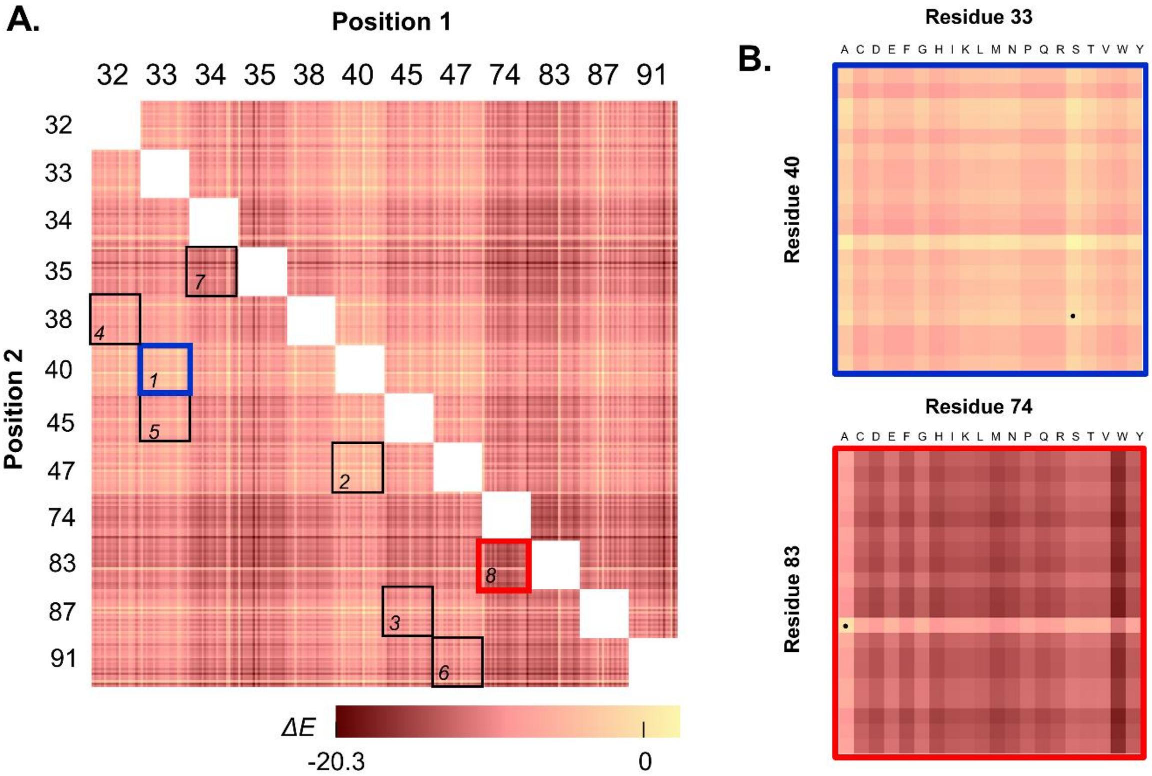


Figure 4.3. The predicted changes in statistical fitness for double mutants. The change in statistical fitness compared to the WT, *ΔE*, for all double mutants with diversity at positions 32, 33, 34, 35, 38, 40, 45, 47, 74, 83, 87, and/or 91 was computed using PLMC with inputs from MSA group *Firmicutes*$_{\textit{stringent-2k}}$ (Table 9). (A) The eight libraries chosen for creation are boxed. (B) A closer look at the predicted *ΔE* values for Library 1 (the library with the highest average *ΔE* value, -4 ± 2) and 8 (the library with the lowest average *ΔE* value, -12 ± 3). The dotted squares represent WT residues.

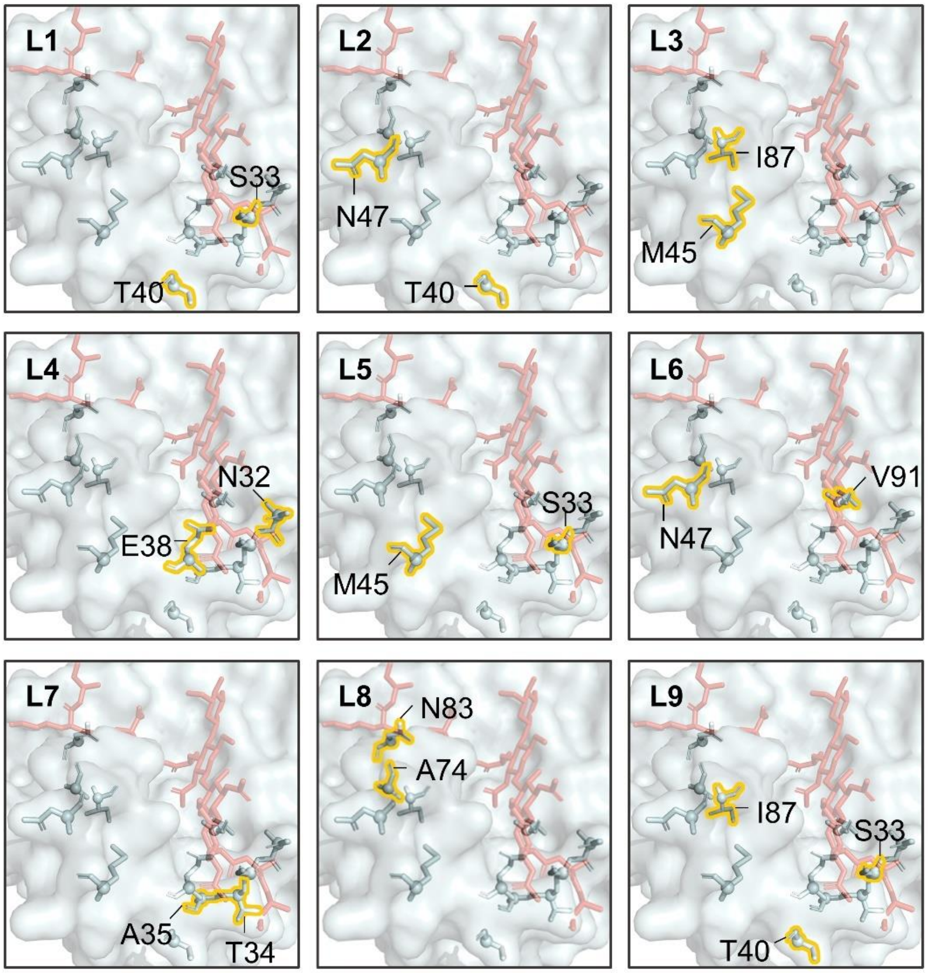


Figure 4.4. Location of residues in each library (L) relative to superimposed ligand in red

One additional library was designed with amino acid diversity constrained based on the predicted statistical fitness. A matrix of the predicted ΔE values from PLMC for single mutants occurring at each of the twelve sites was discretized and input into SwiftLib, an algorithm that specifies a degenerate codon library to yield the desired amino acids at several positions based on a user-defined array of integers describing a favoring or disfavoring of all amino acids (here, based on ΔE)[174]. The resulting optimal library (Library 9) diversified the same two positions as Library 1[127,174], but also included the single mutation I87L. This mutation had a positive predicted ΔE value (+0.21) and was thus highly favored; only three positive ΔE values were observed across the single mutants in general, the other two of which occurred at site 40 (T40N and T40D).

To generate the libraries, gene fragments of the WT were amplified via PCR with mutation-encoding primers. Overlapping fragments were combined via Gibson assembly to yield a collection of plasmids encoding the entire LysEFm5 gene with two randomized codons at the desired sites for each library[175]. Upon assembly, clones from each library were transformed into high efficiency electrocompetent cells. Sequencing of random colonies confirmed that the libraries encoded the entire LysEFm5 gene with diversity at the expected sites.

4.3.2 VRE halo assay to screen LysEFm5 variants for catalytic activity

Recombinant plasmids encoding LysEFm5 mutants were transformed into Escherichia coli for protein expression, and individual clones were assayed for their ability to digest autoclaved vancomycin-resistant *Enterococcus faecium* 8-E9 (VRE) by plating the transformed E. coli on top of agar plates containing autoclaved VRE and isopropyl β-D-1-thiogalactopyranoside (IPTG) used to induce lysin expression. The plating density was such that the majority of individual colonies were easily distinguishable and separate from their neighbors. Upon incubation, colonies expressing an active lysin variant formed a visible halo due to degradation of the surrounding VRE leading to a localized decrease in optical density (Figure 4.5). Halo formation did not occur when E. coli transformed with a plasmid encoding a phage lysin with specific activity against Clostridium perfringens[176] were plated in an identical manner (Figure 4.15). This observation supports the hypothesis that autoclaving the VRE prior to use did not increase its susceptibility to native lysozymes produced by E. coli, or to the activity of a lysin with an alternative specificity. The size of the observed halo is the result of a number of physical properties of the expressed lysins including stability, expression and degradation rates, per-molecule activity, diffusivity, etc.[177]. Therefore, this format does not result in an equal amount of protein produced and subsequently released from each colony, and halo size cannot be directly correlated to specific activity. This assay was instead used in a binary sense to designate a variant as either halo-forming or non-halo-forming, with halo-forming variants assumed to be a subset of all active variants. Assays similar to the one described here have been previously used to screen expression libraries and identify endolysin-producing clones[178], as well as to confirm the production in *E. coli* of two phage-derived lysins with broad activity against multiple strains of *E. faecium* and *E. faecalis*[179]. Though these previous studies chemically permeabilized the *E. coli* to release expressed proteins, the method presented herein relied on only the intrinsic leakage of the host. The mechanism of this release is not known but hypothesized to be the result of cell lysis upon death or increased permeabilization as a result of the overexpression of lysY.

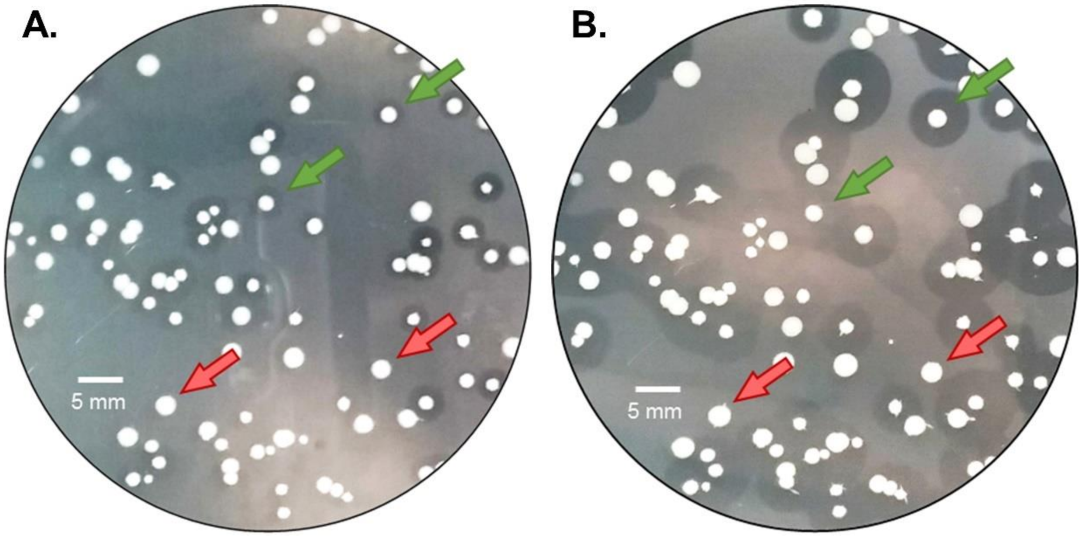


Figure 4.5. Halo formation over time. Libraries were plated on top of LB + kan/VRE/IPTG plates. (A) At approximately 16 – 18 hr following incubation at 37°C, discernable halos

appeared around catalytically active variants (green arrows) and not around catalytically inactive variants (red arrows). (B) At longer times (> 18 hr), the halo radii continued to grow.

Three parallel runs were performed for each library, in which each of approximately 375 colonies were classified as halo-forming or non-halo-forming. This resulted in a total of six bins, corresponding to replicate number and halo-formation designation, per library. DNA isolated from these six bins was sequenced using Illumina MiSeq. Sequences with less than 100 reads were deemed to be erroneous and excluded. Protein sequences translated from the remaining DNA reads were excluded if present in only a single replicate or if they lacked a majority of either halo-forming or non-halo-forming designations. Applying the latter constraint reduced the number of usable data points from 1,731 unique sequences to 873, while greatly improving the classification accuracy of the statistical fitness associated with this method (Figure 4.18).

4.3.3 Secondary site restriction allows for focused library design resulting in a high retention of catalytic activity

Sequencing results showed a rate of activity of between 84 – 100% per library for all classified variants in Libraries 1-7 and 9 (Table 10). There is no general trend in the experimental retention of catalytic activity and the average statistical fitness of these libraries (Figure 4.6). However, Library 8, which had the lowest predicted statistical fitness and diversity at positions N83 and A74, demonstrated a low activity retention of 30%. N83 in LysEFm5, the only primary residue considered in this analysis, is structurally-analogous to N79 in LytA. N79 is a ligand-binding residue that is highly conserved across multiple prokaryotic and eukaryotic-derived peptidoglycan recognition proteins (PGRPs), both with and without amidase activity[173]. In AmiE (the amidase domain of the major autolysin of

Staphylococcus epidermidis) and in human PGRP-Iα, this conserved asparagine residue was shown to hydrogen bond with the carbonyl groups in the second and third amino acids in the peptide stem of MurNAc-L-Ala-D-isoGln-L-Lys, a peptidoglycan analog[161,180]. A74, in contrast, is not predicted to be a primary residue from direct structural comparison, but the average changes in statistical fitness associated with independently mutating A74 and N83 were similar (ΔE = -6 ± 2 for both, compared to ΔE = -3 ± 2 on average for the other ten residues). Even if not directly bound to the ligand and/or playing a pivotal role in stabilizing the transitional state between substrate and product, A74 may stabilize one or more neighboring primary residue(s) in an essential way. The low rate of activity retention observed for Library 8 provides evidence that the statistical fitness parameter was able to predict highly detrimental mutations at a key, conserved site critical for the function of the protein.

Table 10. The number of active and inactive classified mutants in each library.

		Double mutants		All	
Library No.	Diversified Positions	Active (% of total)	Inactive	Active (% of total)	Inactive
1	33, 40	184 (94%)	11	211 (95%)	11
2	40, 47	83 (98%)	2	114 (98%)	2
3	45, 87	53 (82%)	12	67 (84%)	13
4	32, 38	96 (96%)	4	109 (96%)	4
5	33, 45	92 (100%)	0	114 (100%)	0
6	47, 91	24 (100%)	0	45 (98%)	1
7	34, 35	18 (90%)	2	36 (95%)	2
8	74, 83	12 (21%)	46	22 (30%)	51
9	33, 40, 87[b]	74[a] (96%)	3	313 (95%)	15

[a] Triple mutants.

[b] Specific mutation I87L, not implementation of a randomized codon, at this site.

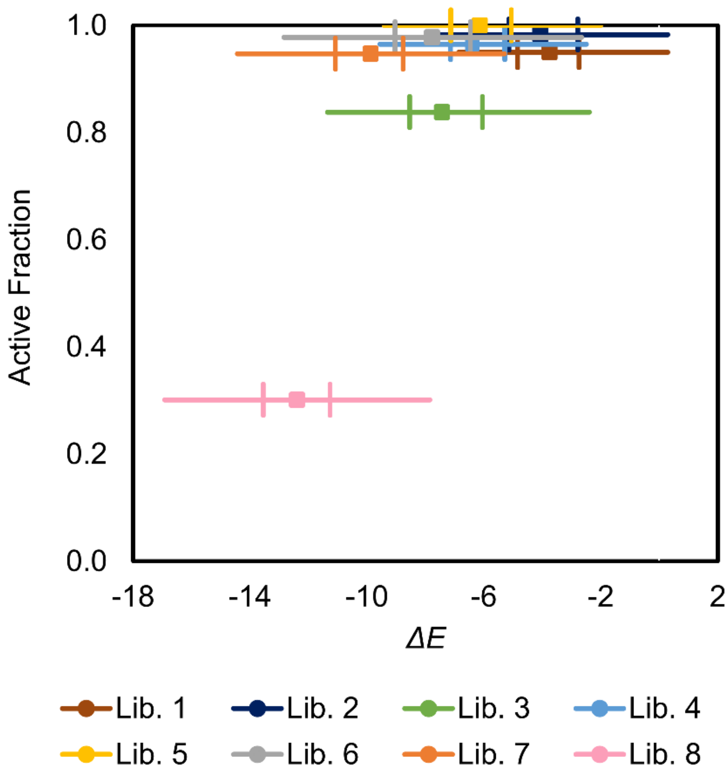


Figure 4.6. Active fraction of variants in library as a function of average ΔE. The difference between the statistical fitness of each variant (using MSA group *Firmicutes*$_{stringent\text{-}2k}$) and the WT, *ΔE*, is plotted against the active fraction of classified mutants in a library. The central square is the median *ΔE* value and the left and right-most vertical lines represent the 25th (Q_1) and 75th (Q_3) percentile. Whiskers extend to the most extreme points not considered outliers; outliers are defined as values less than $Q_1 - W_M(Q_3 - Q_1)$ or greater than $Q_3 + W_M(Q_3 - Q_1)$ where W_M is the maximum whisker length.

The rate of activity retention for Library 9 (with constrained diversity at sites 33 and 40 and the mutation I87L) and Library 1 (with full diversity at sites 33 and 40) were nearly identical at 95%. Further analysis revealed that there were 73 triple mutants present in Library 9 with analogous sequences in Library 1 (sequences with the same diversity at sites 33 and 40, but with the WT residue at site 87). Of these, 73/73 were active in Library 1 (100%) compared to 72/73 in Library 9 (99%). Taken together, these results highlight the flexibility in amino acid identity of the residue at site 87, from the WT I to L.

A similar post-facto analysis was performed for all remaining libraries by comparing the active fraction of classified double mutant sequences to the active fraction of sequences in a hypothetical constrained library (built using a discretized matrix of predicted independent mutation outcomes at each of the two library-specific sites) (Table 11). For libraries 6, 7, and 8, SwiftLib predictions were so constrained that none of the allowable sequences were among those that were experimentally observed. Among the remaining five libraries, only the constrained subset of Library 3 showed any substantial improvement in activity retention (from 82% to 100%).

Table 11. The active fraction for all double mutants in a library compared to active fraction of the subset predicted by SwiftLib.

Library No.	Diversified Positions	SwiftLib Codon at Pos. 1[a]	SwiftLib Codon at Pos. 2[a]	SwiftLib Theoretical AA diversity	Overall Active Fraction	SwiftLib Active Fraction (No. of Observations)
1	33, 40	VNC	NNS	252	0.94	0.97 (173)
2	40, 47	NNS	NDC	252	0.98	0.97 (68)
3	45, 87	RNS	NWC	96	0.82	1.00 (17)
4	32, 38	NNM	BRS	190	0.96	0.92 (38)
5	33, 45	NNS	RBG	126	1.00	1.00 (48)
6	47, 91	NNM	GTA	19	1.00	N/A (0)
7	34, 35	AVC	GCA	3	0.90	N/A (0)
8	74, 83	GCA	AAC	1	0.21	N/A (0)

[a]N = G, T, A, or C; V = G, T, or A; B = G, T, or C; M = A or C; R = A or G; W = A or T; S = G or C; K = G or T.

Figure 4.7 shows the fraction of halo-forming sequences of all classified single, double, and triple mutants based on the identity of the amino acid at a specified position (note that the same positions were mutated in multiple libraries). A deeper analysis of Library 8, which is the only library that included mutations at sites A74 and N83, revealed that the rate of activity retention of single mutants (67%, 10/15) was higher than that of double mutants (21%, 12/58). When the WT residue was retained at A74 and only N83 was mutated, the active fraction was 78% (7/9); conversely, when the WT residue was retained at N83 and only A74 was mutated, the active fraction was 50% (3/6). The intolerance to simultaneous mutations occurring at both sites, but moderate tolerance to single mutations at either site suggests that the retention of one of these two WT residues (or of a close analog with similar polarity and size characteristics) is critical for catalytic activity. In general, the polar, uncharged residues serine, cysteine, and glutamine were the most well-tolerated at site N83 across all mutants (63% (5/8), 50% (1/2), and 50% (2/4), respectively). Mutation to the alanine analog valine was additionally well-tolerated at site A74 (100% (4/4)). In contrast, mutation to the positively-charged, hydrophilic residues arginine and lysine, or to proline and tryptophan, was not tolerated at either A74 or N83. Mutation to arginine was also not tolerated at sites A35, I87, or V91 ((0/2), (0/5), and (0/1), respectively).

Position

Amino Acid	N32	S33	T34	A35	E38	T40	M45	N47	A74	N83	I87	V91
F	1 1. 00	3 1.00			2 1.00	12 0.92	4 1.00	4 1.00			1 1.00	
W	1 1.00	5 0.60			8 1.00	12 0.92	7 1.00	4 1.00	1 0.00	2 0.00	3 1.00	
Y	1 1.00	7 1.00			2 1.00	13 1.00	2 1.00	7 1.00	1 0.00	1 1.00		
P	14 1.00	41 1.00	6 1.00	10 1.00	6 1.00	24 1.00	12 0.58	13 1.00	5 0.00	7 0.00	5 0.60	2 1.00
M	3 0.67	5 0.80			4 1.00	19 0.95	700 0.89	3 1.00	1 1.00	1 0.00	1 1.00	
I	2 1.00	25 1.00	1 1.00		4 1.00	12 0.92	1 1.00	5 1.00			685 0.90	
L	11 1.00	35 1.00	2 1.00	1 1.00	10 1.00	31 0.94	18 1.00	4 1.00	5 0.20	3 0.33	128 0.98	
V	8 1.00	40 1.00	3 1.00	4 1.00	9 1.00	26 0.96	14 0.93	13 1.00	4 1.00	4 0.00	11 1.00	837 0.90
A	10 1.00	42 1.00	9 0.89	847 0.90	7 1.00	27 0.96	21 1.00	9 1.00	809 0.95	4 0.50	8 0.88	15 1.00
G	12 1.00	41 1.00	7 0.86	2 1.00	10 1.00	31 0.97	21 0.95	18 0.94	5 0.20	9 0.44	6 0.50	16 1.00
C	2 0.50	4 0.75	1 1.00		2 1.00	16 0.88	5 1.00	5 1.00	4 0.50	2 0.50	1 1.00	
S	10 1.00	486 0.85	2 1.00	4 1.00	12 1.00	31 0.97	11 1.00	9 0.89	9 0.56	8 0.63	4 0.75	1 1.00
T	10 1.00	32 1.00	834 0.90	3 1.00	8 1.00	489 0.85	14 0.93	9 1.00	5 0.20	3 0.33	7 1.00	1 1.00
N	767 0.89	15 1.00	1 1.00		1 1.00	14 1.00		738 0.88	2 0.00	806 0.95	1 1.00	
Q	4 1.00	7 1.00			5 1.00	16 1.00	6 1.00	1 1.00		4 0.50	4 1.00	
D	2 1.00	21 0.95	1 1.00		4 1.00	19 1.00	4 1.00	4 1.00		5 0.00		
E	3 1.00	9 1.00	1 1.00		762 0.89	21 1.00	7 1.00	5 1.00	6 0.00	3 0.00	2 0.50	
H	2 1.00	31 1.00	2 1.00		6 1.00	16 0.88	2 1.00	6 1.00	1 0.00	3 0.67	1 1.00	
K	2 0.50	3 0.67			3 1.00	15 1.00	3 1.00	1 1.00	3 0.00	3 0.00		
R	8 0.88	21 0.62	3 1.00	2 0.00	8 0.50	29 0.93	21 0.81	15 1.00	12 0.00	5 0.00	5 0.00	1 0.00

Figure 4.7. Average active fraction for a sequence based on the amino acid at the specified position. For each position of interest, the amino acid at that site is related to the active fraction of sequences having that particular mutation. The total number of sequences considered in the calculation of the active fraction value is given in the top-right corner of each cell. Note that this is based on sequence data for single, double, and triple mutants across libraries mutating redundant positions, and is therefore not a canonical heat map of independent mutation outcomes.

4.3.4 Increasing the diversity of protein sequences used in the MSA improves the binary classification ability of the statistical fitness property

The halo-forming or non-halo-forming designation of each sequence was compared to the predicted statistical fitness calculated using different sets of sequence inputs in the starting MSA and assessed via a receiver operating characteristic (ROC) curve.

All six ROC curves for the initial homology sequence sets yielded area under the ROC curve (AUC) values between 0.840 (Firmicutes$_{\text{stringent-2k}}$) and 0.894 (All$_{\text{lax-29k}}$), demonstrating the ability of the statistical fitness to discriminate between active and inactive variants (Figure 4.8). Moreover, relaxing the restrictions placed on the protein sequences in the starting MSA, in terms of both acceptable sequence length (stringent → lax cut-off) and acceptable taxonomy (Firmicutes → Bacteria → All), was shown to consistently improve the AUC and thus increase the reliability of the predictive model (Figure 4.9.A). This agrees with previous findings by Hopf et al., who found that progressively excluding evolutionarily distant sequences led to poorer PLMC predictive performance across 34 sets of data, 21 of which involved proteins[128].

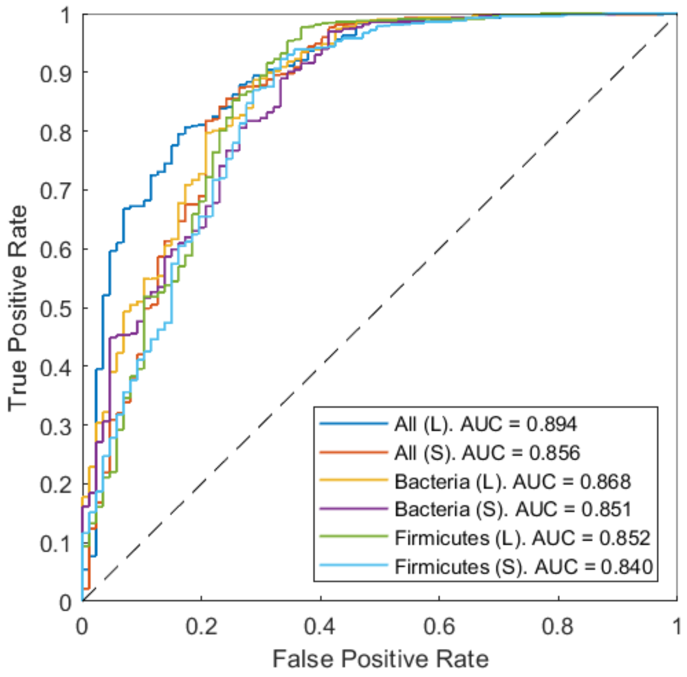


Figure 4.8. ROC curves demonstrating that the statistical fitness is predictive of variant activity. The experimentally-observed outcome (retention or ablation of catalytic activity) for each qualifying variant was evaluated against the variant's statistical fitness, calculated using one of the six sets of protein sequences in the starting MSA (Table 9). The AUC consistently improved when restrictions placed on the protein sequences used in the MSA, in terms of both allotted sequence length and taxonomy, were relaxed. The best predictive method employed the least constrained MSA containing the most sequence information.

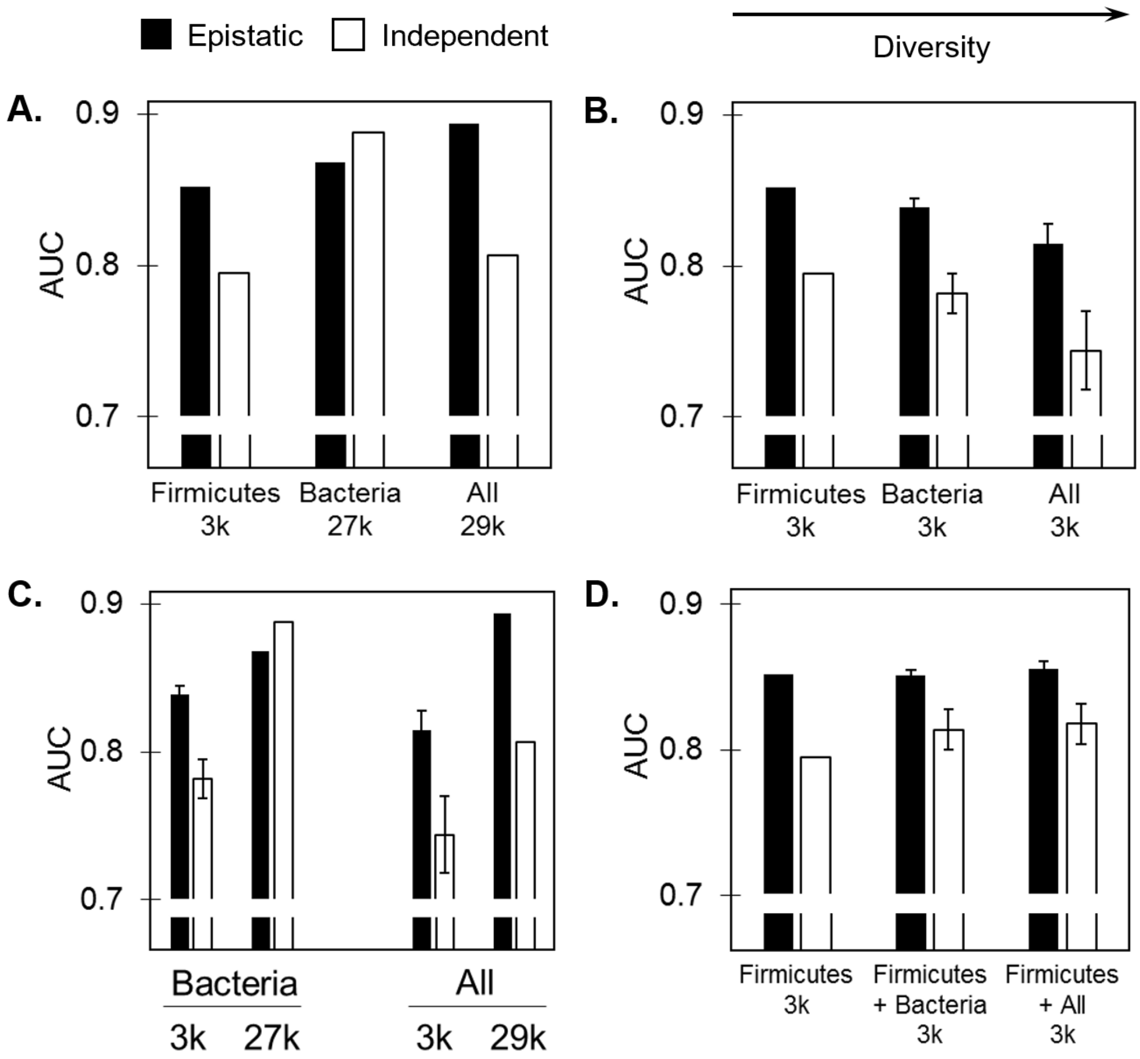


Figure 4.9. Effects of varying sequence diversity and depth on the predictive performance of the statistical fitness. (A – D) Predictive performance of the statistical fitness when different groups of homologous sequences are used in the starting MSA (summarizing key results from Table 9, with the average AUC given for designations consisting of twenty sub-sampled groups containing 3k sequences, and error bars representing the standard error). (A, B) Comparison of different diversity sources including (A) every available sequence or (B) only 3,037 sequences within each category. (C) Comparison of performance at different sequence depth for *Bacteria* and *All.* (D) Comparison of performance for 1.5k *Firmicute* sequences plus 1.5k additional sequences from *Firmicutes*, *Bacteria*, or *All*. All data are for the *lax* length threshold. Black bars result from the epistatic model. White bars result from the independent model.

Including increasingly evolutionarily-distant sequences in the starting MSA also increased the number of sequences under consideration. To decouple the individual effects that the evolutionary distance and sequencing depth have on predictive performance, the groups All$_{\text{lax-29k}}$ and Bacteria$_{\text{lax-27k}}$ were randomly sampled twenty times each to create subgroups containing the same number of sequences as in Firmicutes$_{\text{lax-3k}}$ (3,037). All subgroups were independently aligned with PROMALS3D, and PLMC was used to generate a Potts model for each. Additionally, epistatic coupling was considered in the model as before, or toggled off by omitting pairwise contributions during model inference. An AUC value for the ROC curve relating the statistical fitness to the experimental results was calculated for each of these subgroups (Table 9).

When the sequencing depth was fixed, including sequences closest to the WT phylogenetically, i.e. less diverse, led to the best predictive performance. This was true of both the epistatic and independent models (Figure 4.9.B). Conversely, when the acceptable diversity was fixed, including more sequences in the starting MSA improved the predictive performance of both models (Figure 4.9.C).

Notably, 91% of the 2.9 x 10^4 sequences in the All$_{\text{lax-29k}}$ group are bacteria, and 89% of these bacterial sequences are non-firmicutes. For the epistatic model, supplementing the 3,037 firmicute-only sequences with 2.4 x 10^4 additional non-firmicute bacterial sequences (to yield the group Bacteria$_{\text{lax-27k}}$) led to a +0.016 improvement in AUC. Supplementing further with only 2,548 non-bacterial sequences (to yield the group All$_{\text{lax-29k}}$) led to a further +0.026 improvement in AUC. The same was not true in the independent model: supplementing the group Firmicutes$_{\text{lax-3k}}$ to yield Bacteria$_{\text{lax-27k}}$ improved the AUC from 0.795 to 0.888 (+0.093), but further supplementing the group Bacteria$_{\text{lax-27k}}$ to yield All$_{\text{lax-29k}}$ led to a decrease in the AUC from 0.888 to 0.807 (-0.081). These results suggest that epistasis must be considered in order for highly diverse sequences to be beneficial and improve predictive performance, otherwise, they can have a negative impact if incorporated into the starting MSA. The group Firmicutes$_{\text{lax-3k}}$ was additionally supplemented with 2,548 non-bacterial sequences and the inclusion of the non-firmicute bacterial sequences was circumvented entirely. As compared to the group Bacteria$_{\text{lax-27k}}$, this improved the AUC slightly, from 0.852 to 0.865 (+0.013) in the epistatic case and from 0.795 to 0.799 (+0.004) in the independent case. This suggests that divergent sequences outside of bacteria offer predominantly sparse information, and their contribution is significant only when pairwise information is considered.

Thus, although using a small set of sequences that were phylogenetically similar to the WT led to a superior predictive performance as compared to using a small set of diverse sequences (AUC Fimicutes$_{\text{lax-3k}}$ > AUC Bacteria$_{\text{lax-3k}}$ > AUC All$_{\text{lax-3k}}$), once the MSA was seeded with a collection of close homologs, including more diverse sequences further improved the predictive performance of the epistatic model (AUC All$_{\text{lax-29k}}$ > AUC Bacteria$_{\text{lax-27k}}$). To further evaluate the relative benefits of appending sequences with different diversities, a set of 1.5 x 10^3 firmicute sequences was supplemented with 1.5 x 10^3 sequences either subsampled from All$_{\text{lax-29k}}$ (to yield the subgroup Firmicutes+All$_{\text{lax-3k}}$) or Bacteria$_{\text{lax-27k}}$ (to yield Firmicutes+Bacteria$_{\text{lax-3k}}$), once again resulting in a total of 3,037 sequences per subgroup. The performance of both subgroups was compared to Firmicutes$_{\text{lax-3k}}$. For the epistatic model, the predictive performance of each group was similar. For the independent model, including more diverse sequences was slightly more advantageous as compared to only including more firmicute sequences (Figure 4.9.D).

4.3.5 A mid-throughput binary screen, coupled with a computationally-informed library design, resulted in the efficient isolation of lysin mutants with improved specific activity and/or thermal stability

Ten variants (Table 12) belonging to libraries 1 – 8 were randomly selected during the halo plate assay to more sensitively quantify their specific activity. Nine (variants 1 – 7, 9, and 10) were halo-forming and one (variant 8) was not. Variant 2 (R17L/T40L) had an off-target mutation that was not found in any libraries and was therefore not pursued further. When expressed, sufficient amounts of protein were able to be recovered for eight of the nine remaining variants, with final purities ranging from 90% – 99% (median: 97%) (Figure 4.17). Variant 3 (M45E/I87D, Library 3) was unable to be produced in sufficient quantities and was excluded from further testing. Activity for the remaining eight variants and the WT was assayed using the turbidity reduction method. 0.1 µg, 0.3 µg, or 0.5 µg of each variant was co-incubated with crude VRE cell wall material. The OD600 of each sample was monitored over the course of four hours (Figure 4.10). The largest slope over a fixed timeframe was calculated for each replicate as a proxy for specific catalytic activity (Figure 4.11). Across all starting conditions, the WT exhibited a normalized change in OD600 of 0.041 ± 0.003 OD600/min/µg (n = 16). Four of the eight variants (4, 5, 6, and 10) exhibited activities that were moderately diminished as compared to the WT. Two variants (1 and 9) exhibited activity that was statistically indistinguishable from the WT. Variant 7 exhibited a markedly improved activity of 0.09 ± 0.01 OD600/min/µg ($p < 0.001$). Lastly, variant 8, which was non-halo-forming, had an activity of 0.003 ± 0.002 OD600/min/µg (n = 16), which was statistically indistinguishable from the performance of the buffer negative control, 0.003 ± 0.003 (n = 24).

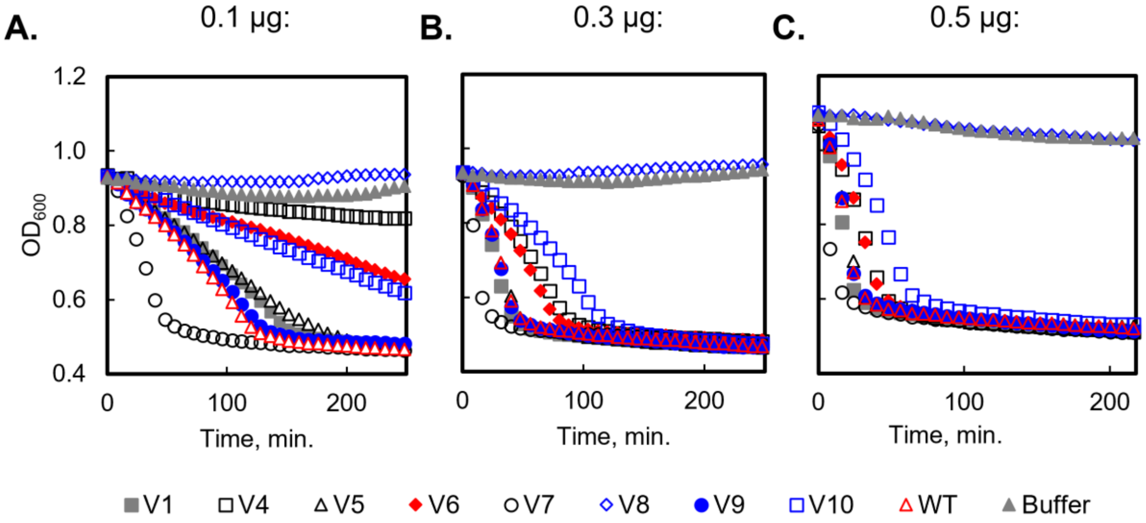


Figure 4.10. LysEFm5 variant and WT activity against VRE cell wall fragments. (A) 0.1 µg (n = 4), (B) 0.3 µg (n = 4) or (C) 0.5 µg (n = 8) of each variant was combined with VRE cell wall fragments in 200 µL of PBS (total). OD_{600} was monitored over time, with data collected every two minutes. Every third data point collected is shown here. Error bars are not shown for visual clarity; a measure of the uncertainty between replicates is given in Figure 4.11 as the standard deviations in the slope of the linear regions.

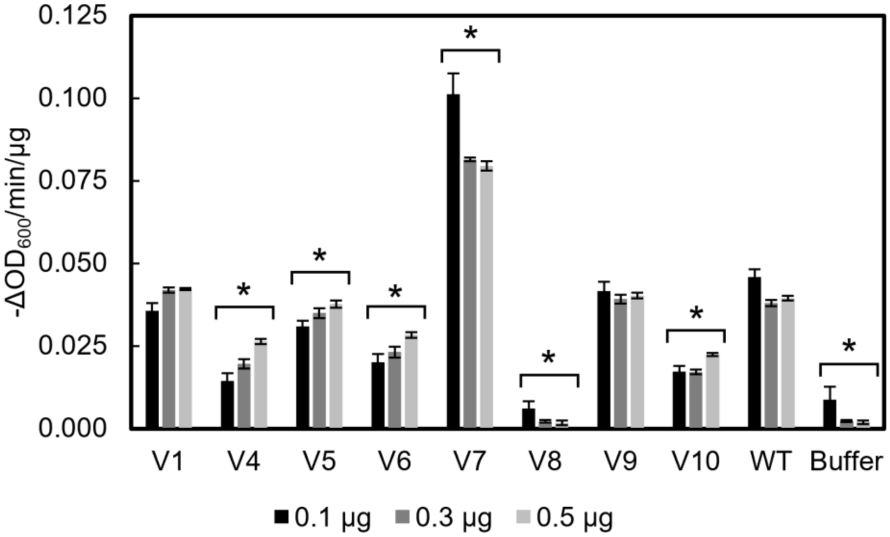
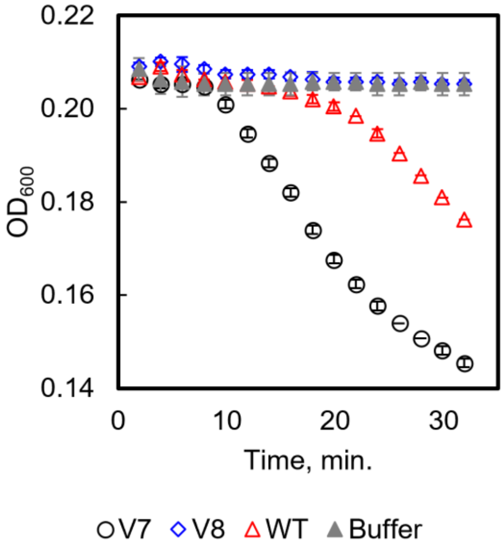


Figure 4.11. Quantified activity for variants, WT, and the buffer negative control. The maximum change in OD_{600} over a 20 minute period (alternatively, 10 minutes for variant 7 at the 0.3 and 0.5 µg conditions owing to rapid kinetics) was calculated from the turbidity assay results (Fig. 10) for each replicate. *p < 0.005 (compared to the WT) for a two-tailed, two-sample heteroscedastic Student’s t-test with a Bonferroni correction applied (η = 10).

The activity of variant 8 was additionally statistically insignificant from the buffer negative control (p > 0.05).

Additionally, variant 7 (most active), variant 8 (least active), WT, and buffer were tested against live 8-E9 VRE in exponential phase to evaluate if the turbidity reduction assay, which used purified cell wall material, correlated to the killing of live cells. The reduction in OD600 of cultured VRE cells was found to agree well with the results of the turbidity reduction assay (Figure 4.12).



	Viable Cells (10^6 CFU)
Condition	**Mean CFU ± SE**
V7	0.92 ± 0.05
V8	6.3 ± 1.0
WT	4.2 ± 0.2
Buffer	5.1 ± 0.7

Figure 4.12. Bacteriolytic activity of variants 7 and 8, the WT, and buffer against live 8-E9 VRE. 0.5 µg of lysin was applied to mid-exponential phase *Enterococcus faecium* 8-E9 resuspended in PBS. Cell lysis was monitored dynamically via OD_{600} reduction (left). Killing activity was assessed by plating serially diluted cell suspensions after approximately 30 minutes (right). Data are presented as mean ± standard error.

To determine whether the perceived change in catalytic activity of any of the variants could be attributed in part to a change in the marginal thermal stability of the WT at assay conditions, the stability of variants 4 – 10 and the WT was assessed by Sypro Orange thermal denaturation assay (Figure 4.13). Variant 1 was unable to be produced in sufficient quantities to use in this assay and was not tested. The melting temperature, Tm, of the WT was 43.4 ± 0.5 °C. Variant 4 did not exhibit a signal consistent with unfolding, perhaps resulting from a low Tm or increased disorder in the molecule. Variants 9 and 10 exhibited Tm that were statistically indistinguishable from the WT. Variants 5, 6, 7, and 8 exhibited improved Tm at or above 46.5 ± 0.4 °C, with variant 7 exhibiting the highest value, 54.9 ± 0.6 °C, an 11.5 ± 0.8 °C improvement over the WT. This variant comprises chemically homologous mutations – T34S and A35V – at adjacent sites (Figure 4.4). The retention of amino acid characteristics at these two sites may be key to the observed improvements in both activity and stability.

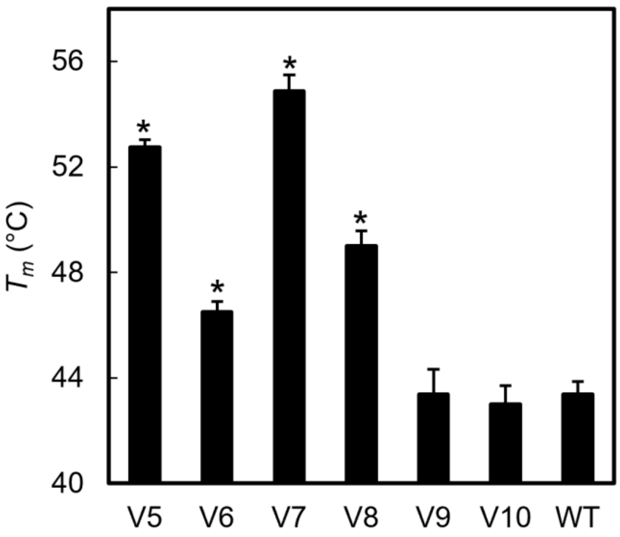


Figure 4.13. Lysin thermal stability. The midpoint of thermal denaturation was measured for lysin variants 5 – 10 and WT by Sypro Orange Thermal Denaturation assay. *p < 0.001 for a two-tailed, two-sample heteroscedastic Student’s t-test.

4.4 Discussion

Improvements to the specific activity of a lysin allow for a lower required dose to achieve the same therapeutic effect, potentially reducing dose-related toxicity and mitigating the immune response to lysin-specific antibodies produced upon administration of the lysin in vivo[181]. Improvements in lysin stability allow for more flexibility in the protein production process, a longer shelf-life, and a reduction in the tendency to unfold thereby reducing aggregate formation[182]. Improvements in one or both characteristics can contribute to heightened bioavailability in an infected host, which can increase treatment efficacy. Of the seven randomly-selected, halo-forming lysin variants that were assayed for catalytic activity, three exhibited activity that was indistinguishable from the WT (0.041 ± 0.003 OD600/min/µg) or improved. Five of the six variants assayed for thermal stability exhibited a Tm that was indistinguishable from the WT (43.4 ± 0.5 °C) or improved. Variant 7 in particular demonstrated both a considerably higher catalytic activity of 0.09 ± 0.01 OD600/min/µg and melting temperature of 54.9 ± 0.6 °C. The improved characteristics of this variant suggest that it could be used as a starting point for future LysEFm5 engineering efforts.

The notion that one of the eight randomly selected halo-forming variants tested was able to be produced in sufficient quantities and exhibited improvements in catalytic activity and stability, and that two others exhibited improvements in stability and retained a fraction of the catalytic activity, is a promising result given the limited sampling. Extending the study presented herein, additional clones could be sampled for more sensitive testing to improve the characterization of the PLMC-informed libraries. This extension may reveal variants with physical properties that are further improved in comparison with those described here. Ultimately, this platform may be able to expedite the discovery process by requiring sensitive testing of a focused set of pre-screened variants rather than uninformed libraries orders of magnitude larger in size.

Of the eight double mutant protein libraries that were studied, Library 8, with diversity at the putative primary residue N83 and secondary residue A74, was predicted to be the worst-performing library ($\Delta E = -12 \pm 3$) and demonstrated the lowest experimental rate of activity (30%). The higher rate of activity among single mutants in this library (67%), as compared to double mutants (21%) suggests that the retention of at least one of these two WT residues is necessary for lysin activity. Mutations at either site to the positively-charged, hydrophilic residues arginine and lysine, or to the structurally-disruptive residues proline and tryptophan, resulted in a consistent loss of activity. The remaining seven libraries showed high rates of activity retention (84 – 100%), but no discernable trend in average statistical fitness. Assuming that the results of the halo assay are a monotonic function of the total lysin activity, the relationship between halo-forming variants as a function of the statistical fitness is expected to be sigmoidal in nature[176]. The observed similar fractions of active variants for Libraries 1 – 7 suggests that the WT lies to the far right of this sigmoid (corresponding to high fitness), such that meaningful differences in the fraction of halo-forming variants would only be seen at considerably lower statistical fitness values, such as the value observed for Library 8. Alternatively, or in addition, it is possible that using structural information in combination with double mutant ΔE data to constrain site selection led to the construction of libraries with relatively low penalties of mutation, and subsequently high observed rates of activity retention.

Constraining the diversity of the libraries using the SwiftLib tool, which is used to specify codon selection based on a known metric, generally did not substantially impact the already high rates of activity retention observed for double mutants in Libraries 1, 2, 4, or 5 (94 – 100%). The activity retention of Library 3, however, was improved from 82% to 100%. Thus, constraint is a useful tool to design combinatorial libraries based on the effective accuracy of the statistical fitness parameter (the metric in this case).

Sequencing results of 873 unique variants across all libraries showed that the statistical fitness parameter was predictive of the experimental loss or retention of catalytic activity of LysEFm5. Our findings support the previously-observed notion that including a large set of homologous protein sequences in the starting MSA leads to the best predictive performance of the Potts model[128,183,184]. Because MSAs aim to create aligned sites that represent related positions in the protein structure, and because all sequences to be used are initially restrained by a relatively strict relevance cut-off, including a large number of sequences in the MSA does not “dilute” information as only relevant portions of these sequences end up being considered – provided that epistatic coupling is taken into account. Conversely, if only a relatively small set of sequences (on the order of hundreds to a few thousand) will be considered in the starting MSA, or if epistatic coupling is not considered, then including diverse sequences may worsen predictive performance.

For future protein engineering efforts, utilizing a second-order Potts models to select beneficial sites for mutation can constitute a useful approach, provided that certain criteria are met. The structure of the domain or functional site in the protein of interest must be evolutionarily conserved, allowing for investigation into the dominant sequence constraints acting on familial sequences, and ideally on the order of tens of thousands of homologous protein sequences must be available for use in the starting MSA. As such, this approach is especially well-suited to engineer the catalytic domain of members of the lysin family with N-acetylmuramidase activity, such as LysEFm5, which are known to have distinctly conserved functions and structural features[114].

As the need for alternative or supplemental strategies to treat bacterial infections resistant to conventional antibiotics increases, computationally-informed methodologies such as this that allow for the expedited discovery of antimicrobial proteins with improved properties are of great relevance.

4.5 Materials and Methods

4.5.1 Bacteria used and culture conditions

Escherichia coli cells were grown in either a liquid culture of lysogeny broth (LB), or on a solid LB agar plate containing 1.5 v/v% agar, and supplemented with 50 g/mL kanamycin (kan). All cultures were grown at 37 °C with liquid cultures shaken at 250 rpm, unless otherwise noted. *Enterococcus faecium* 8-E9 cells were grown in a liquid culture of brain heart infusion (BHI) medium at 37 °C with shaking at 250 rpm.

4.5.2 Inputs for PLMC

A search in Jackhmmer was performed to determine significant query matches in the UniProtKB database to the WT amidase domain of LysEFm5 (AA 1-185) (Figure 4.2.D) with the taxonomy restricted to *E. faecium*. The consensus sequence of the top three results was used as the seed sequence in a subsequent search, with the acceptable taxonomy set to either firmicutes only, bacteria only, or all, and the minimum expectation value (E-value) set to 1.0 x 10^{-5}.

A two-component Gaussian mixture model was constructed to describe the distribution of sequence lengths in the Bacteria and All groupings (Figure 4.14). Each sequence length was assigned a membership score for two component curves, one describing the main distribution and the other describing the tail of short, trailing sequences, presumably outliers. The lax cut-off retained sequences with a component 1 (main distribution) membership score ≥ 0.80 and the stringent cut-off required a component 1 membership score ≥ 0.95. The same criteria were applied to generate the range of acceptable sequence lengths for the Bacteria and All groups; differences between the mean and the spread of each data set resulted in different specific bounding lengths. For the Firmicutes group, there existed a clear outlier length (133 AA), likely due to oversampling. Three Gaussian distributions (components 1-3) were fitted such that component 3 represented the set of outliers. A new membership score was calculated for each sequence by weighing scores from component 1 (main distribution) and component 3. These weights were selected such that the outlier length was the lower bound for the lax cut-off. The ranges of acceptable sequence lengths are summarized in Table 9.

PROMALS3D[137] was used to generate an MSA for each of the resulting sets of protein sequences. The PLMC algorithm[128] was run using the recommended regularization parameters for the single-site and pairwise coupling constraints, without the inclusion of gap states. The strength of L2-regularization was set to $\lambda_h = 0.01$ and $\lambda_J = 36.8$, and sequences were re-weighted to account for redundancy based on an 80% sequence identity cut-off, as recommended.

4.5.3 Design of NNK Libraries

The amidase domain of the major pneumococcal autolysin LytA was identified as a homolog of the amidase domain of LysEFm5 initially via sequence alignment. Thirteen primary ligand-binding residues and ten secondary residues were identified in the putative secondary interaction shell of the LytA crystal structure having the inactivating mutations C60A, H133A, and C136A[173]. Twelve structurally-analogous residues (one primary and eleven secondary) that occupied the same space in the 3D structure of the LysEFm5 protein were selected.

The most restrictive MSA (using sequences from group Firmicutesstringent-2k) was used in PLMC to predict changes in the statistical fitness of mutants arising from all possible double mutations at any two of the sites of interest (Figure 4.3). This heat map was used to select eight sets of two discrete sites for NNK library creation (Table 10), sampling a range of average ∆E values.

Library 9 was designed using the SwifLib tool[174]. The matrix of predicted change in statistical fitness was discretized into an integer matrix using the following criteria: $-10 < \Delta E < -8$, $f(\Delta E) = -5$; $-3 < \Delta E < 0$ $f(\Delta E) = 1$; $\Delta E > 1$, $f(\Delta E) = 10$; $-8 < \Delta E <-6$, $f(\Delta E) = -2$; $\Delta E == 0$, $f(\Delta E) = 2$; $-6 < \Delta E < -3$, $f(\Delta E) = 0$; $0 < \Delta E < 1$, $f(\Delta E) = 5$.

The resulting library, which had a specified maximum size of 252, showed diversity at three positions: 33, 40, and 87. The remaining eight positions of interest encoded the WT residues.

4.5.4 Plasmid creation

A gBlock gene fragment (Integrated DNA Technologies) encoding the entire 341 AA LysEFm5 gene[156] was amplified via Q5 PCR (New England Biolabs, NEB), visualized on an agarose gel, and purified. The product was Gibson assembled[175] (NEBuilder® HiFi DNA Assembly, NEB) into a pET-24 vector modified to include a C-terminal 6xHIS tag[39] containing a selection marker for kanamycin, which was digested with BamHI and NdeI (NEB). The assembled product was transformed into NEB® 5-alpha competent *E. coli* (NEB), which were cultured at 37°C on a lysogeny broth (LB) agar plate with 50 µg/mL kan for selection overnight. A flask was inoculated with cells from a colony on the plate that encoded the LysEFm5 gene and was left to incubate at 37°C overnight. Plasmid DNA was then isolated from the culture.

4.5.5 Construction of NNK Libraries

The nine NNK libraries were each created by assembling two or three overlap extension PCR products. For each library, the universal forward primer (PRIM01) was used in conjunction with a second primer encoding the randomized codon(s) at one or more of the desired sites. The second primer was designed to anneal to the site immediately downstream of the second codon (in the case of Libraries 1, 2, 4, 5, 7, and 8) or the first codon (in the case of Libraries 3, 6, and 9). In the second reaction, a reverse primer annealing to the site immediately upstream of the first or second randomized codon, respectively, and the universal reverse primer (PRIM02) were used. For Libraries 3, 6, and 9, a third reaction producing a fragment containing the two randomized codons was required as the distance between the diversified codons did not allow for all to be reasonably encoded on the same primer. Summaries of the reactions performed and primer identities are given in Table 13 and Table 14. The expected size of each PCR product was confirmed by running a DNA gel and the bands were gel purified to yield the final DNA fragments.

Each of the two or three PCR products per library were independently combined using Gibson assembly (NEB). Fragments were combined into a pET-24 vector digested with BamHI and NdeI (NEB). Each reaction was cleaned up using a PCR cleanup kit (Epoch Life Science), then transformed into 25 µL of MC1061F- electrocompetent cells (Lugicen) quenched with 975 µL of Recovery Medium. 950 µL of the transformation product was used to inoculate a 3 mL culture of LB + kan, which was left to grow at 37 °C and 250 rpm overnight. The remaining 50 µL of transformation product was plated onto an LB + kan agar selection plate. Transformation efficiency was confirmed to be on the order of tens of thousands of transformants for each library. Two to three monoclonal transformants were additionally outgrown and sequenced from each library to confirm the expected diversity.

4.5.6 Halo Plate Creation

100 mL of BHI broth was inoculated with VRE and grown overnight at 250 rpm and 37 °C. The flask was autoclaved and centrifuged, and the VRE was washed repeatedly with phosphate-buffered saline (PBS). The final mass concentration of this stock was approximately 0.3 g of cell material per mL of PBS. The stock was stored at 4 °C until future use.

The effect of the concentration of IPTG, % agar, and time on halo radius were investigated in a separate experiment prior to conducting the halo plate assay (Figure 4.16). All conditions tested were sufficient to produce results that allowed for the easy identification of halo-forming colonies. It was determined that an IPTG concentration of 0.05 mM, 1.0 w/v% agar, and an incubation time of 19 hr were appropriate. Additionally, a saturated culture of *E. coli* expressing a phage lysin with specific activity against Clostridium perfringens[176] was plated alongside a positive control. At no time did halos form on the plate containing the C. perfringens-specific lysin (Figure 4.15).

To create each LB + kan/VRE/IPTG plate used in the halo plate assay, 2.0 w/v% LB and 1.0 w/v% agar were combined in an Erlenmeyer flask along with enough deionized water to constitute 15 mL of total volume per plate, and the solution was microwaved until it boiled. 50 µg/mL of kan, 0.05 v/v% of the stock autoclaved VRE, 0.05 mM IPTG (final concentrations) were added to the solution after boiling.

4.5.7 Halo Plate Assay

For each library, 0.5 - 2 µL of plasmid DNA isolated from a saturated culture of MC1061F- electrocompetent cells was transformed into 25 µL of T7 Express lysY/Iq Competent *E. coli* (NEB). 975 µL of LB + kan media were added, and each transformation product underwent a 1 hr outgrowth at 37 °C and 250 rpm. Afterwards, the cells were spun down, re-suspended in 100 µL of LB + kan, and plated onto an LB + kan selection plate. Plates were incubated overnight at 37 °C. Between 16 and 18 hr later, the lawn of cells on each plate was re-suspended in approximately 10 mL LB + kan, and the cell density was estimated using absorbance measurements taken using a microplate reader (averaged for 1:10, 1:50, and 1:100 dilutions), and an empirically-determined coefficient (7.90 x 10^8 colony forming units/OD600/mL). This was used to determine the serial dilution scheme needed to obtain 125 colony forming (CFU) units/50 µL of cell material. 50 µL of cell material for each library was then plated onto LB + kan/VRE/IPTG agar plates (nine each; three plates per each triplicate). Each plate was incubated for 19 hr at 37 °C. This procedure was performed for two or three libraries (eighteen or twenty-seven plates) at a time.

Following incubation, all colonies belonging to one library were designated as halo-forming if there was visible clearance around the colony, or non-halo-forming if there was not, then systematically plucked and placed into one of six, library-specific bins based on the replicate number (1 – 3) and halo-forming designation. Cell material from each bin was stored at -20 °C for between one to four days. Afterwards, 500 µL of MX1 re-suspension buffer (Epoch Life Science) was added to each of the sixty samples. 20 µL aliquots from each sample belonging to the same replicate number and halo-forming designation were pooled across all nine libraries. Plasmid DNA was extracted from each of the six resulting 200 µL samples.

4.5.8 High-throughput sequencing (Illumina MiSeq) sample preparation

Both sets of five forward (FA1-5) and five reverse (RA1-5) primers were independently premixed at equal ratios (primer sequence identities are given in Table S4). A 50 µL PCR was performed with a final FA1-5 and RA1-5 primer concentration of 500 uM and 2 µL of plasmid DNA. 4 U of exonuclease I (ExoI) was then added to catalyze the degeneration of single-stranded DNA. The samples were incubated at 37°C for 30 minutes, then heat inactivated at 80 °C for 20 minutes. Afterwards, 1 µL of the ExoI-digested product was used as a template in a second, 50 µL PCR with a final FB and RB1-6 sequencing primer concentration of 500 µM. Each product was run on a gel to confirm that it was the expected size. Bands were gel-purified and the concentration of DNA from each was measured using a NanoDropTM Spectrophotometer (Thermo Fisher). The amount of contributing DNA from groups 1-6 was weighted based on the estimated diversity of the group and combined to yield a total of 500 ng of DNA in 100 µL of nuclease-free water. The sample was submitted for a MiSeq 2X300 bp paired-end-read run with version 3 chemistry.

4.5.9 Production of variants

During the agar plate assay, one halo-forming variant each was isolated from libraries 2, 3, 4, 5, 6, and 7. Two halo-forming variants were isolated from library 1 and one additional non-halo-forming variant was isolated from library 8. All variants were confirmed to encode the full LysEFm5 protein, with diversity at the expected sites.

For each clone, a cell culture tube containing 3 mL of LB + kan was inoculated with cells then incubated at 37 °C and 250 rpm overnight. The day after, 100 mL of LB was inoculated with 100 µL of confluent culture. The OD600 was monitored using a plate reader spectrophotometer until it was within the range of 0.6 – 0.8, at which point IPTG was added at a final concentration of 0.5 mM and the culture was left to incubate at 30°C and 250 rpm for six hours. The culture was then spun down, the supernatant was discarded, and 1 mL of lysis buffer (137 mM NaCl, 2.7 mM KCl, 8 mM Na2HPO4, 2mM PBS, 5% glycerol, 3.1 g/L CHAPS, 1.7 g/L imidazole, with a Pierce™ Protease Inhibitor Mini Tablet, EDTA-free (1 tablet per 10 mL buffer)) was added. Each culture was then supplemented with MgSO4 to a final concentration of 20 mM as well as 2 U of DNase I (New England Biolabs) and 10 µg of RNase A (Thermo Scientific). The cell pellet underwent four freeze/thaw cycles at -80 °C/room temperature, respectively. The cell material was then spun down, and the supernatant was filtered and diluted with 1 volume of wash buffer (50 mM sodium phosphate, 300 mM NaCl, 10 mM imidazole, 5% glycerol), applied to 200 µL of HisPur cobalt resin (Thermo Scientific), and rotated end-over-end at room temperature for 30 minutes. This mixture was then applied progressively to spin columns. Three applications of wash buffer were performed followed by three elutions (with 50 mM sodium phosphate, 300 mM NaCl, 150 mM imidazole, 5% glycerol) to constitute the protein sample, in a volume of ~ 1200 µL. Proteins were further purified by application to an ÄKTAprime plus configured with a Superdex 75 Increase 10/300 GL size exclusion column. Samples were run at 0.2 mL/min with PBS + 5% glycerol as eluent. Appropriate fractions were collected, mixed, and divided into 100 µL aliquots which were snap frozen. All subsequent analysis was performed on aliquots thawed on ice immediately before use.

4.5.10 Quantification of variant and WT concentrations

SDS-PAGE was performed to quantify the produced protein concentration of each variant and the WT. 50 µg/mL of bovine serum albumin (BSA) was used as a standard. 12 µL of each variant and the WT, in addition to the BSA standard, was combined with 4 µL of 4X LDS buffer then denatured at 90 °C for 12 minutes. The sample were loaded onto a NuPAGE Bis-Tris 4-12% Protein Gel (Thermo Fisher) along with a PageRuler Unstained Protein Ladder (Thermo Fisher). The gel was run at 200 V for 50 minutes, then stained with SimplyBlue SafeStain (Thermo Fisher). ImageJ was used to determine the intensity of each band corresponding to the BSA standard and protein variants (having an expected molecular weight of ~37 kDa). The relative intensity of the BSA standard was used to determine the unknown variant concentrations.

4.5.11 SYPRO Orange Thermal Denaturation Assay

Variants were diluted to a concentration of 5 µM, and 45 µL was aliquoted into optically clear PCR tubes. The stock solution of Sypro Orange (Thermo Fisher) was diluted to 200x in PBS, 5 µL of which was added to each PCR tube. These solutions were heated from 25 °C to 98 °C in 0.5 °C increments with equilibration time set to 30 seconds after each temperature elevation in a CFX Connect Real-Time PCR Detection System. The fluorescence of the Sypro Orange dye was detected via 450-490 nm excitation and 560-590 nm emission. The maximum change of fluorescence with temperature (defined to be the T_m) was determined via smoothing with local second-degree polynomials having widths of 2.5 °C using the Savitzsky-Golay filter of the sklearn package in Python.

4.5.12 Quantification of variant and WT activity

One hundred mL of BHI was inoculated with a 1000x dilution of VRE grown overnight at 37 °C with agitation. When this culture reached an OD600 of ~ 0.5, it was placed on ice and chilled for 15 minutes. Cells were then pelleted via centrifugation at 6000 × g for 5 minutes. The supernatant was removed and re-suspended in 1 mL of 50 mM Tris-HCl, then added drop-wise to 20 mL of boiling 5% w/v sodium dodecyl sulfate. This solution was boiled with stirring for 15 minutes, then allowed to cool to room temperature and centrifuged at 6000 × g for 5 minutes. The pellet was re-suspended in 1 mL of 1 M NaCl and centrifuged at 17,000 × g for an additional 5 minutes. This was repeated an additional time with 1 mL of 1 M NaCl, then seven times with pure water, and the pellet was finally re-suspended in PBS and stored at 4 °C until use and hereafter referred to as “crude cell wall”.

In a 96-well plate, 0.5 µg of each variant or blank in 5 µL of PBS with 5% glycerol was combined with 195 µL of crude cell wall diluted to approximately an OD600 of 1. Each sample was tested with replication of 4 – 8 with randomized well positions. Measurements of the Abs600 were taken every 2 minutes for multiple hours.

4.5.13 Assessment of cell lysis and killing activity

Enterococcus faecium 8-E9 was streaked onto BHI agar plates and grown overnight at 37 °C. The following morning, a colony was used to inoculate 3 mL of BHI and was incubated at 37 °C with shaking at 250 rpm. When the culture reached mid-exponential phase (OD600 ~ 0.8), cells were washed 2x with sterile PBS with centrifugation of 3000 × g for 3 minutes. Cells were then diluted into 3 mL PBS and 195 µL was applied to 5 µL of 0.5 µg of purified proteins in PBS + 5% glycerol in a 96 well plate. The plate was then incubated at 37 °C with shaking in a spectrophotometer with an OD600 measurement taken every 2 minutes. After 30 minutes, the plate was removed, and cell suspensions serially diluted into BHI. CFU were then determined by enumeration of colonies after plating of dilution series onto BHI agar.

4.5.14 Acknowledgements

This work was supported by a grant from the National Institutes of Health (R01 GM121777).

4.6 Supporting Information

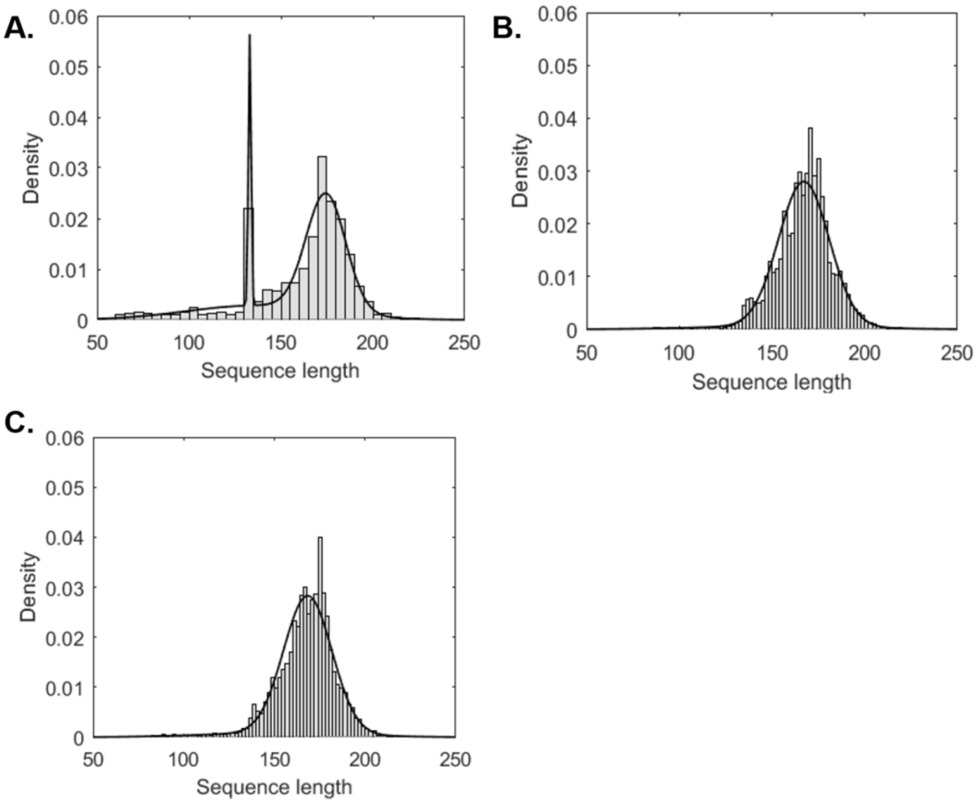


Figure 4.14. Modeling sequence length distributions to identify outliers. Two- or three-component Gaussian mixture models (GMMs) fitted to the distribution of sequence lengths resulting from a jackhmmer search of the UniProtKB database for homologs to the wild-type amidase domain of LysEFm5. (A) The taxonomy in the search was restricted to

firmicutes only (three-component GMM). (B) The taxonomy in the search was restricted to bacteria only (two-component GMM). (C) The taxonomy in the search was not restricted (two-component GMM).

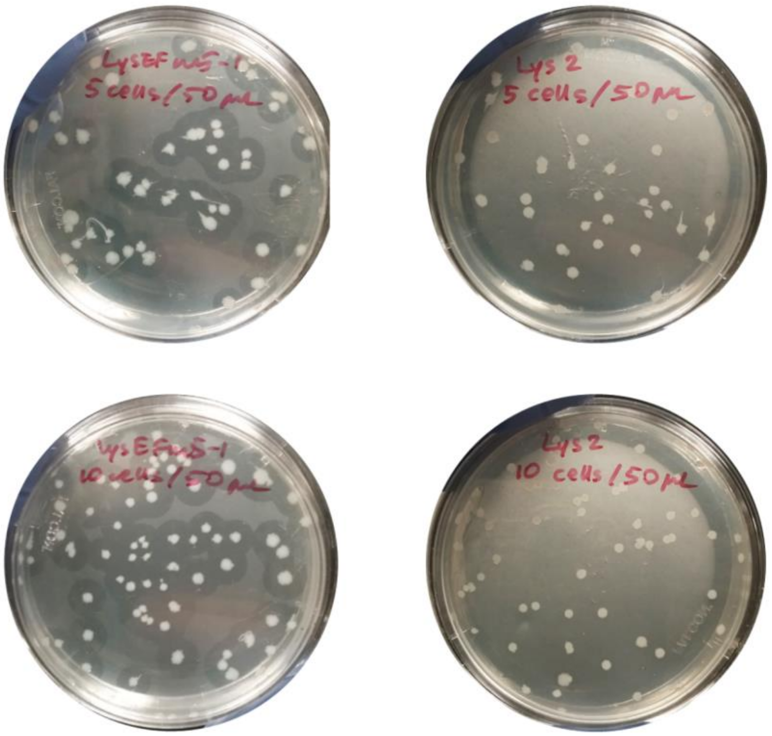


Figure 4.15. E. coli expressing a lysin with alternative specificity do not form halos on LB+kan/VRE/IPTG plates. E.coli containing a plasmid encoding LysEFm5 (left) and Lys2, a *C. perfringens*-specific lysin (right) were plated in an identical manner on LB+kan/VRE/IPTG plates at two densities (top: 0.1 cells/µL; bottom: 0.2 cells/µL). After 19 - 20 hr of incubation at 37 °C, halos were observed on the LysEFm5 plates, but not on the Lys2 plates.

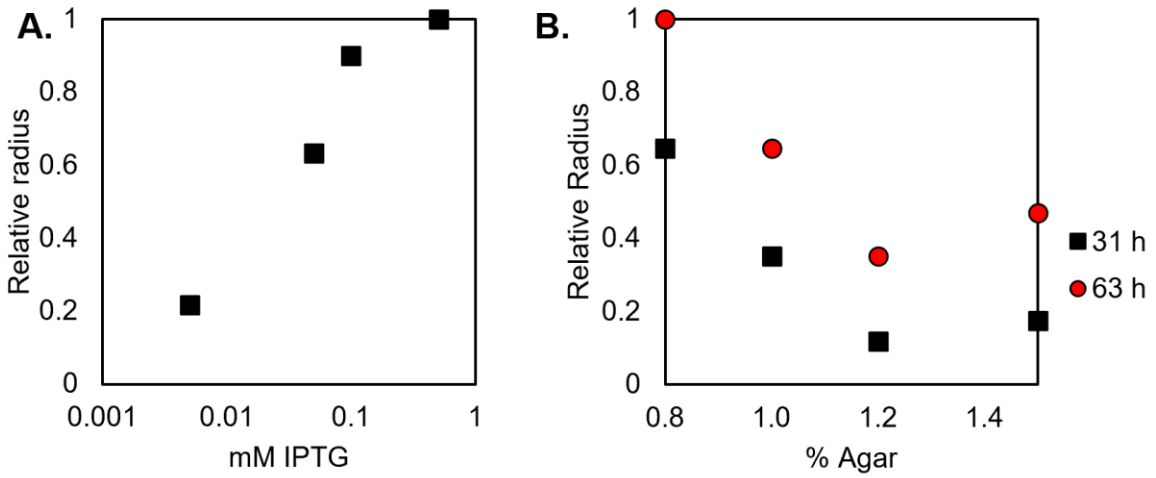


Figure 4.16. Effect of IPTG, agar content, and time on halo radius. Expression-competent *E. coli* containing pET:LysEFm5 were plated on LB + kan agar plates made with 2 w/v% LB, 50 µg/mL kan, 0.05 v/v% of stock autoclaved VRE (~0.3 g/mL cell material suspended in PBS), deionized water, and (A) 0.005, 0.05, 0.1, and 0.5 mM IPTG (and 1.0 w/v% agar) or (B) 0.8, 1.0, 1.2, and 1.5 w/v% agar (and 0.05 mM IPTG) in 15 mL of total volume. Measurements were taken 19 hr after the start of incubation at 37°C for the IPTG-variable plates, and 31 and 63 hr after the start of incubation at 37°C for the agar percentage-variable plates. The radius of the halo around each colony was measured and standardized according to the largest radius measured in that group. For both groups, *n* = 1, as this experiment was conducted only to determine an appropriate set of assay conditions. A direct relationship between IPTG concentration/time and halo size, and an inverse relationship between agar w/v% and halo size was observed. It was determined that a final IPTG concentration of 0.05 mM, 1.0 w.v% agar, and an incubation time of 19 hr or greater was sufficient for the purposes of this assay.

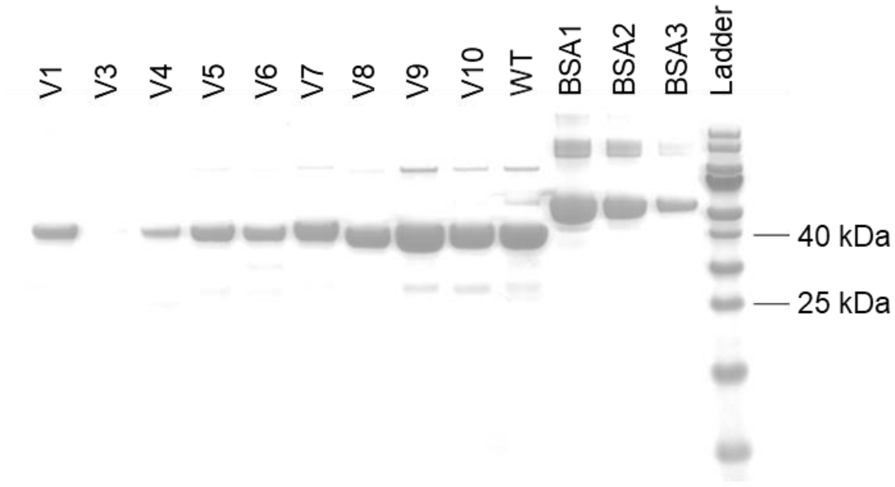


Figure 4.17. SDS-PAGE used to determine variant concentrations. The bands of variants at ~37 kDa correspond to the expected size of LysEFm5. Purities were assessed to be: V1 (99%); V3 (N/A); V4 (97%); V5 (97%); V6 (96%); V7 (98%); V8 (99%); V9 (90%); V10 (93%); WT (91%).

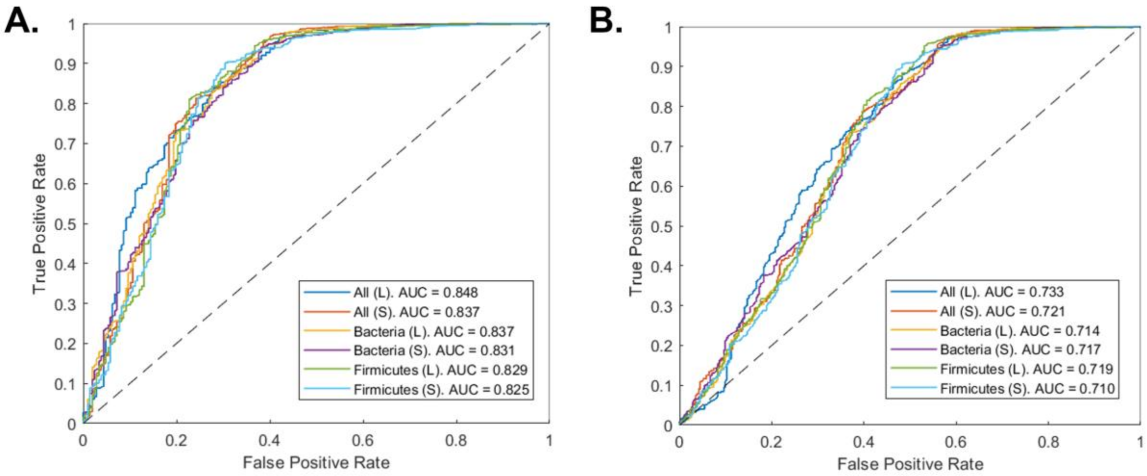


Figure 4.18. Excluding sequences that lack a majority of inactive or active designations improves the classification accuracy of the statistical fitness. ROC curves that quantify the binary classification ability of the statistical fitness parameter are plotted for each starting MSA (Table 9), provided that all 1731 sequences experimentally read at least 100 times are considered in the analysis (including those with the same number of active and inactive observations). (A) Results when sequences with an active fraction of 0.5 are considered active. The AUC ranges from 0.825 – 0.848, depending on the starting MSA. (B) Results when sequences with an active fraction of 0.5 are considered inactive. The AUC ranges from 0.710 – 0.733. Because on the order of hundreds, and up to a thousand, individual colonies were plucked per library, it was expected that there would be contamination from neighboring cell material some fraction of the time. Therefore, observing the same number of active and inactive observations was ultimately attributed to the contamination of one or more bins as a result of human error, leading to the dismissal of these sequences from the analysis. When sequences with an active fraction of 0.5 are excluded entirely, the AUC ranges from 0.840 – 0.894 (Figure 4.8).

Table 12. Additional information for purified variants and the WT.

Variant	Library	Mutations	-ΔOD$_{600}$/min/μg
1	2	T40P, N47V	0.041 ± 0.003
2	N/A	R17L, T40L	N/A
3	3	M45E, I87D	N/A
4	4	N32G, E38T	0.022 ± 0.005
5	5	S33P, M45W	0.035 ± 0.003
6	6	N47A, V91P	0.025 ± 0.004
7	7	T34S, A35V	0.09 ± 0.01
8	8	A74R, N83M	0.003 ± 0.002
9	1	T40A	0.040 ± 0.002
10	1	S33G, T40S	0.020 ± 0.003
WT	N/A	N/A	0.041 ± 0.004

Table 13. List of reactions and primers used to create NNK libraries 1-9.

Lib.	Rxn	Pos. 1	Pos. 2	Pos. 3	Fragment size [bp]	Primer 1	Primer 2
1	1-1	33	40		132	PRIM01	PRIM03
	1-2				936	PRIM12	PRIM02
2	2-1	40	47		153	PRIM01	PRIM04
	2-2				915	PRIM13	PRIM02
3	3-1	45	87		147	PRIM01	PRIM05
	3-2				150	PRIM21	PRIM24
	3-3				774	PRIM14	PRIM02
4	4-1	32	38		126	PRIM01	PRIM06
	4-2				939	PRIM15	PRIM02
5	5-1	33	45		147	PRIM01	PRIM07
	5-2				936	PRIM16	PRIM02
6	6-1	47	91		153	PRIM01	PRIM08
	6-2				156	PRIM17	PRIM22
	6-3				762	PRIM25	PRIM02
7	7-1	34	35		117	PRIM01	PRIM09
	7-2				933	PRIM18	PRIM02
8	8-1	74	83		261	PRIM01	PRIM10
	8-2				813	PRIM19	PRIM02
9	9-1	33	40	87	132	PRIM01	PRIM11
	9-2				186	PRIM20	PRIM23
	9-3				774	PRIM26	PRIM02

Table 14. NNK primer sequence identities.

Primer Name	Library	Sequence identity (5’ → 3’)[a]
PRIM01	1	AAGAAGGAGATATACATATGGTTGAG
PRIM02	1	CAGTGATGATGGTGATGGTGGCATCCNNNNNNNNNNTTATTAATGGTGGTGATGGTG
PRIM03	1	CGCCGCAAGGCGMNNTGCTTCTTGTTTTGCAGTMNNATTACCCCAAGT
PRIM12	1	ACTTGGGGTAATNNKACTGCAAAACAAGAAGCANNKCGCCTTGCGGCG
PRIM04	2	GGCCAGCTGGTTMNNATTCATCGCCGCAAGGCGMNNTGCTTCTTGTTT
PRIM13	2	AAACAAGAAGCANNKCGCCTTGCGGCGATGAATNNKAACCAGCTGGCC
PRIM05	3	GGTTATTATTMNNCGCCGCAAGG
PRIM14	3	TGGTAATATGAACTATNNKGGATATGAAGTCTGTG
PRIM21	3	CCTTGCGGCGNNKAATAATAACC
PRIM24	3	CACAGACTTCATATCCMNNATAGTTCATATTACCA
PRIM06	4	AAGGCGAGTTGCMNNTTGTTTTGCAGTTGAMNNACCCCAAGTATT
PRIM15	4	AATACTTGGGGTNNKTCAACTGCAAAACAANNKGCAACTCGCCTT
PRIM07	5	GGTTATTATTMNNCGCCGCAAGGCGAGTTGCTTCTTGTTTTGCAGTMNNATTACCCCAAG
PRIM16	5	AATACTTGGGGTNNKTCAACTGCAAAACAANNKGCAACTCGCCTT
PRIM08	6	CAGCTGGTTMNNATTCATCGCC
PRIM17	6	GGCGATGAATNNKAACCAGCTG
PRIM22	6	CGTTGCCACAMNNTTCATATCCG
PRIM25	6	CGGATATGAANNKTGTGGCAACG
PRIM09	7	TGCTTCTTGTTTMNNMNNTGAATTACCCCA
PRIM18	7	TGGGGTAATTCANNKNNKAAACAAGAAGCA
PRIM10	8	GATATAGTTCATMNNACCATCGCCATTGGCAGTGTGCCAMNNACCATTGAACGT
PRIM19	8	ACGTTCAATGGTNNKTGGCACACTGCCAATGGCGATGGTNNKATGAACTATATC
PRIM11	9	CGCCGCAAGGCGSNNTGCTTCTTGTTTTGCAGTGNBATTACCCCAAGT
PRIM20	9	ACTTGGGGTAATVNCACTGCAAAACAAGAAGCANNSCGCCTTGCGGCG
PRIM23	9	GACTTCATATCCTAGATAGTTCATATT
PRIM26	9	AATATGAACTATCTAGGATATGAAGTC

Chapter 5 Concluding Remarks and Future Work

As has been discussed throughout this thesis, the problems created by antibiotic resistance warrant the development and application of toolkits from which novel solutions may emerge. The described engineering of an antimicrobial protein for improved specificity (Chapter 2) and two multi-domain cell-wall-degrading antimicrobial enzymes targeting *C. perfringens* (Chapter 3) and *E. faecium* (Chapter 4) for improved stability and/or activity are examples of how protein engineering can be applied to improve the functionality of antimicrobial proteins.

The success of the engineering efforts of LysCP2 and LysEFm5 were made possible by augmenting library design using second-order Potts models whose parameters were inferred using thousands of homologous sequences of their catalytic domains. Interestingly, post-facto analysis of the predictive performance of the statistical fitness of sequences computed by the inferred models for LysCP2 showed that there may be structural features of different regions of the protein which increase the influence of stability on the statistical fitness. Future studies may explore this relationship across different multi-function proteins, for example enzymes, to disentangle how different structural features may be used to further filter regions of the protein most amenable to mutagenesis through library design guided by statistical fitness.

Where possible, the projects discussed within this thesis have utilized homologous information from a single source - that is - highly related sequences. In addition to the source of the information, all models derived from these data have utilized undirected fully-observable graphs to describe the relationships between different amino acids at different positions[128]. Future efforts will benefit from incorporating additional information into models which utilize not only more complex feature representations but also more prior information in their structures. To elaborate on sources of information, though they were used sparingly in the design of libraries for LysEFm5, homology models of the structures of different related sequences could be of significant benefit. Such models could be used to rationally reduce the number of parameters that require inference based upon features such as proximity[127]. Going beyond sequences within the families of those we wish to engineer, there is a wealth of information within the sequences and structures of proteins which has been gathered globally. Incorporating these data requires more sophisticated modeling frameworks, such as neural networks, but has been demonstrated for both sequence[185] and structural[186] frameworks. As was demonstrated in this work, these a priori models will, with increasing accuracy, drive the generation of designed libraries which cover larger and larger amounts of protein sequence space. This process will then feed back into itself, as analysis of designs, in increasingly informative high-throughput assays, will be incorporated to drive to improve accuracy of predictions across physical properties of interest to the protein engineer.

Chapter 6 References

1. Macfarlane, G. *Alexander Fleming: the man and the myth*. (Harvard University Press, 1984).
2. Schatz, A., Bugie, E. & Waksman, S. A. Streptomycin, a Substance Exhibiting Antibiotic Activity Against Gram-Positive and Gram-Negative Bacteria. *Proc. Soc. Exp. Biol. Med.* **55**, 66–69 (1944).
3. Lewis, K. Platforms for antibiotic discovery. *Nat. Rev. Drug Discov.* **12**, 371–87 (2013).
4. Blair, J. M. A., Webber, M. A., Baylay, A. J., Ogbolu, D. O. & Piddock, L. J. V. Molecular mechanisms of antibiotic resistance. *Nat. Rev. Microbiol.* **13**, 42–51 (2014).
5. O'Neill, J. Antimicrobial Resistance : Tackling a crisis for the health and wealth of nations. *Rev. Antimicrob. Resist.* (2014).
6. Springer, B. *et al.* Mechanisms of Streptomycin Resistance : Selection of Mutations in the 16S rRNA Gene Conferring Resistance. *Antimicrob. Agents Chemother.* **45**, 2877–2884 (2001).
7. Lavigne, J. *et al.* An adaptive response of *Enterobacter aerogenes* to imipenem: regulation of porin balance in clinical isolates. *Int. J. Antimicrob. Agents* **41**, 130–136 (2013).
8. Villagra, A. *et al.* Porin alterations present in non-carbapenemase-producing *Enterobacteriaceae* with high and intermediate levels of carbapenem resistance in Chile. *J. Med. Microbiol.* **61**, 1270–1279 (2012).
9. Geldart, K. & Kaznessis, Y. N. Characterization of Class IIa Resistance in Enterococcus faecium. *Antimicrob. Agents Chemother.* **61**, 1–17 (2017).
10. Gold, H. S. Vancomycin-Resistant Enterococci: Mechanisms and Clinical Observations. *Clin. Infect. Dis.* **33**, 210–219 (2001).
11. Gardete, S. & Tomasz, A. Mechanisms of vancomycin resistance in *Staphylococcus aureus*. *J. Clin. Invest.* **124**, 2836–2840 (2014).
12. Melnyk, A. H., Wong, A. & Kassen, R. The fitness costs of antibiotic resistance mutations. *Evol. Appl.* **8**, 273–283 (2015).
13. Rappe, M. S. & Giovannoni, S. J. The Uncultured Microbial Majority. *Annu. Rev. Microbiol.* **57**, 369–394 (2003).
14. Schloss, P. D. & Handelsman, J. Status of the Microbial Census. *Microbiol. Mol. Biol. Rev.* **68**, 686–691 (2004).
15. Nichols, D. *et al.* Use of ichip for high-throughput in situ cultivation of 'uncultivable' microbial species. *Appl. Environ. Microbiol.* **76**, 2445–2450 (2010).
16. Ling, L. L. *et al.* A new antibiotic kills pathogens without detectable resistance. *Nature* **517**, 455–459 (2015).
17. Blin, K. *et al.* antiSMASH 5.0: updates to the secondary metabolite genome mining pipeline. *Nucleic Acids Res.* **47**, 81–87 (2019).
18. Owens, R. C. & Ambrose, P. G. Antimicrobial Safety: Focus on Fluoroquinolones. *Clin. Infect. Dis.* **41**, S144–S157 (2005).
19. van den Bogaard, A. E., Bruinsma, N. & Stobberingh, E. E. The effect of banning avoparcin on VRE carriage in The Netherlands. *J. Antimicrob. Chemother.* **46**, 145–153 (2000).

20. Zeevi, D. *et al.* Personalized Nutrition by Prediction of Glycemic Responses. *Cell* **163**, 1079–1094 (2015).
21. Rea, D. *et al.* Microbiota effects on cancer: from risks to therapies. *Oncotarget* **9**, 17915–17927 (2018).
22. Sochocka, M. *et al.* The Gut Microbiome Alterations and Inflammation-Driven Pathogenesis of Alzheimer's Disease — a Critical Review. *Mol. Neurobiol.* **56**, 1841–1851 (2019).
23. Lessa, F. C. *et al.* Burden of *Clostridium difficile* Infection in the United States. *N. Engl. J. Med.* **372**, 825–834 (2015).
24. Hegemann, J. D., Zimmermann, M., Xie, X. & Marahiel, M. A. Lasso Peptides: An Intriguing Class of Bacterial Natural Products. *Acc. Chem. Res.* **48**, 1909–1919 (2015).
25. Adelman, K. *et al.* Molecular mechanism of transcription inhibition by peptide antibiotic Microcin J25. *Mol. Cell* **14**, 753–762 (2004).
26. Kimura, K. *et al.* Propeptin, a New Inhibitor of Prolyl Endopeptidase Produced by *Microbispora*. *J. Antibiot. (Tokyo).* **50**, 373–378 (1997).
27. Tietz, J. I. T. *et al.* A new genome mining tool redefines the lasso peptide biosynthetic landscape. *Nat. Chem. Biol.* **In press.**, 470–478 (2017).
28. Oliveira, H., São-José, C. & Azeredo, J. Phage-derived peptidoglycan degrading enzymes: Challenges and future prospects for in vivo therapy. *Viruses* **10**, E292 (2018).
29. Vidová, B., Zuzana, Š., Tišáková, L., Oravkinová, M. & Godány, A. Bioinformatics analysis of bacteriophage and prophage endolysin domains. *Biologia (Bratisl).* **69**, 541–556 (2014).
30. Fernández-ruiz, I., Coutinho, F. H. & Rodriguez-valera, F. Thousands of Novel Endolysins Discovered in Uncultured Phage Genomes. **9**, 1–8 (2018).
31. Beeby, M., Gumbart, J. C., Roux, B. & Jensen, G. J. Architecture and assembly of the Gram-positive cell wall. **88**, 664–672 (2013).
32. Darwin, C. *On The Origin of Species by Means of Natural Selection, or Preservation of Favoured Races in the Struggle for Life*. (John Murray, 1859).
33. Romero, P. a & Arnold, F. H. Exploring protein fitness landscapes by directed evolution. *Nat. Rev. Mol. Cell Biol.* **10**, 866–876 (2009).
34. Otwinowski, J., Mccandlish, D. M. & Plotkin, J. B. Inferring the shape of global epistasis. *Proc. Natl. Acad. Sci.* **115**, E7550–E7558 (2018).
35. Visser, J. A. G. M. De & Krug, J. Empirical fitness landscapes and the predictability of evolution. *Nat. Publ. Gr.* **15**, 480–490 (2014).
36. Sarkisyan, K. S. *et al.* Local fitness landscape of the green fluorescent protein. *Nature* **533**, 1–11 (2016).
37. Gray, V. E., Hause, R. J., Luebeck, J., Shendure, J. & Fowler, D. M. Quantitative Missense Variant Effect Prediction Using Large-Scale Mutagenesis Data. *Cell Syst.* **6**, 116-124.e3 (2018).
38. Woldring, D. R., Holec, P. V., Zhou, H. & Hackel, B. J. High-Throughput Ligand Discovery Reveals a Sitewise Gradient of Diversity in Broadly Evolved Hydrophilic Fibronectin Domains. *PLoS One* **10**, e0138956 (2015).
39. Woldring, D. R., Holec, P. V, Stern, L. A., Du, Y. & Hackel, B. J. A Gradient of Sitewise Diversity Promotes Evolutionary Fitness for Binder Discovery in a Three-Helix Bundle Protein Scaffold. *Biochemistry* **56**, 1656–1671 (2017).

40. James, M. *et al.* GROMACS: High performance molecular simulations through multi-level parallelism from laptops to supercomputers. *SoftwareX* **1**, 19–25 (2015).
41. Buß, O., Rudat, J. & Ochsenreither, K. FoldX as Protein Engineering Tool: Better Than Random Based Approaches? *Comput. Struct. Biotechnol. J.* **16**, 25–33 (2018).
42. Alford, R. F. *et al.* The Rosetta All-Atom Energy Function for Macromolecular Modeling and Design. *J. Chem. Theory Comput.* **13**, 3031–3048 (2017).
43. Childers, M. C. & Daggett, V. Insights from molecular dynamics simulations for protein design. *Mol. Syst. Des. Eng.* **2**, 9–33 (2017).
44. Sormanni, P., Aprile, F. A. & Vendruscolo, M. The CamSol method of rational design of protein mutants with enhanced solubility. *J. Mol. Biol.* **427**, 478–490 (2015).
45. Zambrano, R. *et al.* AGGRESCAN3D (A3D): Server for prediction of aggregation properties of protein structures. *Nucleic Acids Res.* **43**, W306–W313 (2015).
46. Jokinen, E., Heinonen, M. & Lähdesmäki, H. MGPfusion: Predicting protein stability changes with Gaussian process kernel learning and data fusion. *Bioinformatics* **34**, i274–i283 (2018).
47. Laimer, J., Hiebl-Flach, J., Lengauer, D. & Lackner, P. MAESTROweb: A web server for structure-based protein stability prediction. *Bioinformatics* **32**, 1414–1416 (2016).
48. Dorr, B. M., Ham, H. O., An, C., Chaikof, E. L. & Liu, D. R. Reprogramming the specificity of sortase enzymes. *Proc. Natl. Acad. Sci. U. S. A.* **111**, 13343–8 (2014).
49. Cheng, J., Randall, A. & Baldi, P. Prediction of protein stability changes for single-site mutations using support vector machines. *Proteins* **62**, 1125–1132 (2006).
50. Paladin, L., Piovesan, D. & Tosatto, S. C. E. SODA: prediction of protein solubility from disorder and aggregation propensity. *Nucleic Acids Res.* **8**, 1045–1063 (2017).
51. Salvat, R. S. *et al.* Computationally optimized deimmunization libraries yield highly mutated enzymes with low immunogenicity and enhanced activity. *Proc. Natl. Acad. Sci.* 201621233 (2017). doi:10.1073/pnas.1621233114
52. Wu, Y. The Potts model. *Rev. Mod. Phys.* **54**, (1982).
53. Martens, E. & Demain, A. L. The antibiotic resistance crisis, with a focus on the United States. *J. Antibiot. (Tokyo).* **70**, 520–526 (2017).
54. Van Boeckel, T. P. *et al.* Global trends in antimicrobial use in food animals. *Proc. Natl. Acad. Sci.* **112**, 5649–5654 (2015).
55. Scallan, E. *et al.* Foodborne illness acquired in the United States-Major pathogens. *Emerg. Infect. Dis.* **17**, 7–15 (2011).
56. The White House Administration. National Action Plan for Combating Antibiotic-Resistant Bacteria. *Open Gov. Natl. Action Plans* 1–63 (2015).
57. Ochman, H., Lawrence, J. G. & Groisman, E. a. Lateral gene transfer and the nature of bacterial innovation. *Nature* **405**, 299–304 (2000).
58. Guinane, C. M. & Cotter, P. D. Role of the gut microbiota in health and chronic gastrointestinal disease: understanding a hidden metabolic organ. *Therap. Adv. Gastroenterol.* **6**, 295–308 (2013).
59. Lopez, F. E., Vincent, P. A., Zenoff, A. M., Salomón, R. A. & Farías, R. N. Efficacy of microcin J25 in biomatrices and in a mouse model of *Salmonella* infection. *J. Antimicrob. Chemother.* **59**, 676–680 (2007).
60. Forkus, B., Ritter, S., Vlysidis, M., Geldart, K. & Kaznessis, Y. N. Antimicrobial

Probiotics Reduce *Salmonella enterica* in Turkey Gastrointestinal Tracts. *Sci. Rep.* **7**, 40695 (2017).

61. Hwang, I. Y. *et al.* Engineered probiotic *Escherichia coli* can eliminate and prevent *Pseudomonas aeruginosa* gut infection in animal models. *Nat. Commun.* **8**, 15028 (2017).

62. Geldart, K., Borrero, J. & Kaznessis, Y. N. A Chloride-Inducible Expression Vector for Delivery of Antimicrobial Peptides Against Antibiotic-Resistant *Enterococcus faecium*. *Appl. Environ. Microbiol.* **81**, 3889–3897 (2015).

63. Geldart, K., Forkus, B., McChesney, E., McCue, M. & Kaznessis, Y. N. pMPES: A Modular Peptide Expression Syste for the delivery of Antimicrobial Peptides. *Pharmaceuticals* **9**, 60 (2016).

64. Borrero, J., Chen, Y., Dunny, G. M. & Kaznessis, Y. N. Modified Lactic Acid Bacteria Detect and Inhibit Multiresistant Enterococci. *ACS Synth. Biol.* **4**, 299–306 (2014).

65. Vincent, P. A., Delgado, M. A., Farías, R. N. & Salomón, R. A. Inhibition of *Salmonella enterica* serovars by microcin J25. *FEMS Microbiol. Lett.* **236**, 103–107 (2004).

66. Sablé, S., Pons, A. M., Gendron-Gaillard, S. & Cottenceau, G. Antibacterial activity evaluation of microcin J25 against diarrheagenic Escherichia coli. *Appl Env. Microbiol* **66**, 4595–4597 (2000).

67. Maksimov, M. O., Pan, S. J. & James Link, a. Lasso peptides: structure, function, biosynthesis, and engineering. *Nat. Prod. Rep.* **29**, 996 (2012).

68. Duquesne, S. & Destoumieux-Garzón, D. Microcins, gene-encoded antibacterial peptides from enterobacteria. *Nat. Prod. …* **24**, 75005 (2007).

69. Severinov, K., Semenova, E., Kazakov, A., Kazakov, T. & Gelfand, M. S. Low-molecular-weight post-translationally modified microcins. *Mol. Microbiol.* **65**, 1380–1394 (2007).

70. Li, Y., Zirah, S. & Rebuffat, S. *Lasso Peptides: Bacterial Strategies to Make and Maintain Bioactive Entangled Scaffolds*. (Springer-Verlag New York, 2015). doi:10.1007/978-1-4939-1010-6

71. Pan, S. J., Cheung, W. L., Fung, H. K., Floudas, C. A. & Link, A. J. Computational design of the lasso peptide antibiotic microcin J25. *Protein Eng. Des. Sel.* **24**, 275–282 (2011).

72. Yan, K. P. *et al.* Dissecting the Maturation Steps of the Lasso Peptide Microcin J25 in vitro. *ChemBioChem* **13**, 1046–1052 (2012).

73. Mukhopadhyay, J., Sineva, E., Knight, J., Levy, R. M. & Ebright, R. H. Antibacterial peptide Microcin J25 inhibits transcription by binding within and obstructing the RNA polymerase secondary channel. *Mol. Cell* **14**, 739–751 (2004).

74. Pavlova, O., Mukhopadhyay, J., Sineva, E., Ebright, R. H. & Severinov, K. Systematic structure-activity analysis of microcin J25. *J. Biol. Chem.* **283**, 25589–25595 (2008).

75. Pan, S. J. & Link, a. J. Sequence diversity in the lasso peptide framework: Discovery of functional microcin J25 variants with multiple amino acid substitutions. *J. Am. Chem. Soc.* **133**, 5016–5023 (2011).

76. Ducasse, R. *et al.* Sequence determinants governing the topology and biological activity of a lasso peptide, microcin J25. *ChemBioChem* **13**, 371–380 (2012).

77. McClintock, M. K., Kaznessis, Y. N. & Hackel, B. J. Enterocin A mutants identified by saturation mutagenesis enhance potency towards vancomycin-resistant

Enterococci. *Biotechnol. Bioeng.* **113**, 414–423 (2016).

78. Tominaga, T. & Hatakeyama, Y. Determination of essential and variable residues in pediocin PA-1 by NNK scanning. *Appl. Environ. Microbiol.* **72**, 1141–1147 (2006).
79. Liu, W. & Hansen, J. N. Enhancement of the chemical and antimicrobial properties of subtilin by site-directed mutagenesis. *J. Biol. Chem.* **267**, 25078–25085 (1992).
80. Healy, B. *et al.* Intensive mutagenesis of the nisin hinge leads to the rational design of enhanced derivatives. *PLoS One* **8**, e79563 (2013).
81. Field, D. *et al.* Saturation mutagenesis of selected residues of the α-peptide of the lantibiotic lacticin 3147 yields a derivative with enhanced antimicrobial activity. *Microb. Biotechnol.* **6**, 564–575 (2013).
82. Boakes, S. *et al.* Generation of an actagardine A variant library through saturation mutagenesis. *Appl. Microbiol. Biotechnol.* **95**, 1509–1517 (2012).
83. Avram, S. *et al.* Evaluation of antimicrobial activity of new mastoparan derivatives using QSAR and computational mutagenesis. *Int. J. Pept. Res. Ther.* **17**, 7–17 (2011).
84. Tominaga, T. & Hatakeyama, Y. Development of innovative pediocin PA-1 by DNA shuffling among class IIa bacteriocins. *Appl. Environ. Microbiol.* **73**, 5292–5299 (2007).
85. Molloy, E. M. *et al.* Saturation Mutagenesis of Lysine 12 Leads to the Identification of Derivatives of Nisin A with Enhanced Antimicrobial Activity. *PLoS One* **8**, e58530 (2013).
86. Haugen, H. S. *et al.* Mutational Analysis of Residues in the Helical Region of the Class IIa. *Appl. Environ. Microbiol.* **77**, 1966–1972 (2011).
87. Kazazic, M., Nissen-Meyer, J. & Fimland, G. Mutational analysis of the role of charged residues in target-cell binding, potency and specificity of the pediocin-like bacteriocin sakacin P. *Microbiology* **148**, 2019–2027 (2002).
88. Pan, S. J., Cheung, W. L. & Link, A. J. Engineered gene clusters for the production of the antimicrobial peptide microcin J25. *Protein Expr. Purif.* **71**, 200–206 (2010).
89. Espah Borujeni, a., Channarasappa, a. S. & Salis, H. M. Translation rate is controlled by coupled trade-offs between site accessibility, selective RNA unfolding and sliding at upstream standby sites. *Nucleic Acids Res.* **42**, 2646–2659 (2014).
90. Nakamura, Y., Gojobori, T. & Ikemura, T. Codon usage tabulated from the international DNA sequence databases; its status 1999. *Nucleic Acids Res.* **27**, 292 (1999).
91. Lai, P. K. & Kaznessis, Y. N. Free Energy Calculations of Microcin J25 Variants Binding to the FhuA Receptor. *J. Chem. Theory Comput.* **13**, 3413–3423 (2017).
92. Mathavan, I. *et al.* Structural basis for hijacking siderophore receptors by antimicrobial lasso peptides. *Nat. Chem. Biol.* **10**, 340–2 (2014).
93. Johnson, J. R., Gajewski, A., Lesse, A. J. & Russo, T. a. Extraintestinal Pathogenic *Escherichia coli* as a Cause of Invasive Nonurinary Infections Extraintestinal Pathogenic Escherichia coli as a Cause of Invasive Nonurinary Infections. *Society* **41**, 5798–5802 (2003).
94. Tawfik, D. S. Accuracy-rate tradeoffs: How do enzymes meet demands of selectivity and catalytic efficiency? *Curr. Opin. Chem. Biol.* **21**, 73–80 (2014).
95. Pomares, M. F. *et al.* Potential applicability of chymotrypsin-susceptible microcin J25 derivatives to food preservation. *Appl. Environ. Microbiol.* **75**, 5734–5738 (2009).

96. Knappe, T. A. *et al.* Introducing lasso peptides as molecular scaffolds for drug design: Engineering of an integrin antagonist. *Angew. Chemie - Int. Ed.* **50**, 8714–8717 (2011).
97. Wilson, K.-A. *et al.* Structure of microcin J25, a peptide inhibitor of bacterial RNA polymerase, is a lassoed tail. *J. Am. Chem. Soc.* **125**, 12475–12483 (2003).
98. Pastagia, M., Schuch, R., Fischetti, V. a. & Huang, D. B. Lysins: The arrival of pathogen-directed anti-infectives. *J. Med. Microbiol.* **62**, 1506–1516 (2013).
99. Fenton, M., Ross, P., Mcauliffe, O., O'Mahony, J. & Coffey, A. Recombinant bacteriophage lysins as antibacterials. *Bioeng. Bugs* **1**, 9–16 (2010).
100. Fischetti, V. A. Bacteriophage lysins as effective antibacterials. *Curr. Opin. Microbiol.* **11**, 393–400 (2008).
101. Swift, S. *et al.* A Thermophilic Phage Endolysin Fusion to a *Clostridium perfringens*-Specific Cell Wall Binding Domain Creates an Anti-Clostridium Antimicrobial with Improved Thermostability. *Viruses* **7**, 3019–3034 (2015).
102. Blázquez, B., Fresco-taboada, A., Iglesias-bexiga, M. & Abedon, S. T. PL3 Amidase , a Tailor-made Lysin Constructed by Domain Shuffling with Potent Killing Activity against Pneumococci and Related Species. *Front. Microbiol.* **7**, 1–13 (2016).
103. Heselpoth, R. D. & Nelson, D. C. A New Screening Method for the Directed Evolution of Thermostable Bacteriolytic Enzymes. *J. Vis. Exp.* **69**, 1–8 (2012).
104. Heselpoth, R. D., Yin, Y., Moult, J. & Nelson, D. C. Increasing the stability of the bacteriophage endolysin PlyC using rationale-based FoldX computational modeling. *Protein Eng. Des. Sel.* **28**, 85–92 (2015).
105. São-José, C. Engineering of Phage-Derived Lytic Enzymes: Improving Their Potential as Antimicrobials. *Antibiotics* **7**, 29 (2018).
106. Scanlon, T. C., Dostal, S. M. & Griswold, K. E. A high-throughput screen for antibiotic drug discovery. *Biotechnol. Bioeng.* **111**, 232–243 (2014).
107. Steipe, B. *et al. Sequence statistics reliably predict stabilizing mutations in a protein domain*. *Journal of Molecular Biology* **240**, 188–192 (Academic Press, 1994).
108. Cochran, J. R., Kim, Y.-S., Lippow, S. M., Rao, B. & Wittrup, K. D. Improved mutants from directed evolution are biased to orthologous substitutions. *Protein Eng. Des. Sel.* **19**, 245–253 (2006).
109. Jäckel, C., Bloom, J. D., Kast, P., Arnold, F. H. & Hilvert, D. Consensus protein design without phylogenetic bias. *J. Mol. Biol.* **399**, 541–546 (2010).
110. Case, B. A. & Hackel, B. J. Synthetic and natural consensus design for engineering charge within an affibody targeting epidermal growth factor receptor. *Biotechnol. Bioeng.* **113**, 1628–1638 (2016).
111. Magliery, T. J. Protein stability: Computation, sequence statistics, and new experimental methods. *Curr. Opin. Struct. Biol.* **33**, 161–168 (2015).
112. Porebski, B. T. & Buckle, A. M. Consensus protein design. *Protein Eng. Des. Sel.* **29**, 245–251 (2016).
113. Reynolds, K. A., Russ, W. P., Socolich, M. & Ranganathan, R. Evolution-based design of proteins. *Methods Enzymol.* **523**, 213–235 (2013).
114. Broendum, S. S., Buckle, A. M. & Mcgowan, S. Catalytic diversity and cell wall binding repeats in the phage- encoded endolysins. *Mol. Microbiol.* **110**, 879–896 (2018).
115. Bershtein, S., Goldin, K. & Tawfik, D. S. Intense Neutral Drifts Yield Robust and

Evolvable Consensus Proteins. *J. Mol. Biol.* **379**, 1029–1044 (2008).

116. Bloom, J. D., Labthavikul, S. T., Otey, C. R. & Arnold, F. H. Protein stability promotes evolvability. *Proc. Natl. Acad. Sci. U. S. A.* **103**, 5869–5874 (2006).
117. Tokuriki, N., Stricher, F., Schymkowitz, J., Serrano, L. & Tawfik, D. S. The Stability Effects of Protein Mutations Appear to be Universally Distributed. *J. Mol. Biol.* **369**, 1318–1332 (2007).
118. Boder, E. T. & Wittrup, K. D. Yeast surface display for screening combinatorial polypeptide libraries. *Nat. Biotechnol.* **15**, 553–557 (1997).
119. Gai, S. A. & Wittrup, K. D. Yeast surface display for protein engineering and characterization. *Curr. Opin. Struct. Biol.* **17**, 467–473 (2007).
120. Traxlmayr, M. W. & Obinger, C. Directed evolution of proteins for increased stability and expression using yeast display. *Arch. Biochem. Biophys.* **526**, 174–180 (2012).
121. Hackel, B. J., Kapila, A. & Dane Wittrup, K. Picomolar Affinity Fibronectin Domains Engineered Utilizing Loop Length Diversity, Recursive Mutagenesis, and Loop Shuffling. *J. Mol. Biol.* **381**, 1238–1252 (2008).
122. Chen, I., Dorr, B. M. & Liu, D. R. A general strategy for the evolution of bond-forming enzymes using yeast display. *Proc Natl Acad Sci U S A* **108**, 11399–11404 (2011).
123. Schmitz, J. E., Ossiprandi, M. C., Rumah, K. R. & Fischetti, V. a. Lytic enzyme discovery through multigenomic sequence analysis in *Clostridium perfringens*. *Appl. Microbiol. Biotechnol.* **89**, 1783–95 (2011).
124. Biasini, M. *et al.* SWISS-MODEL: Modelling protein tertiary and quaternary structure using evolutionary information. *Nucleic Acids Res.* **42**, 252–258 (2014).
125. Tamai, E. *et al.* X-ray structure of a novel endolysin encoded by episomal phage phiSM101 of *Clostridium perfringens*. *Mol. Microbiol.* **92**, 326–337 (2014).
126. Rocklin, G. J. *et al.* Global analysis of protein folding using massively parallel design, synthesis, and testing. **175**, 168–175 (2017).
127. Miyazawa, S. Selection originating from protein stability/foldability: Relationships between protein folding free energy, sequence ensemble, and fitness. *J. Theor. Biol.* **433**, 21–38 (2017).
128. Hopf, T. A. *et al.* Mutation effects predicted from sequence co-variation. *Nat. Biotechnol.* **35**, 128–135 (2017).
129. Finn, R. D. *et al.* HMMER web server: 2015 Update. *Nucleic Acids Res.* **43**, W30–W38 (2015).
130. Bateman, A. *et al.* UniProt: The universal protein knowledgebase. *Nucleic Acids Res.* **45**, D158–D169 (2017).
131. Donovan, D. M. *et al.* The Cell Lysis Activity of the Streptococcus agalactiae Bacteriophage B30 Endolysin Relies on the Cysteine , Histidine-Dependent Amidohydrolase / Peptidase Domain. *Appl. Environ. Microbiol.* **72**, 5108–5112 (2006).
132. Donovan, D. M. *et al.* Peptidoglycan Hydrolase Fusions Maintain Their Parental Specificities. *Appl. Environ. Microbiol.* **72**, 2988–2996 (2006).
133. Donovan, D. M., Lardeo, M. & Foster-frey, J. Lysis of staphylococcal mastitis pathogens by bacteriophage phi11 endolysin. *FEMS Microbiol. Lett.* **265**, 133–139 (2006).
134. Stern, L. A. *et al.* Geometry and expression enhance enrichment of functional yeast-displayed ligands via cell panning. *Biotechnol. Bioeng.* **113**, 2328–2341 (2016).
135. Ng, A. Y. & Jordan, M. I. On Discriminative vs. Generative classifiers: A comparison

of logistic regression and naive Bayes. *Adv. Neural Inform. Process. Syst.* **14**, 605–610 (2001).

136. Halabi, N., Rivoire, O., Leibler, S. & Ranganathan, R. Protein Sectors: Evolutionary Units of Three-Dimensional Structure. *Cell* **138**, 774–786 (2009).

137. Pei, J., Kim, B. H. & Grishin, N. V. PROMALS3D: A tool for multiple protein sequence and structure alignments. *Nucleic Acids Res.* **36**, 2295–2300 (2008).

138. Finn, R. D. *et al.* The Pfam protein families database: Towards a more sustainable future. *Nucleic Acids Res.* **44**, D279–D285 (2016).

139. Marks, D. S., Hopf, T. A. & Sander, C. Protein structure prediction from sequence variation. *Nat. Biotechnol.* **30**, 1072–80 (2012).

140. Edgar, R. C. Search and clustering orders of magnitude faster than BLAST. *Bioinformatics* **26**, 2460–2461 (2010).

141. Barriere, S. L. Clinical , economic and societal impact of antibiotic resistance. *Expert Opin. Pharmacother.* **16**, 151–153 (2015).

142. Langdon, A., Crook, N. & Dantas, G. The effects of antibiotics on the microbiome throughout development and alternative approaches for therapeutic modulation. *Genome Med.* **8**, 1–16 (2016).

143. Francino, M. P. Antibiotics and the Human Gut Microbiome: Dysbioses and Accumulation of Resistances Increased Susceptibility to Infections. *Front. Microbiol.* **6**, 1–11 (2016).

144. Young, R. Y. Bacteriophage Lysis : Mechanism and Regulation. *Microbiol. Rev.* **56**, 430–481 (1992).

145. Loeffler, J. M., Nelson, D. & Fischetti, V. A. Rapid killing of Streptococcus pneumoniae with a bacteriophage cell wall hydrolase. *Science (80-.).* **294**, 2170–2172 (2001).

146. Schuch, R., Nelson, D. & Fischetti, V. A. A bacteriolytic agent that detects and kills Bacillus anthracis. *Nature* **418**, 440–445 (2002).

147. Schmelcher, M., Donovan, D. M. & Loessner, M. J. Bacteriophage endolysins as novel antimicrobials. *Future Microbiol.* **7**, 1147–71 (2012).

148. Nelson, D., Loomis, L. & Fischetti, V. a. Prevention and elimination of upper respiratory colonization of mice by group A streptococci by using a bacteriophage lytic enzyme. *Proc. Natl. Acad. Sci. U. S. A.* **98**, 4107–4112 (2001).

149. Wang, Q., Euler, C. W., Delaune, A. & Fischetti, V. A. Using a novel lysin to help control *Clostridium difficile* infections. *Antimicrob. Agents Chemother.* **59**, AAC.01357-15 (2015).

150. Marr, A. K., Gooderham, W. J. & Hancock, R. E. W. Antibacterial peptides for therapeutic use : obstacles and realistic outlook. *Curr. Opin. Pharmacol.* **6**, 468–472 (2006).

151. Loessner, M. J. Bacteriophage endolysins — current state of research and applications. *Curr. Opin. Microbiol.* **8**, 480–487 (2005).

152. Leonard, E. *et al.* Combining metabolic and protein engineering of a terpenoid biosynthetic pathway for overproduction and selectivity control. *Proc. Natl. Acad. Sci.* **107**, 13654–12659 (2010).

153. Vermassen, A. *et al.* Cell Wall Hydrolases in Bacteria : Insight on the Diversity of Cell Wall Amidases , Glycosidases and Peptidases Toward Peptidoglycan. *Front. Microbiol.* **10**, 1–27 (2019).

154. Diez-Martinez, R. *et al.* A novel chimeric phage lysin with high in vitro and in vivo

bactericidal activity against *Streptococcus pneumoniae*. *J. Antimicrob. Chemother.* **70**, 1763–1773 (2015).

155. Croux, C., Ronda, C., Lopez, R. & Garcia, J. L. Interchange of functional domains switches enzyme specificity: construction of a chimeric pneumococcal-clostridial cell wall lytic enzyme. *Mol. Microbiol.* **9**, 1019–1025 (1993).

156. Gong, P. *et al.* Characterization of Enterococcus faecium bacteriophage IME-EFm5 and its endolysin LysEFm5. *Virology* **492**, 11–20 (2016).

157. Evans Patterson, J. *et al.* An analysis of 110 serious enterococcal infections. Epidemiology, antibiotic susceptibility, and outcome. *Medicine (Baltimore).* **74**, 191–200 (1995).

158. DiazGranados, C. A., Zimmer, S. M., Klein, M. & Jernigan, J. A. Comparison of Mortality Associated with Vancomycin-Resistant and Vancomycin- Susceptible Enterococcal Bloodstream Infections: A Meta-analysis. *Clin. Infect. Dis.* **41**, 327–333 (2005).

159. Edmond, M. B., Ober, J. F., Dawson, J. D., Weinbaum, D. L. & Wenzel, R. P. Vancomycin-Resistant Enterococcal Bacteremia: Natural History and Attributable Mortality. *Clin. Infect. Dis.* **23**, 1234–1239 (1996).

160. Low, L. Y., Yang, C., Perego, M., Osterman, A. & Liddington, R. C. Structure and lytic activity of a Bacillus anthracis prophage endolysin. *J. Biol. Chem.* **280**, 35433–35439 (2005).

161. Zoll, S. *et al.* Structural Basis of Cell Wall Cleavage by a Staphylococcal Autolysin. *PLoS Pathog.* **6**, e1000807 (2010).

162. Yoong, P., Schuch, R., Nelson, D. & Fischetti, V. A. Identification of a Broadly Active Phage Lytic Enzyme with Lethal Activity against Antibiotic-Resistant Enterococcus faecalis and Enterococcus faecium. *J. Bacteriol.* **186**, 4808–4812 (2004).

163. Midelfort, K. S. *et al.* Redesigning and characterizing the substrate specificity and activity of *Vibrio fluvialis* aminotransferase for the synthesis of imagabalin. *Protein Eng. Des. Sel.* **26**, 25–33 (2013).

164. Miklos, A. E. *et al.* Structure-Based Design of Supercharged, Highly Thermoresistant Antibodies. *Chem. Biol.* **19**, 449–455 (2012).

165. Cocco, S., Feinauer, C., Figliuzzi, M., Monasson, R. & Weight, M. Inverse statistical physics of protein sequences : a key issues review. *Reports Prog. Phys.* **81**, 1–17 (2018).

166. Figliuzzi, M., Barrat-Charlaix, P. & Weigt, M. How pairwise coevolutionary models capture the collective residue variability in proteins. *Mol. Biol. Evol.* **35**, 1018–1027 (2018).

167. Miton, C. M. & Tokuriki, N. How mutational epistasis impairs predictability in protein evolution and design. *Protein Sci.* **25**, 1260–1272 (2016).

168. Anishchenko, I., Ovchinnikov, S., Kamisetty, H. & Baker, D. Origins of coevolution between residues distant in protein 3D structures. *Proc. Natl. Acad. Sci.* **114**, 9122–9127 (2017).

169. Levy, R. M., Haldane, A. & Flynn, W. F. Potts Hamiltonian models of protein co-variation, free energy landscapes, and evolutionary fitness. *Curr. Opin. Struct. Biol.* **43**, 55–62 (2017).

170. Jacquin, H., Gilson, A., Shakhnovich, E., Cocco, S. & Monasson, R. Benchmarking Inverse Statistical Approaches for Protein Structure and Design with Exactly Solvable Models. *PLoS Comput. Biol.* **12**, 1–18 (2016).

171. Ekeberg, M., Lovkvist, C., Lan, Y., Weigt, M. & Aurell, E. Improved contact prediction in proteins: Using pseudolikelihoods to infer Potts models. *Phys. Rev. E - Stat. Nonlinear, Soft Matter Phys.* **87**, 1–19 (2013).
172. Reetz, M. T., Wang, L.-W. & Bocola, M. Directed Evolution of Enantioselective Enzymes: Iterative Cycles of CASTing for Probing Protein- Sequence Space. *Angew. Chem. Int. Ed.* **45**, 1236–1241 (2006).
173. Sandalova, T. *et al.* The crystal structure of the major pneumococcal autolysin LytA in complex with a large peptidoglycan fragment reveals the pivotal role of glycans for lytic activity. *Mol. Microbiol.* **101**, 954–967 (2016).
174. Jacobs, T. M., Yumerefendi, H., Kuhlman, B. & Leaver-Fay, A. SwiftLib: Rapid degenerate-codon-library optimization through dynamic programming. *Nucleic Acids Res.* **43**, 1–10 (2015).
175. Gibson, D. G. *et al.* Enzymatic assembly of DNA molecules up to several hundred kilobases. *Nat. Methods* **6**, 12–16 (2009).
176. Ritter, S. C. & Hackel, B. J. Validation and Stabilization of a Prophage Lysin of Clostridium perfringens by Using Yeast Surface Display and Coevolutionary Models. *Appl. Environ. Microbiol.* **85**, 1–18 (2019).
177. Clark, D. P. & Pazdernik, N. J. *Biotechnology: Applying the Genetic Revolution*. (Academic Cell, 2009).
178. Cheng, Q. & Fischetti, V. a. Mutagenesis of a bacteriophage lytic enzyme PlyGBS significantly increases its antibacterial activity against group B streptococci. *Appl. Microbiol. Biotechnol.* **74**, 1284–91 (2007).
179. Proenc, D. *et al.* Phage Endolysins with Broad Antimicrobial Activity Against Enterococcus faecalis Clinical Strains. *Microb. Drug Resist.* **18**, 322–332 (2012).
180. Guan, R. *et al.* Structural basis for peptidoglycan binding by peptidoglycan recognition proteins. *Proc. Natl. Acad. Sci.* **101**, 17168–17173 (2004).
181. Zhang, L. *et al.* LysGH15 kills Staphylococcus aureus without being affected by the humoral immune response or inducing inflammation. *Sci. Rep.* **6**, 1–9 (2016).
182. Liu, Q., Xun, G. & Feng, Y. The state-of-the-art strategies of protein engineering for enzyme stabilization. *Biotechnol. Adv.* 0–1 (2018). doi:10.1016/j.biotechadv.2018.10.011
183. Kosciolek, T. & Jones, D. T. Accurate contact predictions using covariation techniques and machine learning. *Proteins* **84**, 145–151 (2015).
184. Marks, D. S. *et al.* Protein 3D Structure Computed from Evolutionary Sequence Variation. *PLoS One* **6**, e28766 (2011).
185. Rives, A. *et al.* Biological Structure and Function Emerge from Scaling Unsupervised Learning to 250 Million Protein Sequences. *bioRx* **622803**, 1–31 (2019).
186. Ingraham, J., Garg, V. K., Barzilay, R. & Jaakkola, T. Generative Models for Graph-Based Protein Design. *Int. Conf. Learn. Represent.* 1–10 (2019).
187. Zacharias, D. A., Violin, J. D., Newton, A. C. & Tsien, R. Y. Partitioning of Lipid-Modified Monomeric GFPs into Membrane Microdomains of Live Cells. *Science (80-.).* **296**, 913–917 (2002).
188. Chen, X., Zaro, J. L. & Shen, W.-C. Fusion protein linkers : Property , design and functionality. *Adv. Drug Deliv. Rev.* **65**, 1357–1369 (2013).
189. Sternke, M., Tripp, K. W. & Barrick, D. Consensus sequence design as a general strategy to create hyperstable , biologically active proteins. *Proc. Natl. Acad. Sci.*

116, 11275–11284 (2019).

190. Sievers, F. *et al.* Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Mol. Syst. Biol.* **7**, 539–539 (2014).

191. Papadimitriou, C. H. On the Complexity of Integer Programming. *J. Assoc. Comput. Mach.* **28**, 765–768 (1981).

Chapter 7 Appendices

7.1 Description of antimicrobial enzyme LysCP3 and validation of activity

In the process of identifying and validating LysCP2 (Chapter 3), another putative lysin was identified and validated for lytic activity against *Clostridium perfringens*. This lysin, from a prophage domain of *Clostridium perfringens* E str. JGS1987 (Accession WP_003465320) was also predicted to possess an SH3 cell wall binding domain, but unlike LysCP2, its catalytic domain was predicted to be an N-acetylmuramoyl-L-alanine amidase, similar to LysEFm5 (Chapter 4). The sequence of this putative lysin was synthesized and transformed into pETh as described previously (Section 3.5.2). After sequence verification, LysCP3 was expressed and purified via co-affinity chromatography (Section 3.5.3) and evaluated for activity against *Clostridium perfringens* ATCC 12916 (Figure 7.1). LysCP3 displayed activity against *C. perf*., which was less than that of LysCP2 in the context of this preliminary characterization. Future studies of LysCP3 for antimicrobial activity may benefit from exploration in conjunction with LysCP2, in particular to explore the efficacy change resulting from cleaving both the saccharide backbone as well as the stem peptide of the peptidoglycan of the target.

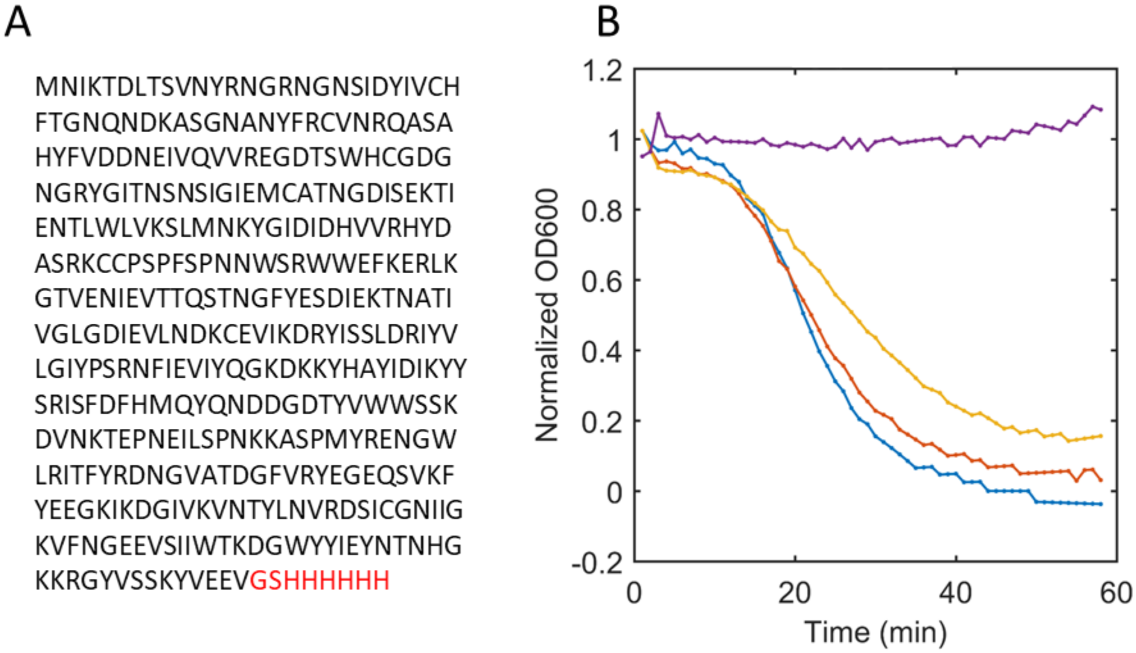


Figure 7.1. Sequence and activity of LysCP3. (A) The amino acid sequence of LysCP3 (black) and a GSHHHHHH tag to facilitate purification (red). (B) The activity of dilutions of purified LysCP3 as measured by reduction in optical density at 600 nm (OD600) of *Clostridium perfringens* ATCC 12916 at 37 °C. Briefly, exponential-phase *C. perf*. were harvested on ice and washed 2x with ice-cold PBS, resuspended and mixed with dilutions of eluted LysCP3 to a final concentration of 90% cell-suspension and 10% lysin dilution. Treatments: 1x lysin (blue); 0.5x lysin (orange); 0.25x lysin (yellow); buffer only (purple).

7.2 Demonstration of capacity of yeast-displayed LysCP2 to bind to *Clostridium perfringens* cell wall fragments via flow cytometry

In the process of progressing with the project outlined in Chapter 3, efforts were made to explore the possibility of directly assaying the enzymatic activity of lysins via fluorescent activated cell sorting on the yeast surface. As a component of this exploration it was desired to know if cell wall fragments could be bound to the yeast-displayed LysCP2. To assess this the cell-wall-binding domain of LysCP3 (Section 7.1) was assembled as a genetic fusion with monomeric Enhanced Green Fluorescent Protein (mEGFP)[187] with a flexible linker[188] (Design and assembly of this labeling reagent was conducted by Mike L.

Yang), hereafter mEGFP-LysCP3_CWBD. Yeast displaying LysCP2 (Section 3.5.11) were labeled with: (1) crude cell wall extract of *Clostridium perfringens* ATCC 12916 (Section 3.5.5), which had first been sonicated (30s on, 30s rest, on ice, 8x) and then centrifuged at 3000 ×*g* for 1 minute to remove large fragments; then (2) mouse anti-c-myc (clone 9E10, see Section 3.5.11); and then (3) goat anti-mouse conjugated with AlexaFluor 647 (see Section 3.5.11) and mEGFP-LysCP3_CWBD. This demonstrated (Figure 7.2) that the displayed LysCP2 retained its ability to bind to cell-wall fragments of the target organism. Beyond the intended function, future studies may explore this utility to enable characterization of specificity of lysin cell-wall-binding domains in high-throughput as well as enable characterization and optimization of physical properties via mutagenesis.

A

MVSKGEELFTGVVPILVELDGDVNGHKFSVSGEGEGDATYGKLTLKF
ICTTGKLPVPWPTLVTTLTYGVQCFSRYPDHMKQHDFFKSAMPEGY
VQERTIFFKDDGNYKTRAEVKFEGDTLVNRIELKGIDFKEDGNILGHK
LEYNYNSHNVYIMADKQKNGIKVNFKIRHNIEDGSVQLADHYQQN
TPIGDGPVLLPDNHYLSTQSKLSKDPNEKRDHMVLLEFVTAAGITLG
MDELYKGSAGSAAGSGEFVENIEVTTQSTNGFYESDIEKTNATIVGL
GDIEVLNDKCEVIKDRYISSLDRIYVLGIYPSRNFIEVIYQGKDKKYHAY
IDIKYYSRISFDFHMQYQNDDGDTYVWWSSKDVNKTEPNEILSPNK
KASPMYRENGWLRITFYRDNGVATDGFVRYEGEQSVKFYEEGKIKD
GIVKVNTYLNVRDSICGNIIGKVFNGEEVSIIWTKDGWYYIEYNTNH
GKKRGYVSSKYVEEVGSHHHHHH

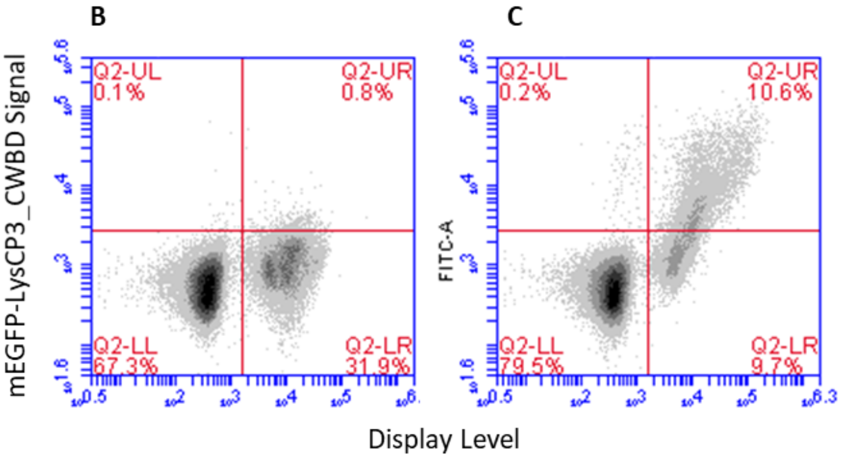
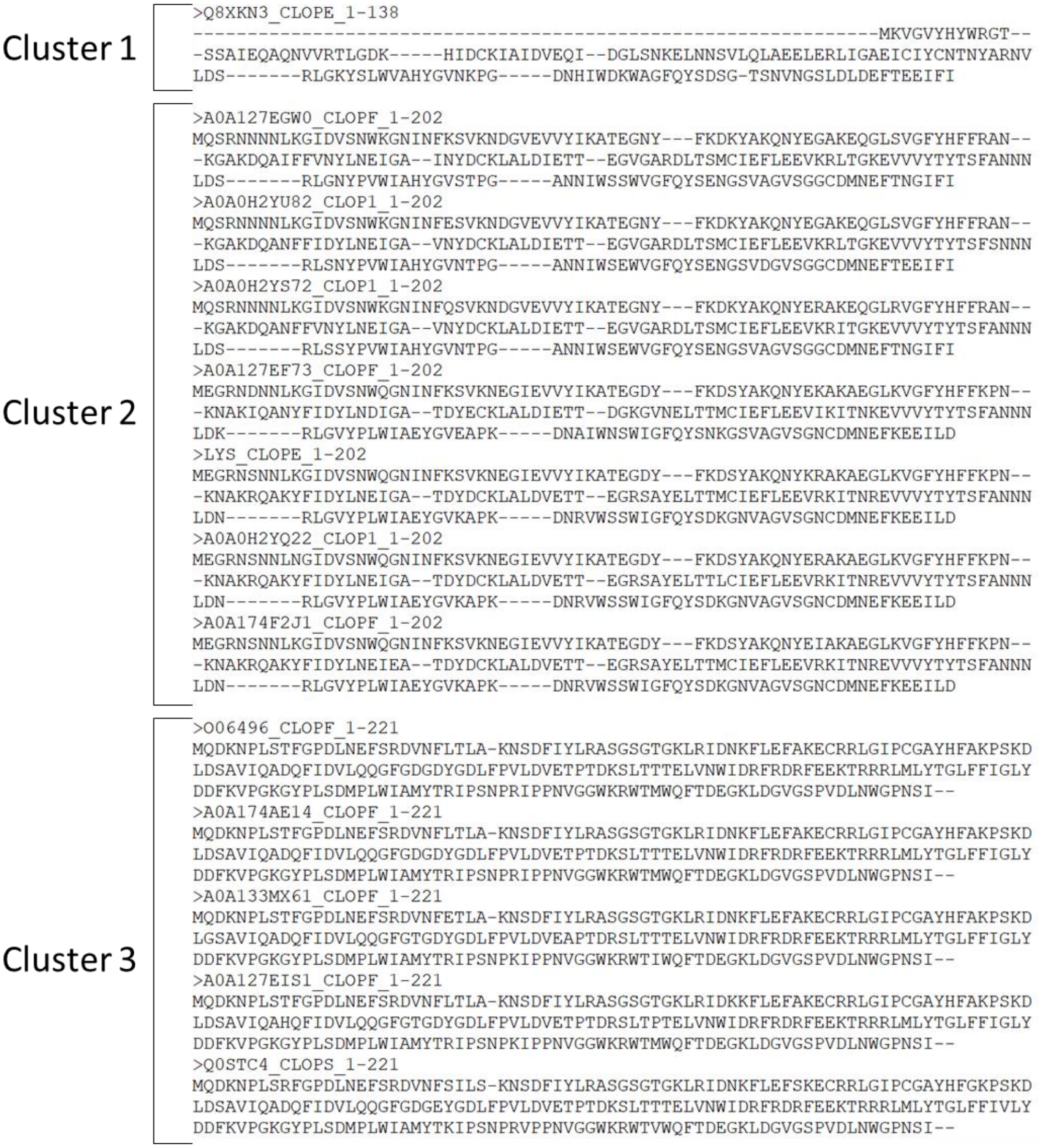


Figure 7.2. Labeling of yeast-displayed LysCP2 with cell wall fragments of Clostridium perfringens. (A) Sequence of mEGFP_LysCP3-CWBD with the mEGFP (green), linker (blue),LysCP3_CWBD (black), and purification tag (red) colored. Yeast displayed LysCP2 were labeled for cell-wall binding, using mEGFP-LysCP3_CWBD, either by first labeling with (C) or without (B) first labeling with *Clostridium perfringens* ATCC 12916 cell wall fragment. The non-expressing sub-population (bottom-left quadrant) is labeled by none of the reagents because it doesn't have any of the target epitopes and paratopes on its surface. Upon induction an expressing subpopulation (right quadrants) emerges. The mEGFP-LysCP3_CWBD reagent is found to only bind to expressing yeast after they have first been labeled with cell wall fragments.

7.3 LysCP2Con and LysCP2Con20: Antimicrobial enzymes designed with sitewise and pairwise homology information with activity against *Clostridium perfringens*

The logical extension of predicting the effects of mutations via second-order Potts models (Chapter 3 and Chapter 4) is to completely redesign a protein using this information. In the broader form of this process, the protein sequence with the maximum model score is identified. In the sitewise case this is considered whole-protein consensus design[189]. Though the solution to this problem becomes NP-complete with order greater than 1, it is worthwhile to consider how optimizing these higher-order models may affect the outcomes of the design. Herein the outcome of a limited study is presented wherein a novel pair of lysins are designed using sitewise and pairwise information of different natures. These proteins, LysCP2ConS and LysCP2ConP are built using sitewise and pairwise information, respectively.

To generate LysCP2Con, a JackHmmer[129] search limited to sequences within UniprotKB was conducted as before (Section 3.5.7) but limited in phylogeny to only *Clostridium perfringens*, yielding 13 sequences. The sequences which resulted were aligned using PROMALS3D[137] and assessed. It was observed that these sequences could be broadly clustered into 3 groups (Figure 7.3). The group within which LysCP2 belonged (Figure 7.3, Cluster 2) was selected for subsequent analysis (7 total sequences). The sitewise consensus of these proteins was determined, and named LysCP2ConS_cat, and was genetically fused to the cell-wall-binding domain of LysCP2 to yield LysCP2ConS.

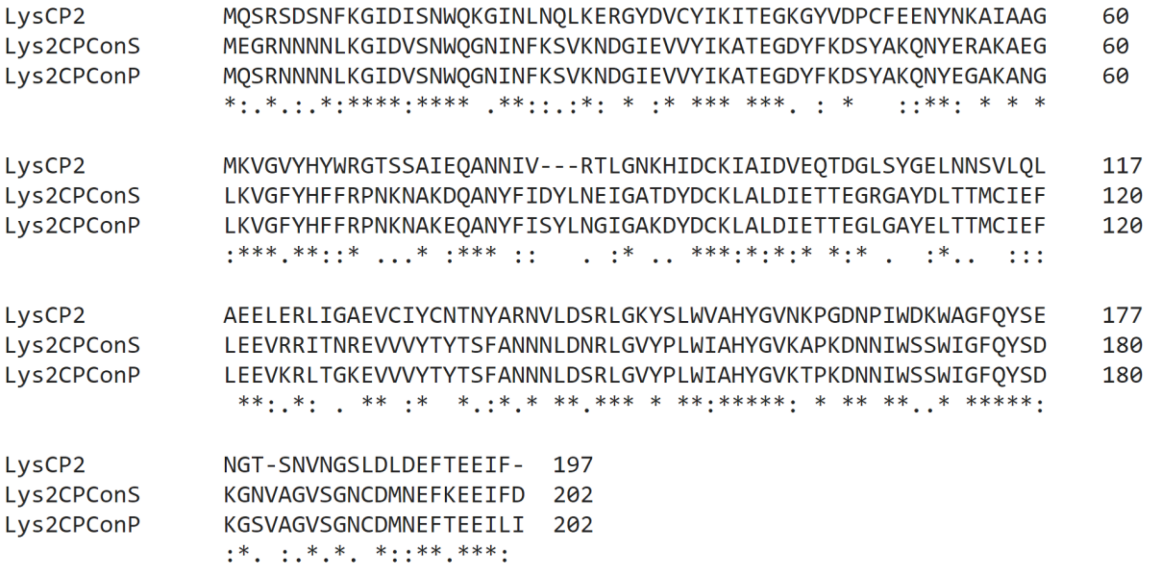

Cluster 1
Cluster 2
Cluster 3

```
>Q8XKN3_CLOPE_1-138
---------------------------------------------------------------MKVGVYHYWRGT--
-SSAIEQAQNVVRTLGDK-----HIDCKIAIDVEQI--DGLSNKELNNSVLQLAEELERLIGAEICIYCNTNYARNV
LDS-------RLGKYSLWVAHYGVNKPG-----DNHIWDKWAGFQYSDSG-TSNVNGSLDLDEFTEEIFI
>A0A127EGW0_CLOPF_1-202
MQSRNNNNLKGIDVSNWKGNINFKSVKNDGVEVVYIKATEGNY---FKDKYAKQNYEGAKEQGLSVGFYHFFRAN--
-KGAKDQAIFFVNYLNEIGA--INYDCKLALDIETT--EGVGARDLTSMCIEFLEEVKRLTGKEVVVYTYTSFANNN
LDS-------RLGNYPVWIAHYGVSTPG-----ANNIWSSWVGFQYSENGSVAGVSGGCDMNEFTNGIFI
>A0A0H2YU82_CLOP1_1-202
MQSRNNNNLKGIDVSNWKGNINFESVKNDGVEVVYIKATEGNY---FKDKYAKQNYEGAKEQGLSVGFYHFFRAN--
-KGAKDQANFFIDYLNEIGA--VNYDCKLALDIETT--EGVGARDLTSMCIEFLEEVKRLTGKEVVVYTYTSFSNNN
LDS-------RLSNYPVWIAHYGVNTPG-----ANNIWSEWVGFQYSENGSVDGVSGGCDMNEFTEEIFI
>A0A0H2YS72_CLOP1_1-202
MQSRNNNNLKGIDVSNWKGNINFQSVKNDGVEVVYIKATEGNY---FKDKYAKQNYERAKEQGLRVGFYHFFRAN--
-KGAKDQANFFVNYLNEIGA--VNYDCKLALDIETT--EGVGARDLTSMCIEFLEEVKRITGKEVVVYTYTSFANNN
LDS-------RLSSYPVWIAHYGVNTPG-----ANNIWSEWVGFQYSENGSVAGVSGGCDMNEFTNGIFI
>A0A127EF73_CLOPF_1-202
MEGRNDNNLKGIDVSNWQGNINFKSVKNEGIEVVYIKATEGDY---FKDSYAKQNYEKAKAEGLKVGFYHFFKPN--
-KNAKIQANYFIDYLNDIGA--TDYECKLALDIETT--DGKGVNELTTMCIEFLEEVIKITNKEVVVYTYTSFANNN
LDK-------RLGVYPLWIAEYGVEAPK-----DNAIWNSWIGFQYSNKGSVAGVSGNCDMNEFKEEILD
>LYS_CLOPE_1-202
MEGRNSNNLKGIDVSNWQGNINFKSVKNEGIEVVYIKATEGDY---FKDSYAKQNYKRAKAEGLKVGFYHFFKPN--
-KNAKRQAKYFIDYLNEIGA--TDYDCKLALDVETT--EGRSAYELTTMCIEFLEEVRKITNREVVVYTYTSFANNN
LDN-------RLGVYPLWIAEYGVKAPK-----DNRVWSSWIGFQYSDKGNVAGVSGNCDMNEFKEEILD
>A0A0H2YQ22_CLOP1_1-202
MEGRNSNNLNGIDVSNWQGNINFKSVKNEGIEVVYIKATEGDY---FKDSYAKQNYERAKAEGLKVGFYHFFKPN--
-KNAKRQAKYFIDYLNEIGA--TDYDCKLALDVETT--EGRSAYELTTLCIEFLEEVRKITNREVVVYTYTSFANNN
LDN-------RLGVYPLWIAEYGVKAPK-----DNRVWSSWIGFQYSDKGNVAGVSGNCDMNEFKEEILD
>A0A174F2J1_CLOPF_1-202
MEGRNSNNLKGIDVSNWQGNINFKSVKNEGIEVVYIKATEGDY---FKDSYAKQNYEIAKAEGLKVGFYHFFKPN--
-KNAKRQAKYFIDYLNEIEA--TDYDCKLALDVETT--EGRSAYELTTMCIEFLEEVRKITNREVVVYTYTSFANNN
LDN-------RLGVYPLWIAEYGVKAPK-----DNRVWSSWIGFQYSDKGNVAGVSGNCDMNEFKEEILD
>O06496_CLOPF_1-221
MQDKNPLSTFGPDLNEFSRDVNFLTLA-KNSDFIYLRASGSGTGKLRIDNKFLEFAKECRRLGIPCGAYHFAKPSKD
LDSAVIQADQFIDVLQQGFGDGDYGDLFPVLDVETPTDKSLTTTELVNWIDRFRDRFEEKTRRRLMLYTGLFFIGLY
DDFKVPGKGYPLSDMPLWIAMYTRIPSNPRIPPNVGGWKRWTMWQFTDEGKLDGVGSPVDLNWGPNSI--
>A0A174AE14_CLOPF_1-221
MQDKNPLSTFGPDLNEFSRDVNFLTLA-KNSDFIYLRASGSGTGKLRIDNKFLEFAKECRRLGIPCGAYHFAKPSKD
LDSAVIQADQFIDVLQQGFGDGDYGDLFPVLDVETPTDKSLTTTELVNWIDRFRDRFEEKTRRRLMLYTGLFFIGLY
DDFKVPGKGYPLSDMPLWIAMYTRIPSNPRIPPNVGGWKRWTMWQFTDEGKLDGVGSPVDLNWGPNSI--
>A0A133MX61_CLOPF_1-221
MQDKNPLSTFGPDLNEFSRDVNFETLA-KNSDFIYLRASGSGTGKLRIDNKFLEFAKECRRLGIPCGAYHFAKPSKD
LGSAVIQADQFIDVLQQGFGTGDYGDLFPVLDVEAPTDRSLTTTELVNWIDRFRDRFEEKTRRRLMLYTGLFFIGLY
DDFKVPGKGYPLSDMPLWIAMYTRIPSNPKIPPNVGGWKRWTIWQFTDEGKLDGVGSPVDLNWGPNSI--
>A0A127EIS1_CLOPF_1-221
MQDKNPLSTFGPDLNEFSRDVNFLTLA-KNSDFIYLRASGSGTGKLRIDKKFLEFAKECRRLGIPCGAYHFAKPSKD
LDSAVIQAHQFIDVLQQGFGTGDYGDLFPVLDVETPTDRSLTPTELVNWIDRFRDRFEEKTRRRLMLYTGLFFIGLY
DDFKVPGKGYPLSDMPLWIAMYTRIPSNPKIPPNVGGWKRWTMWQFTDEGKLDGVGSPVDLNWGPNSI--
>Q0STC4_CLOPS_1-221
MQDKNPLSRFGPDLNEFSRDVNFSILS-KNSDFIYLRASGSGTGKLRIDNKFLEFSKECRRLGIPCGAYHFGKPSKD
LDSAVIQADQFIDVLQQGFGDGEYGDLFPVLDVETPTDKSLTTTELVNWIDRFRDRFEEKTRRRLMLYTGLFFIVLY
DDFKVPGKGYPLSDMPLWIAMYTKIPSNPRVPPNVGGWKRWTVWQFTDEGKLDGVGSPVDLNWGPNSI--
```

Figure 7.3. Sequence alignment and clustering of LysCP2 homologs originating from Clostridium perfringens.

To generate LysCP2ConP, a second-order Potts model was again inferred (Section 3.5.7) but the source of the sequences was found with a JackHmmer search across UniprotKB with phylogeny restricted to Firmicutes and the seed sequence being that of LysCP2ConS_cat. These sequences were then aligned with PROMALS3D and

PLMC used to infer a second-order Potts model. This model was then used to optimize the sequence of LysCP2ConS_cat. Residues selected for mutation were restricted to those that were not perfectly preserved among the 7 sequences that generated LysCP2ConS_cat, with the assumption being that the conserved sites may be critical for some unknown function. A single step of the algorithm was: (1) Compute the complete site-scanning mutagenic landscape using the inferred model with the candidate sequence; (2) Select the best point-mutation and change the candidate sequence to this new sequence. The algorithm was initialized with LysCP2Con_cat being the candidate sequence at the first step, and was repeated for 20 iterations, at which point the score improvements upon mutation became negligible, to yield LysCP2ConP_cat. This catalytic domain was then genetically fused with the cell-wall-binding domain of LysCP2 to yield the final LysCP2ConP. An alignment[190] of these proteins is displayed in Figure 7.4.



```
LysCP2          MQSRSDSNFKGIDISNWQKGINLNQLKERGYDVCYIKITEGKGYVDPCFEENYNKAIAAG    60
Lys2CPConS      MEGRNNNNLKGIDVSNWQGNINFKSVKNDGIEVVYIKATEGDYFKDSYAKQNYERAKAEG    60
Lys2CPConP      MQSRNNNNLKGIDVSNWQGNINFKSVKNDGIEVVYIKATEGDYFKDSYAKQNYEGAKANG    60
                *:.*.:.*:****:**** .**::.:*: * :* *** ***. : *   ::**: * * *

LysCP2          MKVGVYHYWRGTSSAIEQANNIV---RTLGNKHIDCKIAIDVEQTDGLSYGELNNSVLQL    117
Lys2CPConS      LKVGFYHFFRPNKNAKDQANYFIDYLNEIGATDYDCKLALDIETTEGRGAYDLTTMCIEF    120
Lys2CPConP      LKVGFYHFFRPNKNAKEQANYFISYLNGIGAKDYDCKLALDIETTEGLGAYELTTMCIEF    120
                :***.**::* ...* :*** ::   . :* .. ***:*:*:* *:* .  :*..  :::

LysCP2          AEELERLIGAEVCIYCNTNYARNVLDSRLGKYSLWVAHYGVNKPGDNPIWDKWAGFQYSE    177
Lys2CPConS      LEEVRRITNREVVVYTYTSFANNNLDNRLGVYPLWIAHYGVKAPKDNNIWSSWIGFQYSD    180
Lys2CPConP      LEEVKRLTGKEVVVYTYTSFANNNLDSRLGVYPLWIAHYGVKTPKDNNIWSSWIGFQYSD    180
                 **:.*: . ** :*  *.:*.* **.*** * **:*****: * ** **..* *****:

LysCP2          NGT-SNVNGSLDLDEFTEEIF-  197
Lys2CPConS      KGNVAGVSGNCDMNEFKEEIFD  202
Lys2CPConP      KGSVAGVSGNCDMNEFTEEILI  202
                :*. :.*.*. *::**.***:
```

Figure 7.4. Alignment of LysCP2, LysCP2ConS, and LysCP2ConP catalytic domains Displayed is the result of alignment[190] of the catalytic domains of named lysins. Below each column symbols indicate the level of conservation: (*) Perfectly reserved; (:) Strongly similar residues; (.) Weakly similar residues.

Each of the proteins LysCP2, LysCP2ConS, and LysCP2ConP was purified (Section 3.5.3) and analyzed for activity against crude cell wall extract of *Clostridium perfringens* ATCC 12916 (Section 3.5.5). All lysins displayed cell wall degrading activity. Further, the thermal properties of each were assessed both for unfolding (Figure 7.5) and residual activity (Figure 7.6). These analyses indicate that Lys2ConS and Lys2ConP have enhanced thermal properties, compared with LysCP2. In addition, the catalytic domains of Lys2Con and Lys2Con20 appear to retain their thermal characteristics in isolation, which is in contrast to the catalytic domain of LysCP2 which sees its T_m greatly reduced (Figure 7.5). Interestingly, there appears to be a reversal in the trend when comparing the T_m and T_{50} of LysCP2ConS and LysCP2ConP in the denaturation and residual activity assays, respectively. Whereas LysCP2ConS has a T_m higher than LysCP2ConP, LysCP2Con20 has a significantly higher T_{50}. This may be evidence of enhanced refolding properties of LysCP2Con20.

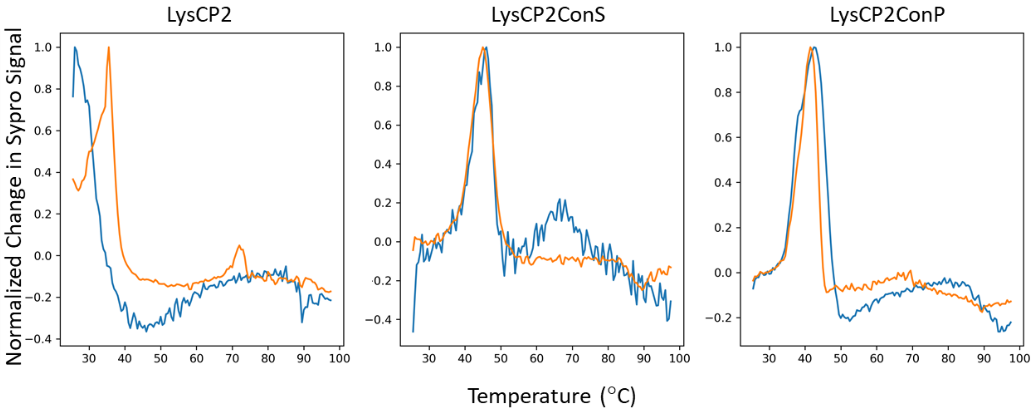


Figure 7.5. Assessment of unfolding temperature of full-length and catalytic domains of LysCP2, LysCP2ConS, and LysCP2ConP. To explore the unfolding properties of the designed proteins, the full-length lysins (orange) as well as the catalytic domains in isolation (blue) were expressed and assessed for stability via a Sypro Thermal Denaturation Assay (Section 3.5.4). These demonstrated T_m's of 38 °C, 45 °C, and 41.5

°C for the full-forms of LysCP2, LysCP2ConS, and LysCP2ConP, respectively. As well, T_m's of <30 °C, 46 °C, and 42.5 °C for the catalytic domains of the same, respectively.

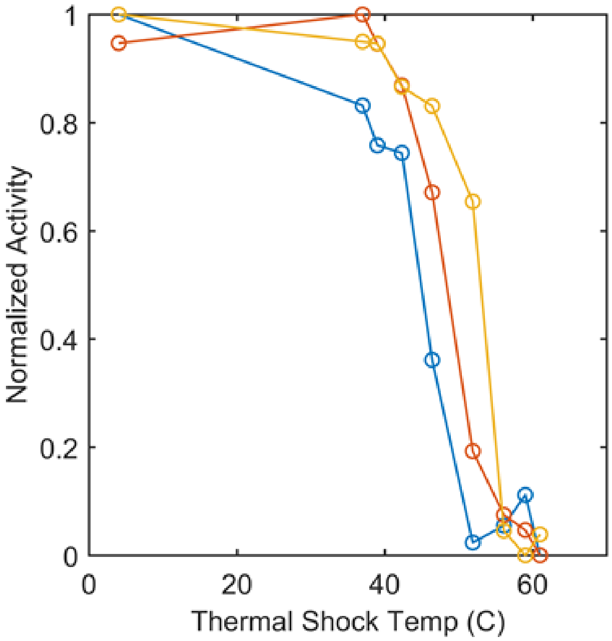


Figure 7.6. Residual activity of LysCP2, LysCP2ConS, and LysCP2ConP after thermal shock. LysCP2 (blue), LysCP2ConS (red), and LysCP2ConP (orange) were incubated at the indicated temperatures for 30 minutes and then rapidly cooled to 4 °C. Lysins were then incubated with crude cell wall of *Clostridium perfringens* ATCC 12916 at 37 °C and activity quantified and normalized lysins held at 4 °C without thermal shock. The data indicate T_{50}'s of 44.9 °C, 48.4 °C, and 53.0 °C for LysCP2, LysCP2ConS, and LysCP2ConP, respectively.

These analyses are intriguing first attempts at using pairwise data for whole-protein redesign to improve the physical properties of lysins. Future studies would be worthwhile to explore how the effects of different choices in the design process affect the properties of the final proteins. In principle these choices may be subdivided into three groups: (1) model architecture; (2) source of sequence information; and (3) sequence optimize protocol.

Throughout this work, second-order Potts models have been prioritized as the model architecture of choice. This choice was made both from a theoretic argument as well as recent experimental successes[128]. The inference of the parameters has been restricted to PLMC, due primarily to its speed at the cost of accuracy of inferred parameters[165]. Future studies may choose to explore how more accurate inference methods, such as Boltzmann Machine Learning, affect the performance of optimized proteins. In addition, all models presented are “centered” around a single focus sequence, in a format where observations are conditioned on gaps in the multiple sequence alignment. As a result of this treatment, only those positions of the multiple sequence alignment corresponding to the focus sequence are possible. Though more challenging for inference, model extension to properly facilitate gaps would permit the optimization process to extend beyond the domain of the target sequence.

As was seen for LysCP2 (Chapter 3) and LysEFm5 (Chapter 4), increasing the amount of sequences improved the predictive performance for stability and activity, respectively. Does this trend hold when examining completely redesigned proteins? Future studies would benefit from generating sitewise and pairwise sequences from different levels of phylogeny. It may be found that conserved residues for *Clostridium perfringens* become diluted with phylogenetically distant sequences and changed during whole-protein optimization.

Lastly, the optimization protocol was a simple steepest descent algorithm with mutations limited only to variable positions. Given that this is an integer linear programming problem, there is no general solution to the optimization procedure and it is considered to be NP-complete[191]. Future studies may explore how different algorithms perform in the context of these coevolutionary models and sample a range of constraints hypothesized to be consistent with functionality, such as conservation.