Engineering Antimicrobial Proteins and Peptides as Next-Generation Antimicrobials via Genome-mining, High-throughput Functional Screening, and Sequence Modeling

A DISSERTATION

SUBMITTED TO THE FACULTY OF THE

UNIVERSITY OF MINNESOTA

BY

Daniel Thomas Tresnak

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY

Benjamin J. Hackel

April 2022

Acknowledgements

This work is the culmination of the effort, support, and mentorship of countless individuals.

Thank you to my advisor, Ben Hackel, for being a role model for me, as a scientist and, more importantly, as a person. I have greatly appreciated your ability to focus on the finest detail of a research proposal one moment, discuss college sports the next, and then race a 100-yard dash. I have grown significantly thanks to your guidance and could not have asked for a better mentor.

Thank you to all past and present members of the Hackel lab, as it has been incredible to work within such a great community and learn from so many talented scientists. Of note, thank you to Seth and Alex for your mentorship and friendship, which I am happy to continue beyond our times in the Hackel lab.

Thank you to my thesis committee, Drs. Samira Azarin, Gary Dunny, and Casim Sarkar. I have learned tremendously from each of you through my coursework, research discussions, and beyond.

Thank you to all of my friends, within CEMS, locally, and across the country. You have been a constant voice of support, a much-needed distraction, and everything in between. Especially thanks to Neil, Wyatt, Abbey, Andreas, Brad and Laurel.

Last, but certainly not least, thank you to my family. Thank you to my parents for their constant support and instillment of my work ethic; thank you to my brothers for providing examples of what to and not to do; and thank you to my aunts, uncles, cousins, and beyond for being a bright part of my life.

Dedication

This work is dedicated to my wife, Clare. Thank you for your constant support, encouragement, and understanding; for every sacrifice you've made and continue to make; for being the bright spot (along with Blue) of every day. Words cannot capture my gratitude and without you I would not be here. Thank you.

Abstract

Antimicrobial resistance is a critical and worsening global health threat, driven by excessive and erroneous use of broad-spectrum antibiotics, coupled with minimal discovery of new antimicrobial therapies. Recent advances within the biotechnology field, in particular DNA synthesis and sequencing technologies and molecular and cellular biology techniques, has empowered protein engineering efforts. The encoding of proteins via efficiently synthesized DNA enables rapid construction and testing of large protein libraries, while combinatorial amino acid diversity provides nearly limitless protein phenotypes. Given this development, ribosomally synthesized antimicrobial proteins are one compelling solution in the design and discovery of new antimicrobial therapies. Yet, development of antimicrobial proteins as clinical therapies has remained limited, in part due to lack of high throughput methods for evaluating antimicrobial activity and statistical models for informing protein design.

This work spans sequence-function mapping of two antimicrobial protein families and the development of efficient strategies for their continued engineering. First, genome-mining approaches are applied to investigate the small antimicrobial protein family of class IIa bacteriocins. A library of class IIa bacteriocin variants was designed and experimentally evaluated for inhibitory activity to six strains of enterococcus. Ridge regression modeling yielded moderate predictive performance and elucidated factors impacting bacterial susceptibility to class IIa bacteriocins. Individual characterization of proteins with inhibitory activity yielded

a collection of variants with high potency and stability which are compelling for further studies. The latter half of this work focuses on the design of lysin catalytic domains, which degrade critical bonds in the peptidoglycan layer of targeted bacteria, and high throughput methods for screening lysin activity and stability. Structural information and epistatic models trained on natural sequence diversity were used to design lysin catalytic domain libraries, and experimental evaluation yielded one variant which displayed 1.8-fold improvement in catalytic activity and an 11.5 °C improvement in melting temperature compared to the parental catalytic domain. This enhanced variant was then used as a lead molecule across an array of protein diversification and library design strategies. A high throughput depletion-based assay was engineered for screening lysin catalytic domain activity and coupled with on-yeast protease stability assays to functionally evaluate ~5□104 lysin catalytic domain variants. Ridge regression modeling was conducted to elucidate sequence-function relationships, compare protein diversification strategies for informing epistatic models, and predict compelling new designs. This work identified several improved variants, expanded the explored lysin catalytic domain sequence space and demonstrated an efficient approach for lysin engineering.

Altogether, the research presented here advances protein engineering strategies broadly, validates the utility of high throughput methods for screening antimicrobial proteins, and empowers their continued development as next-generation antimicrobial therapies.

Table of Contents

List of Tables

List of Figures

Chapter 1 Introduction

1.1 Traditional broad-spectrum antibiotics and the development of antimicrobial resistance

Most antibiotics were originally discovered via screening of soil-derived bacteria for the ability to inhibit growth of test microorganisms[1–3] as was done for discovery of the most famous antibiotic, penicillin, first identified by Alexander Fleming in 1928[4]. These traditional antibiotics are small chemical (non-protein) molecules produced by bacteria to inhibit growth of competing microorganisms. While this screening approach, widely adopted by the pharmaceutical industry, yielded the discovery of most therapeutic antibiotics between 1940-1960 in the Golden Age of Antibiotic Discovery, further identification of new antibiotics drastically diminished after 1960, as this discovery process only yielded products that had been previously observed (Figure 1.1)[1,2,5].

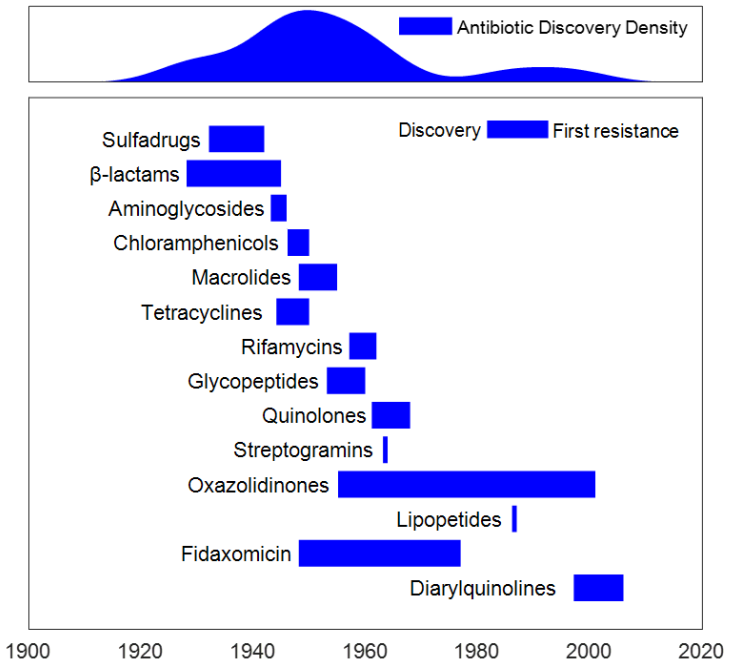


Figure 1.1 Discovery and first observed resistance to classes of traditional antibiotics.

(Top) Density plot showing presence of antibiotics without known resistance over time. (Bottom) Initial discovery of and first observed resistance to dominant classes of traditional antibiotics. Left side of bars represents discovery of antibiotics and right side of bars represents first observed resistance. Data adapted from Lewis, K.[2].

Worsening the problem of decreased antibiotic discovery, bacteria demonstrate a consistent ability to develop resistance to antibiotics via an array of mechanisms, resulting in worse therapeutic efficacy[6]. Bacterial modes of resistance are classified into three mechanisms: (1) reduction in intracellular antibiotic concentrations via efflux or reduced uptake, (2) modification of the antibiotic target via mutation or post-translational modification, and (3) inactivation of the antibiotic, often via enzymatic modification. Notably, the time required for development of these resistance mechanisms can vary significantly depending on

antibiotic target and mechanism complexity. For example, resistance to carbapenem, a beta lactam antibiotic which inhibits bacterial cell wall synthesis, is achieved in some *E. aerogenes* strains via down-regulation of the porin receptor needed for antibiotic uptake[7], whereas some resistant isolates of *Enterobacter spp.* and *K. pneumoniae* have mutations in outer membrane proteins that leave them non-functional[8,9]. Such resistance mechanisms can be observed in lab-grown bacteria within a day[10]. Conversely, resistance to vancomycin requires expression of three enzymes which modify the cell wall stem peptide to prevent vancomycin binding and was not observed clinically for ~30 years after initial approval in 1958[11,12]. This observation is the result of the, often substantial, fitness cost imposed on bacteria by mutations which yield antibiotic resistance. This fitness cost can impact bacterial ability to compete within their environment, and so mutations which pose minor costs are more likely to persist[13]. Thus, in design and development of new antimicrobial therapies, preference should be given to those which target critical cell components that would pose significant fitness costs to modify in an effort to minimize development of resistance.

Exacerbating development of antibiotic resistance, broad spectrum antibiotics, those which inhibit growth of many bacterial species with limited selectivity, are excessively and erroneously used. Broad spectrum activity applies an unnecessary selective pressure on bacterial species to develop resistance, and so development of new antimicrobial therapies capable of selective targeting of pathogenic bacteria are needed to slow further resistance development. Additionally, excessive use of antibiotics, for example in agriculture, which

accounts for ~70% of medically important antibiotics, further applies such a selective pressure[14]. Genetic determinants encoding these resistance mechanisms are then capable of being transferred between bacterial species, including to clinically relevant pathogenic bacteria, via horizontal gene transfer, allowing for accumulation of multiple antibiotic resistance genes [15]. Ultimately, our lack of discovery of new antibiotics and poor antimicrobial stewardship has led to an antimicrobial resistance crisis, with 2.8 million antibiotic-resistant infections and 35000 deaths in the United States[16]. Globally, antimicrobial resistance is projected to cause 10 million deaths annually by 2050[14].

1.2 Antimicrobial peptides and proteins are compelling antimicrobial therapies

Solutions to the current antimicrobial resistance crisis must span both improved antimicrobial stewardship policies, which focus on systematically modifying how and where we use antimicrobial therapies to reduce new development of resistance, and a platform for continued discovery of new, potent, selective antimicrobial therapies to pace development of resistance. Recent efforts taken by international agencies address the shortcomings in antimicrobial stewardship policies and long-term goals to aid in continued improvement in this space[14,16,17]. Towards developing a platform for continued discovery and engineering of antimicrobial therapies, ribosomally synthesized antimicrobial peptides and proteins (AMPs) are a compelling option[18–21].

AMPs are polypeptides, often derived from bacteria, which achieve growth inhibition using a diverse array of mechanisms[22,23]. For example, class IIa bacteriocins are relatively small AMPs (30-50 amino acids in length) that achieve inhibitory activity via membrane receptor-mediated pore formation. These AMPs specifically target bacteria which express a particular family of mannose phosphotransferase (manPTS) receptors, enabling selective targeting of a subset of bacterial species[24,25]. This specificity, coupled with high thermal stability, make class IIa bacteriocins interesting for development as antimicrobial therapies, most likely in combination with other antimicrobials due to moderate rates of resistance. In contrast to class IIa bacteriocins, endolysins (hereafter referred to as lysins) are multi-domain proteins derived from bacteriophage. Lysins are typically composed of distinct catalytic domains (CDs), which degrade critical bonds in peptidoglycan, and cell wall-binding domains (CWBD), which bind to the cell walls of targeted bacteria[26,27]. The modular design of lysins and their ability to be engineered for specific and potent targeting of both Gram-positive and Gram-negative bacteria makes them arguably the most compelling family of AMPs as next-generation antimicrobial therapies. Ultimately, the diverse structures and functions of AMPs are highly compelling for development as antimicrobial therapies, as these can enable selective targeting of clinically relevant pathogens and development as combination therapies comprised of antimicrobials with different modes of action to limit development of resistance.

Ribosomally synthesized AMPs also offer the ability for continued, rapid engineering as compared to traditional small molecule antibiotics. Most traditional

antibiotics are natural products or derivatives thereof, and engineering of alternatives often requires multiple synthesis steps for generation of a small number of new molecules. Further, attempts at synthesis of new small molecule antibiotics has largely proven unsuccessful[5]. Alternatively, ribosomally synthesized AMPs can be encoded via DNA fragments, which can be shuttled into bacterial expression strains which then produce the encoded AMP of interest. Advances in biotechnology have drastically reduced the cost of DNA synthesis over the past two decades, enabling more cost-efficient assembly of large libraries of DNA fragments encoding proteins of interest. These libraries can then be inserted into different bacteria strains, enabling production of many diverse proteins for individual testing. Further, following insertion into bacterial strains, the DNA fragment encoding a protein of interest can be extracted and sequenced to identify the protein sequence and connect this with functional metrics. Advances in DNA sequencing have further empowered this field, as these libraries can now be efficiently sequenced, allowing quantitative evaluation of millions of proteins in some instances[28–32].

1.3 Protein engineering and applications towards antimicrobial proteins

The diversity observed across AMP structure and function, and, notably, the ability for continued AMP engineering, is the result of the immense possible diversity achieved by combining the 20 amino acids into polypeptide chains of varying lengths. Yet, this diversity also poses significant challenges when trying to engineer enhanced proteins, as the possible number of amino acid combinations

(and protein diversity) scales exponentially with protein length, yielding a massive protein sequence space. For example, a relatively small protein of 50 amino acids has 20^{50} (~10^{65}) possible amino acid combinations, more than the number of stars in the universe (~$6*10^{22}$)[33]. Further complicating this challenge, prior work has found that most protein sequences are lacking in functions of interest and that single amino acid substitutions in a protein sequence can result in oblation of protein function[34]. Thus, we require efficient strategies for traversing protein sequence space.

Protein engineering efforts have focused both on development of assays capable of screening large libraries of protein variants, which expand the protein sequence space we are capable of experimentally evaluating, as well as strategies for informing protein design, which focus protein mutagenesis toward more functional protein sequence space. Protein screening strategies have been successfully developed for high throughput (10^5-10^9 variants) evaluation of binding to targets of interest[35–37], thermal and proteolytic stability[28], developability[30], and other functions of interest; yet screening of antimicrobial activity has remained mostly limited (10^2-10^3 variants) given its complex and diverse modes of action[38–41]. Efforts to screen libraries of AMPs often conduct activity testing on agar or 96-well plates, resulting in testing of up to 10^3 variants. Recently, however, two new strategies have been implemented to enable testing of ~10^5-10^6 AMPs. The first uses co-encapsulation of a target bacterial strain with a population of AMP-secreting bacteria[42]. These cells are encapsulated in an agarose droplet, which acts as a physical linkage of protein genotype and

phenotype. Droplets can be fluorescently labeled to identify growth inhibition of the target bacteria and then sorted, via fluorescence-activated cell sorting (FACS), to collect active AMPs. A more recent approach requires engineering of a system where bacterial expression of AMPs with inhibitory activity kills the cells that express them, resulting in depletion of active AMPs from the population over time[31,43]. Depletion in these strategies is then quantified by high throughput sequencing of the initial and final populations. Both of these strategies have significantly enhanced the throughput of screening AMP activity; yet expanding these platforms to new AMP families with distinct modes of action often requires significant developmental work and has not been achieved for most AMP families.

Towards rational design of protein libraries, a myriad of approaches has been implemented. For example, when available, protein structure information can be used to identify potentially beneficial amino acid substitutions[44,45]. Further, use of bottom-up approaches, such as Rosetta[46] and FoldX[47], which leverage empirical energy potentials to sample and predict protein folding and stability, can be used in place of or in addition to structural information to guide protein mutagenesis. These approaches can provide significant benefit in designing proteins, especially for stability engineering efforts; however, until recent advances in protein structure modeling with the development of AlphaFold, structural information has been limited[48]. Even so, predicted structures are not always accurate enough to inform design and structural approaches do not always provide sufficient information to engineer more complex protein function, such as enzymatic activity. Alternatively, natural protein sequences or experimental

sequence-function data can be used to elucidate trends in protein sequences and identify beneficial or deleterious mutations via regression or more complicated machine learning models. Yet, natural protein sequence-based models assume these proteins have an evolutionary fitness sufficient to exist in nature, but this fitness does not always correlate with desired protein phenotypes. Further, models trained on experimental sequence-function data are limited to those functions which we can screen in high throughput. All and combinations of these protein design approaches have yielded enhanced engineering of protein stability[49,50], catalytic activity[51], developability[30], and other molecular properties. However, engineering of AMPs has, until recently, been limited to small, rationally designed libraries of compelling AMPs derived directly from bacteria. Thus, efforts to advance our ability to screen larger and more diverse AMP libraries, coupled with protein modeling strategies to learn critical AMP sequence-function relationships, will significantly aid the design and discovery of new antimicrobial therapies to combat antimicrobial resistance.

1.4 Contributions of Dissertation

This thesis focuses on the discovery and engineering of compelling AMP families and development of screening and design strategies for their continued efficient engineering.

1.4.1 Genome-mining and statistical modeling of class IIa bacteriocins

In Chapter 2, we experimentally evaluate 210 class IIa bacteriocin variants, identified via genome-mining and random mutagenesis, for antimicrobial activity

against 6 strains of enterococci. Inhibitory activity was ridge regressed to AMP sequence and strain information, yielding moderate predictive performance and supporting the dominant role of manPTS as the target of class IIa bacteriocins. Active AMPs were individually tested across eight enterococcus and four *Listeria* strains to elucidate trends in susceptibility and demonstrated that manPTS sequence is informative of susceptibility but other factors, such as membrane composition, also contribute to susceptibility. Ongoing work detailed in Appendix 7.1 investigates other genes in *E. faecalis* strain OG1RF that enable resistance to class IIa bacteriocins, as well as the impact of treatment with these AMPs on biofilm formation. This effort has yielded multiple AMPs displaying high potency and stability with compelling potential for clinical translation and set the stage for ongoing work that will provide mechanistic insights into the development of resistance to class IIa bacteriocins.

1.4.2 Engineering lysins for enhanced activity and stability via epistatic models

In Chapters 3 and 4, we develop and apply high throughput methods for engineering enhanced lysin catalytic domains. We first apply site-saturation mutagenesis at sites selected with guidance by sequence and structural information from homologous proteins (Chapter 3). Experimental evaluation of 873 unique variants yielded one which displayed 1.8-fold improvement in catalytic activity and an 11.5 °C improvement in melting temperature compared to the wildtype LysEFm5 lysin catalytic domain. This variant, termed LysV7, was then used as a lead molecule in design of sequence-diverse libraries (Chapter 4). A

high throughput depletion-based assay was engineered for screening catalytic domain activity and coupled with on-yeast protease stability assays to functionally evaluate lysin libraries. Experimental data identified several compelling variants with superior performance compared to LysV7, while ridge regression modeling elucidated sequence-function relationships, identified efficient strategies for diversifying proteins to inform epistatic models, and predicted compelling variants not observed in our libraries. These advances empowered lysin engineering by validation of a high throughput screening platform, demonstrated an efficient library diversification and analysis strategy for mapping protein sequence-function landscapes, and yielded compelling lysin variants for further evaluation *in vivo*.

Chapter 2 - Mining and statistical modeling of natural and variant class IIa bacteriocins elucidate activity and selectivity profiles across species

Adapted from “Tresnak, D. T. & Hackel, B. J. 2020. ‘Mining and statistical modeling of natural and variant class IIa bacteriocins elucidates activity and selectivity profiles across species.’ *Appl. Environ. Microbiol.* **86**, 1–18.”

2.1 Abstract

Class IIa bacteriocin antimicrobial peptides (AMPs) are a compelling alternative to current antimicrobials because of potential specific activity towards antibiotic-resistant bacteria, including vancomycin-resistant enterococci. Engineering of these molecules would be enhanced by a better understanding of AMP sequence-activity relationships to improve efficacy *in vivo* and limit effects of off-target activity. Towards this goal, we experimentally evaluated 210 natural and variant class IIa bacteriocins for antimicrobial activity against six strains of enterococci. Inhibitory activity was ridge regressed to AMP sequence to predict performance, achieving an area under the curve of 0.70 and demonstrating the potential of statistical models for identifying and designing AMPs. Active AMPs were individually produced and evaluated against eight enterococci strains and four *Listeria* strains to elucidate trends in susceptibility. It was determined that the mannose phosphotransferase system sequence is informative of susceptibility to class IIa bacteriocins, yet other factors, such as membrane composition, also

contribute strongly to susceptibility. A broadly potent bacteriocin variant (lactocin DT1) from a *Lactobacillus ruminis* genome was identified as the only variant with inhibitory activity towards all tested strains, while a novel enterocin variant (DT2) from an *E. faecium* genome demonstrated specificity towards *Listeria* strains. Eight AMPs were evaluated for proteolytic stability to trypsin, chymotrypsin, and pepsin, and three C-terminal disulfide-containing variants, including divercin V41, were identified as compelling for future *in vivo* studies given high potency and proteolytic stability.

2.2 Statement of Importance

Class IIa bacteriocin antimicrobial peptides (AMPs), an alternative to traditional small-molecule antibiotics, are capable of selective activity towards various Gram-positive bacteria, limiting negative side effects associated with broad-spectrum activity. This selective activity is achieved through targeting of the mannose phosphotransferase system (manPTS) of a subset of Gram-positive bacteria, although factors affecting this mechanism are not entirely understood. Peptides identified from genomic data, as well as variants of previously characterized AMPs, can offer insight into how peptide sequence affects activity and selectivity. The experimental methods presented here identify promising potent and selective bacteriocins for further evaluation, highlight the potential of simple computational modeling for prediction of AMP performance, and demonstrate that factors beyond manPTS sequence affect bacterial susceptibility to class IIa bacteriocins.

2.3 Introduction

Antibiotic resistance is a growing threat to global public health. Reports estimate that over 2 million people in the United States suffer from antibiotic-resistant infections annually, resulting in 23,000 deaths and up to $20 billion in excess costs [14,52]. It is projected that 10 million annual deaths globally will be caused by antimicrobial resistance by the year 2050, demonstrating the worsening state of this problem [14]. Improper use of antibiotics and the lack of discovery of new antimicrobial therapies are the two main factors driving this development. High use of antibiotics, including medically relevant antibiotics, in animals for food production creates an unnecessary selective pressure for the development of antibiotic resistance [53,54]. Similarly, broad-spectrum antibiotics, antibiotics which have antimicrobial activity towards a wide array of pathogenic and commensal bacteria alike, are over-prescribed and overused, further contributing to the development of resistance. These mechanisms of antibiotic resistance can then spread to clinically relevant, pathogenic bacterial populations through horizontal gene transfer [15]. Further exacerbating this situation, few new antibiotics have been discovered due to limited financial incentives and high risk associated with their development for pharmaceutical companies [1,2]. Given these circumstances, there is a dire need for novel classes of selective antimicrobial therapies.

Antimicrobial peptides (AMPs) are one compelling alternative to traditional, small-molecule antibiotics, either as individually produced free peptides [55–57] or through engineered probiotic delivery of AMPs at the site of infection [58–61]. AMPs offer a variety of different mechanisms of antimicrobial activity, with a range

of selectivity, potency, and cytotoxicity. While the potency of these molecules demonstrates their ability to replace or support current antibiotics, the high selectivity of these molecules is particularly important, both to minimize unnecessary selective pressure on the broader bacterial populations and to limit negative health effects associated with the broad-spectrum activity of small-molecule antibiotics. For example, it has been demonstrated that broad spectrum antibiotics vastly reshape the flora in the microbiota [62,63], allowing for opportunistic infections, such as *Clostridium difficile*, to develop, while use of more selective antimicrobial therapies, such as the two-peptide bacteriocin thuricin CD, results in much lower perturbation of the gut microbiota [64,65]. While natural evolution has developed families of AMPs with selective activity to various bacterial species, further work is necessary to optimize this activity towards clinically relevant bacterial strains while minimizing off-target activity to broader bacterial populations. Previous work has been conducted towards this aim with several AMPs, such as the screening of random variants of the lantibiotic nisin for improved activity and selectivity to multiple Gram-positive bacterial strains [66], the generation of synthetic AMPs containing species-specific targeting moieties [67], and the evaluation of single- and multi-mutant microcin J25 variants for reduced activity to commensal *E. coli* strains while retaining high activity to pathogenic *Salmonella* [41], among others.

Class IIa bacteriocins are one promising family of AMPs characterized by small size (30-50 amino acids), a highly conserved N-terminal alpha-helical domain with a more variable C-terminal domain, and potent, selective activity to

various strains of Gram-positive bacteria [24,68–70]. Class IIa bacteriocins are known to interact directly with the mannose phosphotransferase systems (manPTSs) of various Gram-positive bacteria. Specifically, class IIa bacteriocins interact with the EIIC and EIID components of manPTSs to induce pore formation in the bacterial membrane and ion leakage, ultimately resulting in cell death [71–73]. Previous work with the class IIa bacteriocin pediocin PA-1 has demonstrated that these AMPs are capable of effectively targeting *Listeria in vivo* with limited impact on the commensal bacteria [57,74]. Further work optimizing class IIa bacteriocins has greatly expanded our understanding of how AMP sequence affects its activity. Multiple studies have focused on mutating natural class IIa bacteriocins to identify residues essential for activity [75], assess the beneficial impact of charged residues on target-cell binding and activity [76], and identify variants with improved activity or stability [77–80]. Additionally, it has been well demonstrated that susceptible bacteria readily develop resistance to class IIa bacteriocins through downregulation of target manPTSs and other mechanisms, such as shifts in metabolic activity or cell membrane adaptation [10,81,82]. However, it is hypothesized that combinatorial treatments of class IIa bacteriocins with AMPs or antibiotics that act through different mechanisms will aid in limiting resistance [83–85].

While class IIa bacteriocins show promise as novel antimicrobial therapies, much work remains to develop efficacious molecules for *in vivo* treatment. Proteolytic stability is a key limitation for peptide therapies *in vivo* [86,87], and the ability to engineer peptides that retain antimicrobial activity with increased

proteolytic stability is lacking. Towards this goal, further work exploring AMP sequence-activity relationships is necessary to use in parallel with the understanding of protein sequence-stability relationships. Furthermore, factors affecting the susceptibility of bacterial strains to particular AMPs must be better understood to aid the engineering of AMPs with selective targeting of harmful bacteria to limit negative health effects caused by broad-spectrum activity and slow the development of resistance. Expansive evaluation of class IIa bacteriocins would inform these elements broadly and identify new candidate therapeutics. Thus, in this study, we evaluated a library of natural and variant class IIa bacteriocins for antimicrobial activity across an array of bacterial strains. We used ridge regression to predict AMP performance from sequence and further analyzed the susceptibility of strains of enterococci and *Listeria* to class IIa bacteriocins to elucidate the extent of the impact manPTS sequence has on susceptibility.

2.4 Results

2.4.1 AMP library design contains broad coverage of class IIa bacteriocin sequence space

We sought to evaluate class IIa bacteriocin AMPs to expand our understanding of sequence-function relationships and to identify promising lead therapeutics. We hypothesized that evaluation of a broad array of sequence diversity would allow us to elucidate structure-activity relationships across different bacterial strains and identify AMPs with desirable selectivity profiles. To achieve the desired breadth, we constructed an AMP library containing 150 previously identified class IIa bacteriocins from the UniProt database [88] as well as random and rationally

designed variants of six previously studied bacteriocins. We included variants that have been previously characterized to serve as positive controls while exploring entirely novel variants identified from genomic information. The high prevalence of randomly mutated class IIa variants (900 random variants out of 1150 total variants within the library) enabled us to explore the class IIa bacteriocin tolerance to random mutagenesis. Rationally designed AMPs were used to evaluate the utility of domain swapping for identifying compelling class IIa bacteriocins and our ability to design active and proteolytically stable variants using PeptideCutter software to identify protease cleavage sites [89]. Domain swapping has been previously used in literature to identify functional class IIa bacteriocins, so we sought to further explore this method [78,90]. Additionally, proteolytic stability is one known limitation of class IIa bacteriocins for *in vivo* applications [18,91], so rational design of highly stable lead molecules would be a useful advance.

Six previously characterized class IIa bacteriocins (Table 2.1) were selected as the basis for generation of random and rational variants. Here, we denote these six AMPs as seed sequences, as they seeded the random and rational AMP libraries. These six bacteriocins were chosen because they have all been previously studied multiple times [75,78,92–95], allowing for comparisons with other studies, and have been shown to have relatively high potencies, suggesting these are good starting points for engineering. While the amino acid sequences used for enterocin A, enterocin NKR-5-3C, enterocin P, and divercin V41 were the natural AMP sequences, inactive variants of pediocin PA-1 (D17N) and sakacin P (K11E) were used as seed sequences [75,78]. The inclusion of pediocin PA-1$_{D17N}$

and sakacin P_{K11E} allowed for the inclusion of known negative controls within the library screen while also exploring whether any gain-of-activity variants could be found.

Table 2.1 Alignment of seed and consensus sequences and identification of head, interior, and tail regions.

	Alignment Position																																																	
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	36	37	38	39	40	41	42	43	44	45	46	47	48	49	50
Consensus	T	T	H	S	G	K	Y	Y	G	N	G	V	Y	C	N	K	K	K	C	W	V	D	W	G	Q	A	W	T	C	I	G	N	N	S	A	N	G	W	A	G	G	-	A	I	P	G	-	-	K	C
Enterocin A	·	·	·	·	·	·	·	·	·	·	·	·	·	·	T	·	N	·	·	T	·	·	·	A	K	·	T	·	·	·	A	G	M	·	I	G	·	F	L	·	·	·	·	·	·	·	·	·	·	·
Enterocin NKR-5-3C	A	·	-	-	-	-	·	·	·	·	·	L	·	·	·	S	·	·	·	·	·	E	·	·	I	T	G	G	·	L	A	Q	Y	A	I	G	·	·	L	·	·	·	·	·	·	·	·	·	·	·
Enterocin P	A	·	-	-	-	R	S	·	·	·	·	·	·	·	·	N	S	·	·	·	·	N	·	·	E	·	K	E	N	·	A	G	I	V	I	S	·	·	·	S	·	L	·	G	M	·	·	·	H	-
Divercin V41	·	K	-	-	-	-	·	·	·	·	·	·	·	·	·	S	·	·	·	·	·	·	·	·	·	·	S	G	·	·	·	Q	T	V	V	G	·	·	L	·	·	·	·	·	·	·	·	·	·	·
Pediocin PA-1_D17N	-	K	-	-	-	-	·	·	·	·	·	·	T	·	G	·	H	S	·	S	·	N	·	·	K	·	T	·	·	·	I	·	·	G	·	M	A	·	·	T	·	·	G	H	Q	·	N	H	·	·
Sakacin P_K11E	-	K	-	-	-	-	·	·	·	·	·	·	H	·	G	E	H	S	·	T	·	·	·	·	T	·	I	G	N	·	·	·	·	A	·	A	N	·	·	T	·	·	G	N	A	·	W	N	·	-
	N-terminal Domain						Internal Domain																																	C-terminal Domain										

Dots represent agreement with the consensus sequence.

To generate class IIa chimeras, the six seed sequences were aligned, N-terminal, internal, and C-terminal domains were defined, and domains were swapped between all six seed sequences. All unique N-terminal domains of seed sequences were swapped between the internal and C-terminal domains of all active seed sequences to generate 20 chimeras. To explore possible interactions between the ATR_{1-3} N-terminal domain sequence and the Y7S mutation unique to enterocin P, chimeras were constructed with the enterocin P N-terminal domain and each interior and C-terminal domain with a Y7S mutation. 30 chimeras were generated by swapping all unique N-terminal and C-terminal domains with the consensus [96–98] sequence of all class IIa bacteriocin identified in the UniProt database. Random variants were generated containing 2, 4, or 6 simultaneous mutations in seed sequences (150 variants for each seed sequence for a total of 900 random variants).

The genes encoding the described natural, rational, and random class IIa bacteriocin variants were synthesized as a pool of oligonucleotides (oligopool).

The genes were amplified, cloned into the pNZC expression vector with a Usp45 signal peptide [99,100], and transformed into *L. lactis* for expression. The chloride-inducible pNZC expression vector was chosen to act as a constitutive expression vector, given the high chloride concentration in brain heart infusion (BHI), and for convenient use in *L. lactis* [99]. L. lactis was chosen for AMP expression because of its status as a model lactic acid bacterium and its demonstrated ability to deliver peptides in vivo in a contained system [101–103]. To sample the library, individual, random colonies were grown in each well of a total of 13 deep 96-well plates. Whole-cell PCR was conducted on each well to append plate, row, and column indices to identify AMP sequences using high-throughput sequencing. Nine hundred forty wells (75%) could be confidently identified, with a total of 309 unique AMP sequences, 166 of which were in the initial oligopool (53%). This low frequency of sequences from the initial oligopool (166/1130 = 15% of oligopool sequences) most likely stemmed from initial bias in the oligopool, which may have been amplified during PCR steps. The remaining 143 unique sequences consisted of point mutants which resulted in sequences highly similar to sequences in the oligopool (44 sequences), insertion/deletion mutants (55 sequences), and point mutants which resulted in premature stop codons (39 sequences), and sequences containing DNA construction errors (5 sequences). While this frequency of erroneous sequences is higher than expected, a total of 210 identified sequences (Table S2.2) were within expected class IIa sequence space and were used for further analysis. While alternative gene synthesis or pool sampling approaches

could enable even deeper coverage of the proposed population, the set of 210 represents a broad set of IIa variants.

2.4.2 AMP library exhibits broad activity towards enterococcus and varied tolerance to random mutagenesis

To evaluate the isolated class IIa bacteriocins, we performed agar diffusion assays to categorize variants as highly active, moderately active, or inactive followed by more precise characterization of all active variants. Six enterococci strains – four *E. faecalis* and two *E. faecium* – were used as indicators for reasonable breadth and to provide an opportunity to assess strain and species selectivity. The selected strains were chosen as a relevant sampling of pathogenic and nonpathogenic strains, many of which have known antibiotic resistance.

Sixteen of 20 (80%) natural class IIa bacteriocins displayed some level of inhibitory activity towards at least one strain (Table S2.2). The high rate of activity is expected, given that several of these variants have been previously shown to be active. However, eight of the tested AMPs have not been previously characterized to our knowledge, five of which were found to be active. Thus, our library design effectively achieved one of its goals of expanding the active IIa repertoire. As for the four natural IIa bacteriocins observed as inactive against all six strains, one was the negative control variant seed sequence of pediocin PA-1_{D17N}, while two others were fragmented variants of enterocin A. These data provide a partial constraint on their known activities in regard to partial selectivity or ineffective expression from the *L. lactis* host while also identifying limitations of genomic annotation to discover active AMPs. Four of the six parental seed sequences were

observed in library screening, with enterocin A and sakacin P_{K11E} being unobserved.

Thirteen of 183 (7%) random variants were active (Table S2.2). Of the 103 random variants whose parental sequence were observed as active, 11% retained some level of inhibitory activity toward at least one indicator strain (Figure 2.1A). The active frequency was substantially higher for double mutants (10/30 = 33%) than quadruple or hexa-mutants (2% and 0%, respectively). These results confirm relatively high tolerance to two mutations but low tolerance to accumulated mutation [104,105]. One hexa-mutant, sakacin $P_{K11E,\ Y2V,\ G4D,\ N5V,\ N24C,\ I25G,\ N27E}$ exhibited activity to one strain; however, since the parental seed sequence sakacin P_{K11E} was unobserved, this AMP was not included in Figure 2.1A and it is unclear whether the parental sequence had any activity.

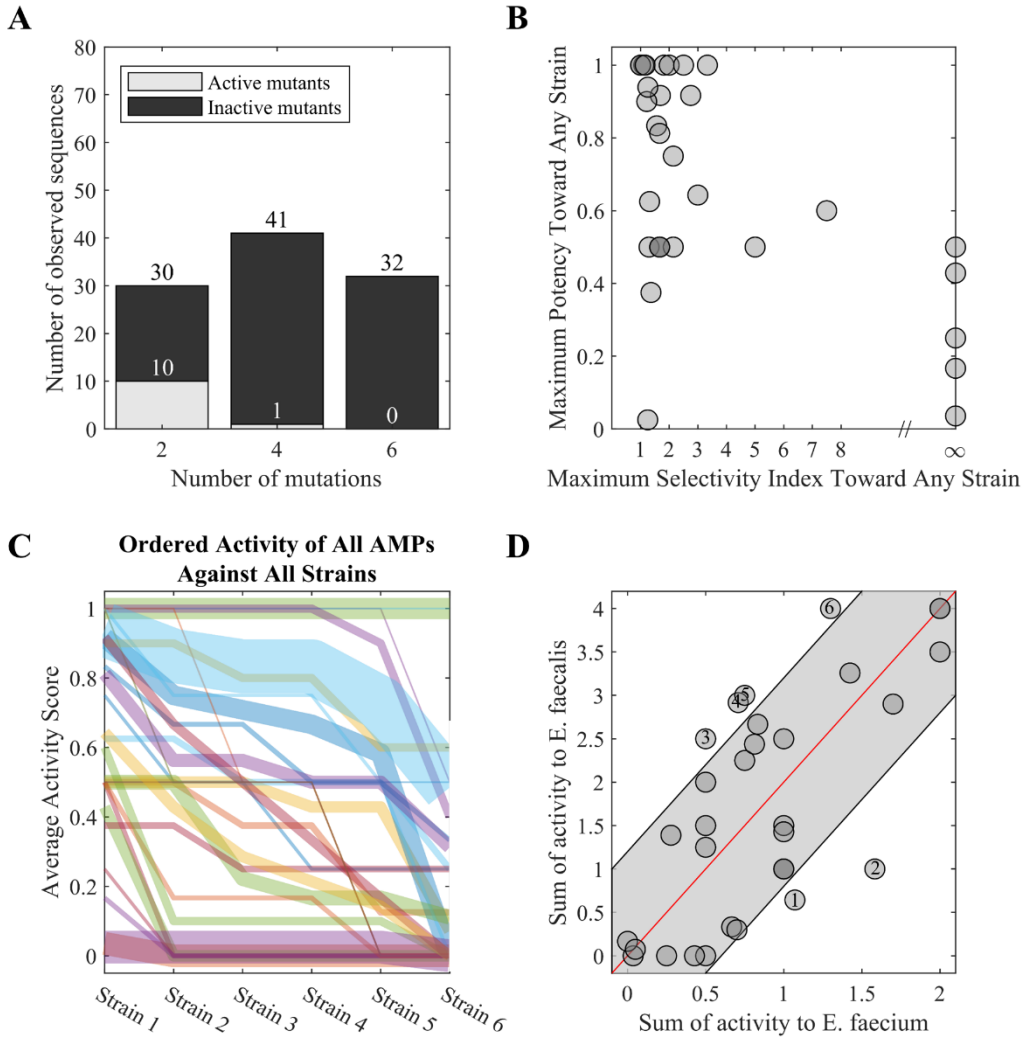


Figure 2.1 Summary of observed AMP activity from library screen displays broad activity to all tested strains.

(A) Activity of observed random mutants suggests bacteriocin IIa AMPs have a relatively low tolerance to mutagenesis. (B) Most AMPs display an increasing potency across all indicator strains with limited selectivity. We define a selectivity index here as the activity score towards one strain divided by the average activity score across the other five strains. (C) Average AMP activity ordered by rank across all six indicator strains. Line thickness indicates number of observations for a particular AMP. (D) Comparison of AMP activity between E. faecalis and E. faecium indicator strains. Sum of activity scores to E. faecalis and E. faecium strains is shown on the y- and x-axes, respectively. Red line represents an equal fraction of activity toward both sets of strains. Gray zone represents zone of 50% selectivity between E. faecalis and E. faecium strains. Six AMPs fall outside the

50% selectivity zone: enterocin DT1 (1) and enterocin DT3 (2) showed selectivity to E. faecium while lactocin DT2 (3), enterocin NKR-5-3CS12A, E18D (4), enterocin NKR-5-3CS12T,T22A (5), and enterocin NKR-5-3C (6) had selectivity towards E. faecalis.

Incomplete library sampling hindered our ability to extensively evaluate the impact of chimeric design or proteolytic stability design. Only six domain-swapped variants and one enterocin A stability variant were observed in library evaluation. Three domain-swapped variants retained activity towards multiple strains, which is suggestive of tolerance, yet the limited sampling of chimeric options precludes a robust conclusion. Interestingly, all three inactive domain-swapped variants included the consensus interior region, suggesting critical residues may have been lost in design of the consensus region. The one observed enterocin A stability variant was inactive across all six observations.

To identify selective AMPs and trends in class IIa bacteriocin activity, the activity scores of all AMPs (0 for inactive, 0.5 for inhibition less than enterocin P, 1.0 for comparable to or more active than enterocin P) was averaged over all observations and analyzed across all six indicator strains (Figure 2.1). Most active AMPs exhibited broad-spectrum activity toward multiple indicator strains with limited selectivity (Figure 2.1BC, Table S2.2). Twenty-three of 32 AMPs exhibited an activity score $\geq$0.2 against at least 4 of 6 strains, while 17 AMPs had an activity score of $\geq$0.5 against at least 4 strains. Yet multiple AMPs exhibited strong specificity: five AMPs were observed with activity towards only one strain, including enterocin $P_{N7V,\ Y10E}$ and enterocin NKR-5-3C$_{G20V,\ I31M}$, with activity scores of 0.5 and 0.43, respectively. In comparison to the parental enterocin P and enterocin

NKR-5-3C activity scores, these mutations drastically reduced observed activity across all six indicator strains, with only activity towards one strain being retained. Converse to these selective AMPs, only five AMPs exhibited full breadth with activity scores >0.5 across all six indicator strains, including enterocin P and two enterocin P variants: G6D, K15G and K15R, V31S (Figure 2.1C, Table S2.2).

When comparing activity across species, most AMPs displayed comparable activity towards both *E. faecalis* and *E. faecium* strains (Figure 2.1D). The strongest exception was enterocin DT3, an AMP from *E. pallens*, which demonstrated higher activity towards the two *E. faecium* indicator strains than the four *E. faecalis* strains (Figure 2.1D). One additional variant exhibits appreciable *E. faecium* preference while four AMPs exhibit *E. faecalis* preference.

2.4.3 Class IIa bacteriocins show limited specificity between species

To more thoroughly determine AMP activity and specificity profiles, all 32 active AMPs were individually produced, and their total activities were quantified against eight strains of enterococci and four strains of *L. monocytogenes*. *L. monocytogenes* strains were included to evaluate specificity trends between enterococci and *Listeria*, and an additional two *E. faecium* strains were added to have an equal sampling of *E. faecium*, *E. faecalis*, and *L. monocytogenes* strains. We chose to produce the unmodified AMPs in *L. lactis* cultures and conduct ammonium sulfate (AS) precipitation to generate more concentrated samples rather than using purification tags which may affect structure or activity of small class IIa AMPs. Given this method of production, we quantified total activity as the minimum inhibitory dilution (MID), which is the lowest dilution of resuspended AS

precipitation solution that inhibited growth. This metric of total activity is the product of an AMP's ability to be produced by *L. lactis* and inhibit growth, which is directly meaningful to intended applications with probiotic delivery. To ensure this was comparable across all AMPs, *L. lactis* growth conditions and AS precipitation resuspension volumes were kept constant. MIDs were determined using serial dilutions of resuspended AS precipitation products in an agar-diffusion experiment, and the lowest dilution which led to formation of any zone of inhibition was determined to be the MID.

Several tested AMPs showed comparable activity to nearly all tested indicator strains, suggesting a lack of specificity. Seventeen of 32 AMPs (53%) displayed inhibitory activity toward at least 10 of the 12 indicator strains. Two AMPs were especially potent, with MIDs of $\leq$0.1 against 10 of the 12 indicator strains, correlating to a 10-fold dilution of resuspended AS precipitation product inhibiting bacterial growth. Only one AMP, lactocin DT1 from *Lactobacillus ruminis*, had activity towards all 12 indicator strains, albeit with limited activity to *E. faecium* strain NRRL B-2354 and *L. monocytogenes* strain V7. Hiracin JM79 had a MID of $\leq$0.04, displaying high potency, towards all strains except *L. monocytogenes* strain V7, towards which it was inactive. Interestingly, *E. faecium* strain NRRL B-2354 and *L. monocytogenes* strain V7 were tolerant to nearly all class IIa bacteriocins tested. This is particularly interesting for *E. faecium* strain NRRL B-2354 because it shares identical and nearly identical manPTS EIIC and EIID genes with *E. faecium* strains 8E9 and 6E6, respectively, both of which were highly susceptible to class IIa bacteriocins (Table 2.2, Figure 2.2).

AMP	Sequence	EF-8E9	EF-6E6	EF-7A	EF-B2354	EF-V583	EF-CH116	EF-Pan7	EF-Com1	LM-AS1775	LM-M03	LM-CDC7762	LM-V7
Consensus	ATKYYGNGVYCNKXKCWVNWGEAKGCIAXIAIGGWAGGALAGKGHVGGR												
Hiracin JM79	AT····L····E····D·NQ···E·GK·IVN··VNHGPWAPRR	0.0025±0.0005	0.006±0.002	0.008±0.001	0.04±0.01	0.02±0.005	0.03±0.01	0.04±0.02	0.035±0.009	0.006±0.002	0.011±0.005	0.011±0.003	>1
Divercin V41	··········SK····D··Q·S···GQTVV···L···IP··C	0.01±0.002	0.2±0.1	0.06±0.04	>1	0.01±0.002	0.01±0.002	0.01±0.01	0.028±0.007	0.003±0.001	0.01±0.002	0.005±0.004	>1
Enterocin P_K15R, V31S	··RS·····NSR·····EN··G·S·S···S·-···M··	0.04±0.03	0.04±0.03	0.7±0.2	>1	0.04±0.03	0.08±0.02	0.1±0.1	0.1±0.1	0.04±0.03	0.25±0.06	0.2±0.1	>1
Lactocin DT1	SR·····T·G·H··T····Q·WT··-GVNRLANFGH·NC	0.04±0.01	0.07±0.01	0.07±0.01	0.7±0.1	0.07±0.01	0.035±0.009	0.05±0.01	0.06±0.02	0.01±0.001	0.011±0.003	0.007±0.002	0.7±0.1
Enterocin NKR-5-3C_S12T, T22A	AT····L···TK····E··I·G··L·QY·····L···VP··C	0.04±0.03	>1	0.08±0.02	>1	0.028±0.007	0.08±0.02	0.25±0.06	0.1±0.1	0.01±0.01	0.06±0.04	0.028±0.007	>1
Carnocin DT1	ST·····S·T·K··S····QSWTE-GVQRW·DHLF·	0.06±0.04	0.25±0.06	0.1±0.1	>1	0.01±0.01	0.01±0.01	0.01±0.002	0.01±0.01	0.001±0.001	0.003±0.002	0.003±0.001	>1
Enterocin NKR-5-3C	AT·····L···SK····E··ITG··L·QY·····L···VP··C	0.06±0.02	0.4±0.1	0.1±0.03	>1	0.017±0.005	0.09±0.03	0.1±0.05	0.08±0.03	0.014±0.004	0.016±0.006	0.02±0.005	>1
Enterocin NKR-5-3C_S12A, E18D	AT·····L···AK····D··ITG··L·QY·····L···VP··C	0.06±0.04	0.4±0.2	0.1±0.1	>1	0.028±0.007	0.04±0.03	0.04±0.03	0.08±0.02	0.01±0.002	0.02±0.01	0.01±0.01	>1
Enterocin NKR-5-3C_LM2	·····L···SK····E··ITG··L·QY·····L···VP··C	0.08±0.02	0.7±0.2	0.2±0.1	>1	0.028±0.007	0.2±0.1	0.04±0.03	0.25±0.06	0.007±0.005	0.02±0.01	0.01±0.002	>1
Enterocin NKR-5-3C_I21C, A27G	AT·····L···SK····E··CTG··LGQY·····L···VP··C	0.08±0.02	>1	0.25±0.06	>1	0.25±0.06	0.4±0.2	0.5±0.2	0.5±0.2	0.08±0.02	0.25±0.06	0.2±0.1	>1
Enterocin P_G6D, K15G	··RS·D·····NSG·····EN··G·V·S···S·-···M··	0.1±0.03	0.18±0.05	0.25±0.04	0.7±0.1	0.08±0.01	0.12±0.04	0.12±0.04	0.15±0.04	0.25±0.04	0.4±0.1	0.5±0.1	>1
Enterocin P_LM2	··S·····NS·····EN··G·V·S···S·-···M··	0.1±0.1	0.25±0.06	0.25±0.06	>1	0.1±0.1	0.4±0.2	0.2±0.1	0.4±0.2	0.1±0.1	0.2±0.1	0.2±0.1	0.7±0.2
Enterocin NKR-5-3C_LM4	·····L···SK····E··ITG··L·QY·····L···VP··C	0.1±0.1	>1	0.4±0.2	>1	0.028±0.007	0.25±0.06	0.08±0.02	0.4±0.2	0.02±0.01	0.06±0.04	0.02±0.01	>1
Lactocin DT2	·····H·G·YS·T·D··T·I·N·GNN·AAN··T·GN··WNK	0.2±0.1	0.25±0.06	0.4±0.2	>1	0.5±0.2	0.25±0.06	0.7±0.2	0.1±0.1	0.02±0.01	0.04±0.03	0.01±0.002	>1
Lactocin DT3	SR·····IT·G·H··T····Q·WT··-GVNRLANFGH·NC	0.2±0.1	0.5±0.2	0.25±0.06	>1	0.25±0.06	0.2±0.1	0.4±0.2	0.25±0.06	0.06±0.04	0.04±0.03	0.04±0.03	>1
Enterocin P	··RS·····NS·····EN··G·V·S···S·-···M··	0.25±0.04	0.4±0.1	0.25±0.04	>1	0.28±0.09	0.3±0.1	0.18±0.05	0.4±0.1	0.21±0.04	0.4±0.1	0.4±0.1	>1
Enterocin DT1	T·····TQ····D·SR·RSETVDRGVKAYVN·FTKVL·G····	0.4±0.1	0.5±0.1	0.6±0.1	>1	0.25±0.04	0.4±0.1	0.4±0.1	0.5±0.1	0.1±0.03	0.12±0.04	0.1±0.03	>1
Curvacin A	ARS·····NK·····R···TQS·IGGM·S···S·-···M	0.4±0.2	0.7±0.2	0.7±0.2	>1	0.5±0.2	0.7±0.2	0.7±0.2	0.7±0.2	0.25±0.06	0.5±0.2	0.5±0.2	>1
Enterocin DT2	T·····TQ····D·SR·RSE·VDRGVKAYVN·FTKVL·G····	0.5±0.2	0.7±0.2	0.7±0.2	>1	0.7±0.2	0.7±0.2	0.7±0.2	0.7±0.2	0.007±0.005	0.01±0.01	0.02±0.01	>1
Divercin V41_Y9S, K13T	·····S··ST····D··Q·S···GQTVV···L···IP··C	0.5±0.2	>1	>1	>1	0.7±0.2	0.5±0.2	0.7±0.2	0.7±0.2	0.08±0.02	0.25±0.06	0.1±0.1	>1
Enterocin DT3	T·····S·G·TT·T·D·AA·GT···Q·S···FL···IP··C	0.7±0.2	0.7±0.2	>1	>1	0.4±0.2	0.5±0.2	0.7±0.2	0.5±0.2	0.25±0.06	0.4±0.2	0.5±0.2	>1
Enterocin DT4	·D··I···NS·····EN··G·V·S···S·-···M··	0.7±0.2	>1	0.7±0.2	>1	0.7±0.2	0.7±0.2	0.7±0.2	0.7±0.2	>1	0.4±0.2	0.7±0.2	0.7±0.2
Avicin A	T·····S···KG·S·D··K·ISI·GNNSAANL·T·GA··WKS	0.7±0.2	>1	0.7±0.2	>1	>1	0.7±0.2	0.7±0.2	0.7±0.2	0.02±0.01	0.04±0.03	0.002±0.001	>1
Plantaracin 423	·····T·G·HS·S····Q·FS·SVSH-LANFGH·KC	0.7±0.2	>1	>1	>1	0.25±0.06	0.7±0.2	0.7±0.2	0.7±0.2	0.028±0.007	0.06±0.04	0.04±0.03	>1
Enterocin NKR-5-3C_G20V, I31M	AT·····L···SK····E·VITG··L·QY·M···L···VP··C	0.7±0.2	>1	>1	>1	0.5±0.2	0.7±0.2	0.7±0.2	0.5±0.2	>1	0.7±0.2	0.7±0.2	>1
Sakacin P_Y2V, G4D, N5V, N24C, I25G, N27E	·V·DV··H·GEHS·T·D··T·I··GGEN·AAN··T·GN··WNK	>1	0.7±0.1	0.7±0.1	>1	0.7±0.1	0.5±0.1	0.6±0.1	0.6±0.1	0.5±0.1	0.4±0.1	0.6±0.1	>1
Leucocin A	·····H·T·SG·S·····FSA-GVHRLANGGN·FW	>1	0.7±0.2	>1	>1	0.7±0.2	0.5±0.2	0.7±0.2	0.5±0.2	0.01±0.01	0.04±0.03	0.06±0.04	>1
Mesentericin Y105	·····H·T·SG·S·····ASA-GIHRLANGGN·FW	>1	>1	0.5±0.2	>1	0.4±0.2	0.25±0.06	0.4±0.2	0.5±0.2	0.06±0.04	0.25±0.06	0.25±0.06	>1
Leucocin A_S15R	·····H·T·SG·R·····FSA-GVHRLANGGN·FW	>1	>1	0.7±0.1	>1	0.6±0.1	0.4±0.1	0.5±0.1	0.7±0.1	0.035±0.009	0.1±0.03	0.08±0.01	>1
Enterocin P_Y5C, K15H	··RSC·····NSH·····EN··G·V·S···S·-···M··	>1	>1	0.7±0.2	>1	0.7±0.2	0.5±0.2	0.7±0.2	>1	>1	>1	>1	>1
Enterocin P_N7V, Y10E	··RS··V··E··NS·····EN··G·V·S···S·-···M··	>1	>1	0.7±0.2	>1	0.7±0.2	0.7±0.2	0.7±0.2	0.7±0.2	>1	0.7±0.2	>1	>1
Enterocin P_S4C, S33D	··RC·····NS·····EN··G·V·D···S·-···M··	>1	>1	>1	>1	0.7±0.2	>1	>1	0.7±0.2	0.2±0.1	0.7±0.2	0.25±0.06	>1
Inactive Controls													
Lactocin DT4	RYH·····S··RYH·R·D·SRSWY··VNR·G·AY·T·GQ·TI·NC	>1	>1	>1	>1	>1	0.7±0.2	0.7±0.2	>1	>1	>1	>1	>1
Enterocin DT5	·····T·N··T·D·AK·TT···GMS···FL···IP··	>1	>1	>1	>1	>1	>1	>1	>1	>1	>1	>1	>1

Figure 2.2 Summary of MIC data identifies AMPs capable of selective activity towards Listeria.
MID values represent the lowest fraction of AMP AS precipitate solution that results in growth inhibition. Blue to red coloring represents most to least potent MID values. AMP names highlighted in gray are identified as selective. Dots in alignment show sequence agreement with consensus sequence.

While no AMPs displayed selective targeting of *E. faecium* strains, several AMPs displayed low activity to *E. faecium* while retaining high activity to several *E. faecalis* and/or *L. monocytogenes* strains (Figure 2.2). Three AMPs displayed activity to seven of eight *E. faecalis* and *L. monocytogenes* strains with limited activity to only one *E. faecium* strain. Six AMPs displayed at least a four-fold increase in potency to the three susceptible *L. monocytogenes* strains relative to any other strains tested, while a natural enterocin variant, enterocin DT2, displayed at least a 25-fold increase in potency to the 3 susceptible *L. monocytogenes* strains over all other strains tested. Of note, only 2 gain-of-activity variants were observed: enterocin $P_{G6D, K15G}$ and enterocin P_{LM2}. To identify class IIa bacteriocin sequence features that affect specificity, we analyzed these selective variants for sequence motifs in the C-terminal domain, as previous literature suggests the C-terminal domain may act as a targeting domain for specific manPTS genes [94,106]. Interestingly, only two active AMPs contained the C-terminal sequence motif GGFGGR, both of which had higher activity towards the three susceptible *L. monocytogenes* strains than the seven susceptible enterococci strains. In comparison, 18 of 32 active AMPs contained either the GGA(I/V)PGKC or GLAGMGH C-terminal sequence motifs, and these AMPs commonly had comparable MIDs across many strains of different species (Figure 2.2). These trends support the hypothesis that C-terminal sequence motifs of class IIa bacteriocins play a strong role in determining their range of activity towards different species. However, it remains unclear if this trend is due to stronger/weaker interactions between these AMP sequence motifs and the

manPTSs, the cell membrane, or another factor specific to various bacterial species. Ultimately, most class IIa bacteriocins displayed comparable potency across most strains tested with limited selectivity (Figure 2.3A). However, when considering activity to species, some AMP variants display equivalently increasing potency and specificity (Figure 2.3B). The y = x line in Figure 2.3B is the limit of maximal selectivity given that the lowest observed activity was a MID of 1 (MID values of >1 were treated as 1 for selectivity calculations to eliminate infinite values). Several AMPs exhibit high potency while falling near the y = x line, exhibiting near the maximum measurable selectivity. These AMP variants suggest that class IIa bacteriocin activity can be tailored to different species.

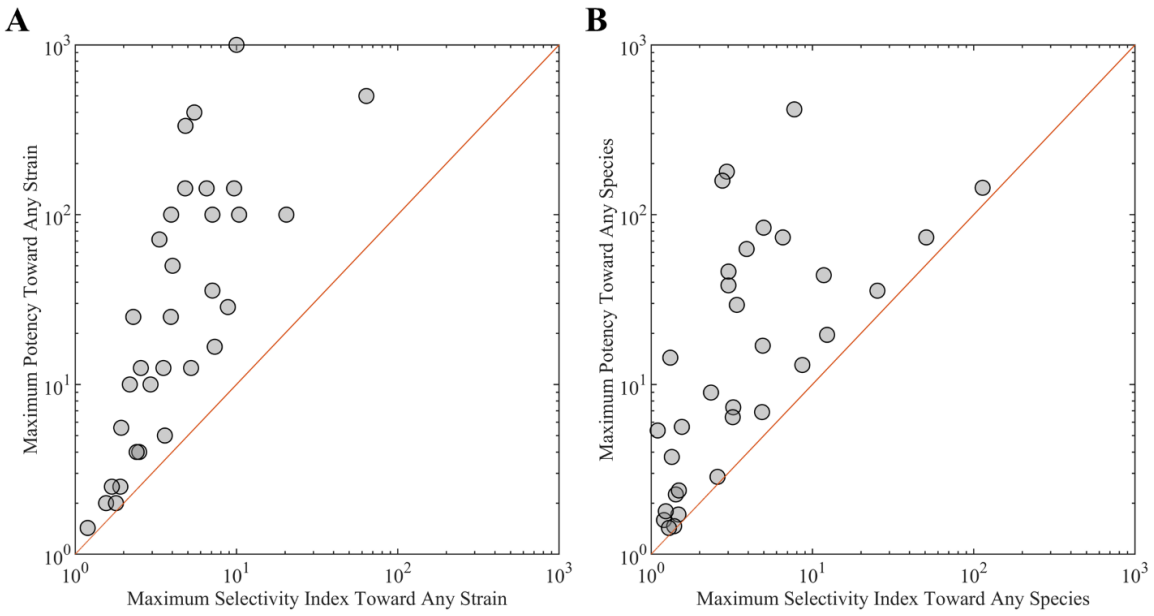


Figure 2.3 AMPs display minor selectivity at a species level but limited selectivity at a strain level.

(A) Maximum AMP potency towards any strain plotted against maximum selectivity towards any strain. All AMPs fall above the line y = x, showing a broadly increasing potency towards several strains with limited selectivity. (B) Maximum average potency towards any species plotted against maximum selectivity towards any

species. Some AMPs fall near the line y = x, displaying increasing potency with selectivity towards a given species.

2.4.4 Ridge regression of AMP sequence-activity data shows manPTS sequence is equivalently predictive as strain identity

Beyond using this broad analysis of class IIa bacteriocin sequence space to identify compelling AMPs, we aimed to advance the understanding of sequence-activity relationships. We hypothesized that a generalized linear model could identify such relationships to predict AMP performance. To test this hypothesis, the activity of all 210 observed AMPs was ridge regressed to AMP sequence in four separate models. Ridge regression was used to minimize overfitting and because initial ridge regression models were shown to offer improved prediction over lasso regression models. The four different models were chosen to test which set of information best trains a ridge regression model and were assessed via a receiver operating characteristic (ROC) curve. In the first model, AMP sequences were regressed to their binary activity towards all indicator strains; *i.e.* if an AMP had activity towards at least one indicator strain, it was classified as active. This model performed only moderately above random, with an area under the ROC curve (AUC) of 0.60±0.02 (Figure S2.4). The second set of models (one for each strain) independently evaluated activity for each strain; a near-equivalent average AUC of 0.59 ± 0.01 indicates that independent strain-specific modeling does not aid predictive power. We then evaluated a model in which strain identity was encoded as an input along with sequence, and all data were jointly modeled. This approach elevated the AUC to 0.70 ± 0.04 ($p = 0.04$ versus model 2), which reveals that strain-specific information is useful in distinguishing AMP activity toward the

different strains tested. The fourth model replaced strain identity by manPTS sequence encoding, which did not further elevate model performance (AUC = 0.69 ± 0.04). Thus, specific manPTS sequence information provided value equivalent to strain identity. The equivalent strength of manPTS sequence information is consistent with its hypothesized dominant role in dictating susceptibility to class IIa bacteriocins. While moderate predictive performance was achieved by models 3 and 4, no obvious trends in charge, hydrophilicity, or polarity of beneficial amino acids at certain positions were identified.

2.4.5 ManPTS sequence does not fully define susceptibility to class IIa bacteriocins

While most class IIa bacteriocins displayed limited selectivity, new molecules were characterized which displayed selective activity to *Listeria* strains, suggesting that class IIa bacteriocin activity can be tailored to certain species. Therefore, we sought to identify any trends in susceptibility to class IIa bacteriocins between species in hopes of elucidating the underlying mechanisms by which selectivity is achieved. We initially compared manPTS sequences between enterococci strains and *L. monocytogenes* strain ATCC 51775, given class IIa bacteriocins are known to interact with manPTS EIIC and EIID domains [71,72]. The manPTS genes from *L. monocytogenes* strain ATCC 51775 were used for this analysis as this is the only tested *Listeria* strain with available genomic sequence data. More significant differences in manPTS sequences between *L. monocytogenes* and enterococci than between *E. faecium* and *E. faecalis* were observed which may contribute to increases in *Listeria* susceptibility to class IIa AMPs (Figure 2.4). However, given

manPTS sequences were only used from one *Listeria* species, more data points are necessary to confirm these trends.

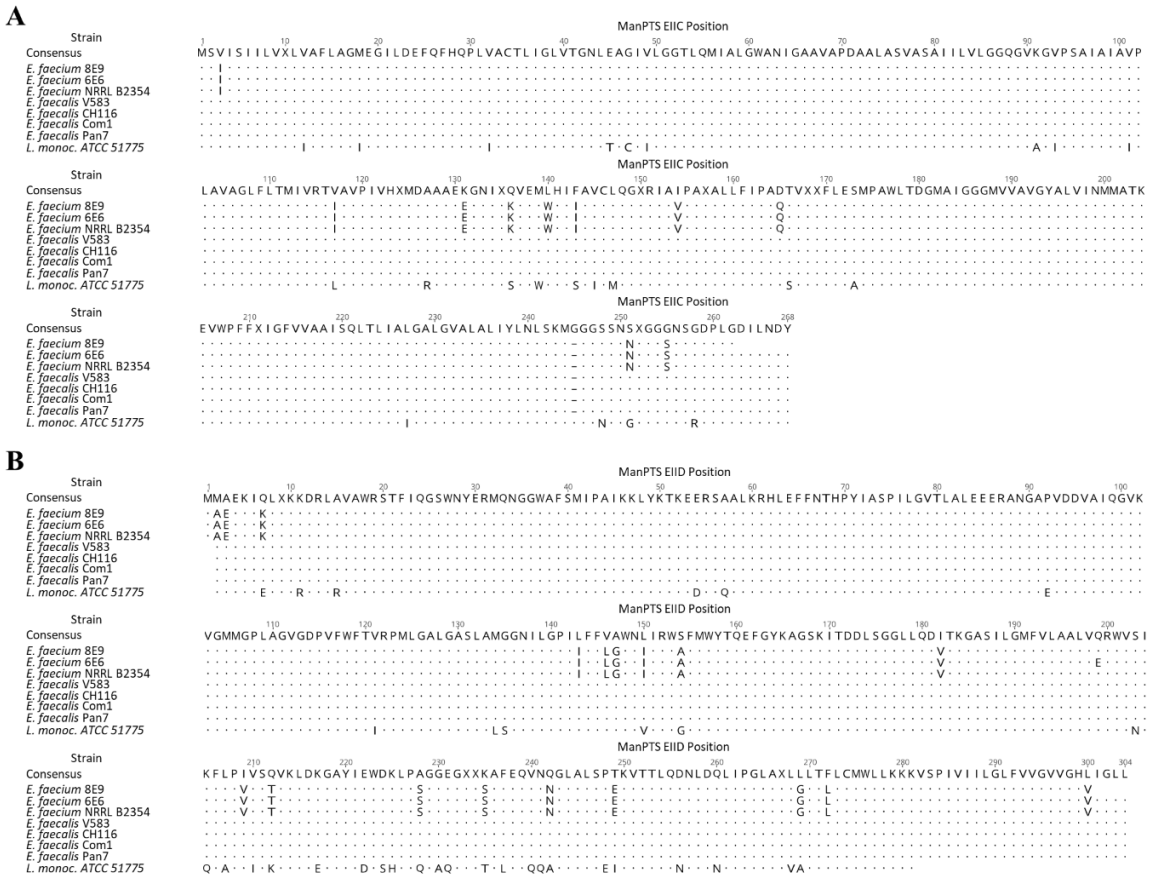


Figure 2.4 Alignment of manPTS EIIC (A) and EIID (B) sequences. The genes for *E. faecium* 8E9 were amplified using primers designed from the *E. faecium* 6E6 manPTS sequences (Table S2.3) and Sanger sequenced. All other manPTS genes were identified from genomic information. Dots in alignment show sequence agreement with consensus sequence.

To further assess the specificity of AMPs across bacterial strains and species, the MID values for all active AMPs were pairwise compared across all tested strains (Figure 2.5). Linear trends appear when comparing the MIDs between strains of the same species (Figure 2.5, Figure S2.5). Given nearly all strains within a particular species share nearly identical manPTS sequences

(Figure 2.4), this is not surprising. However, for some strains with identical manPTS sequences, susceptibility could still differ by an order of magnitude. Examples of this are the MIDs of AMP enterocin NKR-5-3C$_{LM4}$ to the four *E. faecalis* strains, which range from 0.028 to 0.4 despite identical manPTS sequences. More telling is the comparison of activity of several AMPs to *E. faecium* strains 8-E9 and NRRL B-2354, where strain 8-E9 is highly susceptible to class IIa bacteriocins while strain NRRL B-2354 appears tolerant to nearly all class IIa bacteriocins despite having identical manPTS sequences.

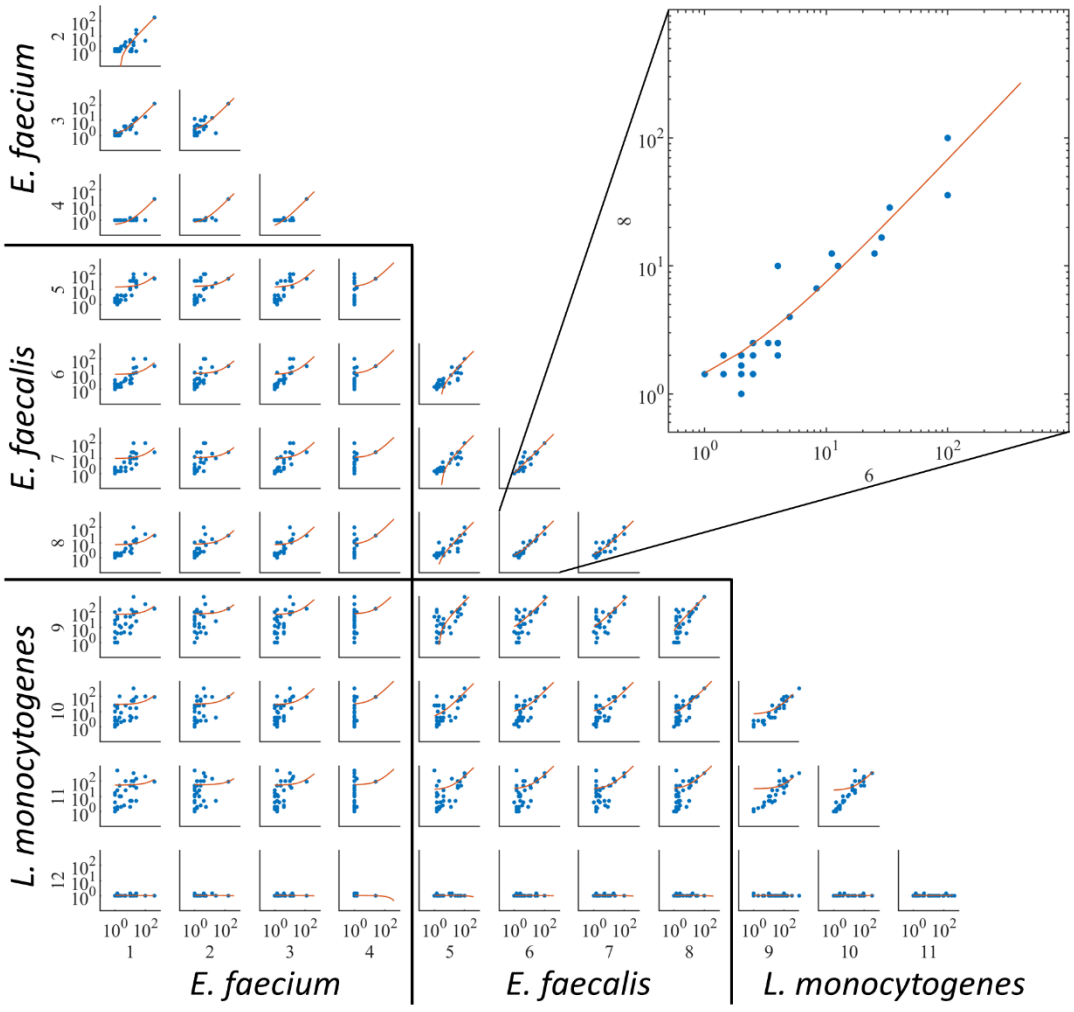


Figure 2.5 Activity of all AMPs plotted between each individual strain shows susceptibility trends between species.

The MICs of each AMP in bacteriocin units was plotted between all individual strains. A bacteriocin unit is defined here as the inverse of the lowest active fraction of AMP AS precipitate solution. One-to-one linear trends appear between most strains of the same species, shown in the subgroups near the diagonal. The inlet shows an example plot of activity of all 32 AMPs towards strains 6 (x-axis) and 8 (y-axis). Given both strains 6 and 8 are *E. faecalis* strains, a strong trend similar to y=x appears. Plots for strains 4 and 12 fall nearly vertical or horizontal due to the low susceptibility observed for these strains. Strains by number: 1: *E. faecium* 8E-9, 2: *E. faecium* 6E-6, 3: *E. faecium* 7A, 4: *E. faecium* NRRL B2354, 5: *E. faecalis* V583, 6: *E. faecalis* CH116, 7: *E. faecalis* Pan7, 8: *E. faecalis* Com1, 9: *L. monocytogenes* ATCC 51775, 10: *L. monocytogenes* M-03-1213B-1, 11: *L. monocytogenes* CDC 7762, 12: *L. monocytogenes* V7. Figure S2.5 shows circle plot displaying correlation coefficients and slopes of linear fit of all sub plots.

When comparing susceptibilities between different species, the tested *E. faecalis* strains (5-8) appear to have larger correlation in susceptibility with *L. monocytogenes* strains (9-12) than *E. faecium* strains (1-4) (Figure S2.5). This trend is observed despite *L. monocytogenes* strain ATCC 51775 manPTS EIIC and EIID sequences having only 88% and 80% sequence similarity with *E. faecalis* manPTS EIIC and EIID sequences, respectively (Figure 2.4, Figure S2.6). Conversely, the minor correlation in susceptibility to class IIa bacteriocins observed between *E. faecium* strains 1-3 and *E. faecalis* strains 5-8 would be expected to be higher given their manPTS EIIC and EIID sequence similarities of 93% and 92%, respectively. *E. faecium* and *L. monocytogenes* strains show very little correlation in susceptibility to class IIa bacteriocins and have 89% and 78% similarity in manPTS EIIC and EIID sequences, respectively. These trends clearly demonstrate that factors other than manPTS sequence significantly impact susceptibility to class IIa bacteriocins.

2.4.6 C-terminal disulfide-containing class IIa bacteriocins are compelling for *in vivo* application

Towards identifying compelling molecules for further study and evaluation *in vivo*, we evaluated the proteolytic stability of eight of the most potent AMPs against trypsin, chymotrypsin, and pepsin (Figure 2.6). While class IIa bacteriocins are known to be very thermally stable, proteolytic stability is a common limitation of peptide therapies for *in vivo* applications [86,87]. Trypsin, chymotrypsin, and pepsin were selected because they are highly prevalent in the human digestive system [107]. The eight AMPs were incubated briefly with varying concentrations of the given protease, heated to inactive the protease, and tested for loss of activity compared to an untreated control in an agar diffusion assay. The eight AMPs selected were chosen as a sampling of potent natural and variant AMPs with high potency towards *L. monocytogenes* strain ATCC 51775. This strain was selected for this assay as it was observed to be the most susceptible to class IIa bacteriocins during MID quantification.

AMP	Sequence	K_{cat}'/K_m			MIC[a]	Number of Cleavage Sites[b]		
		Tryp.	Chym.	Pep.		Tryp.	Chym.	Pep.
Enterocin NKR-5-3C	ATYYGNGLYCNSKKCWVEWGITGGCLAQYAIGGWLGGAVPGKC	<0.009	11.6±3.1	1.7±0.7	0.014±0.004	3	7	15
Enterocin NKR-5-3C$_{LM2}$	TKYYGNGLYCNSKKCWVEWGITGGCLAQYAIGGWLGGAVPGKC	0.2±0.2	0.6±0.5	<0.018	0.007±0.005	4	7	14
Enterocin NKR-5-3C$_{LM4}$	KYYGNGLYCNSKKCWVEWGITGGCLAQYAIGGWLGGAVPGKC		0.2±0.2	<0.018	0.02±0.01	4	7	14
Enterocin NKR-5-3C$_{S12T,\ T22A}$	ATYYGNGLYCNTKKCWVEWGIAGGCLAQYAIGGWLGGAVPGKC	<0.009	3.5±1.9	3.9±1.4	0.01±0.01	3	7	15
Enterocin NKR-5-3C$_{S12A,\ E18D}$	ATYYGNGLYCNAKKCWVDWGITGGCLAQYAIGGWLGGAVPGKC	<0.009	54.8±44.5	2.4±2.2	0.01±0.002	3	7	15
Divercin V41	TKYYGNGVYCNSKKCWVDWGQASGCIGQTVVGGWLGGAIPGKC	<0.009	0.04±0.04	<0.018	0.003±0.001	4	6	9
Hiracin JM79	ATYYGNGLYCNKEKCWVDWNQAKGEIGKIIVNGWVNHGPWAPRR	1.4±0.9	9.7±5.9	8.5±2.6	0.006±0.002	6	7	11
Carnocin DT1	STYYGNGVSCTKKKCSVNWGQSWTEGVQRWGDHLFG		21.8±15.4	0.4±0.3	0.001±0.001	4	6	10

Figure 2.6 k_{cat}'/K_m values of trypsin, chymotrypsin, and pepsin against tested AMPs.

[a]MIC values are those listed in Figure 2.2; [b]All protease sites were identified with PeptideCutter [89]; all k_{cat}'/K_m values are in units of mM_{enzyme}^{-1} mM_{AMP}^{-1} s^{-1}; values for gray boxes could not be quantified due to noise in the assay results. Blue to red coloring represents most to least stable k_{cat}'/K_m values.

Trypsin k_{cat}'/K_m values could only be robustly determined for six of the eight tested AMPs due to experimental noise in the assay. Of these, four AMPs saw no loss in activity across the tested enzyme concentrations (≤ 256 μM); thus, the normalized proteolytic efficiency was <0.009 $mM_{enz}^{-1}mM_{AMP}^{-1}s^{-1}$. The LM2 variant of enterocin NKR-5-3C, which introduces an additional trypsin site via A1T, T2K mutation, exhibits appreciable susceptibility. Hiracin JM79, which lacks a C-terminal disulfide, is even more susceptible. For chymotrypsin and pepsin, the parental enterocin NKR-5-3C had a greater susceptibility than both N-terminal chimeric variants tested (LM2 and LM4). All three variants had the same number of identified chymotrypsin cleavage sites, so this would not explain the observed differences. Additionally, the parental peptide and LM2 variants have identical chymotrypsin cleavage sites, so it is unclear what sequence features enabled higher proteolytic tolerance. The two other enterocin NKR-5-3C variants, S12T, T22A and S12A, E18D, show similar proteolytic susceptibility to the parental. Across all three proteases, divercin V41, which contains two disulfide bridges, had the highest proteolytic stability. Amongst variants, both N-terminal chimeric enterocin NKR-5-3C variants displayed the highest stabilities across the tested proteases. Thus, all three AMPs are compelling molecules for further evaluation *in vivo* given their high potencies and stabilities.

2.5 Discussion

The results shown here expand our database of characterized AMPs for analysis, identify compelling molecules for further study and optimization, demonstrate the utility of statistical modeling in AMP engineering efforts, and highlight the

significance of factors beyond manPTS receptors in determining bacterial susceptibility to class IIa AMPs. Although only a small fraction of our initial library was isolated, 210 unique class IIa bacteriocins were evaluated for inhibitory activity to six enterococci strains in a *L. lactis*-secretion agar-diffusion assay. While several of these AMPs were previously characterized, many more were novel variants identified from genomic data or variants containing multiple amino acid mutations from a parental sequence. The activity data from random variants suggest that random mutagenesis of class IIa bacteriocins should limit the number of simultaneous amino acid mutations in situations where throughput is limited to two or three residues in any variant to increase the likelihood of finding functional variants. Additionally, the enhanced selectivity achieved by double mutants of enterocin P (N7V, Y10E) and enterocin NKR-5-3C (G20V, I31M) may suggest that a more efficient route for generating AMPs with selective activity will be to introduce selectively deleterious amino acid mutations into highly potent AMPs to abolish activity to off-target strains. As for chimeras, our incomplete sampling of chimeric class IIa variants hinders firm conclusions; however, the high success rate, albeit among only six observations, and previous literature utilizing chimeric class IIa bacteriocins suggests that (1) chimeric class IIa bacteriocins are promising stepping stones to generate improved-activity variants [78,108,109] and (2) class IIa bacteriocins may be segmented into domains which each serve distinct functions, as has been previously hypothesized.

While screening was in part conducted to identify highly potent class IIa bacteriocins, a key focus of this study was the investigation of selective activity of

these AMPs. Ridge regression determined that inclusion of strain identity information was beneficial for predicting AMP performance, suggesting that differences in bacterial susceptibility could be predicted. The equivalency of predictive benefit from using manPTS sequences rather than strain identity information is consistent with a strong role for manPTS in determining bacterial susceptibility to class IIa bacteriocins, as was previously hypothesized. Yet, further analysis of bacterial susceptibility to class IIa bacteriocins showed that strains with identical manPTS sequences could have significantly different susceptibilities to the same AMP. While some correlation was observed between manPTS sequence and susceptibility to class IIa AMPs, there are clear outliers which do not agree (Figure S2.6). This suggests other factors, such as bacterial membrane composition or metabolism, play a more significant role in determining bacterial susceptibility to various AMPs than previously thought and should be further investigated in the future.

Activity and stability characterization of individual AMPs in the current study aimed at identifying compelling molecules to further pursue for development as *in vivo* therapeutics. We sought to identify AMPs capable of selectively targeting the individual species of *E. faecium*, *E. faecalis*, or *L. monocytogenes* based on the hypothesis that such AMPs may limit negative effects associated with off-target activity to commensal bacteria of other species. Yet, most active AMPs displayed broad activity profiles towards all tested strains due to our reduced sampling of our AMP library and the broad activity of many previously characterized class IIa AMPs which seeded our library. We suspect that given the narrow range of commensal

bacteria which possess the necessary manPTS genes encoding the target receptors for class IIa bacteriocins, the potent AMPs identified here are still compelling as *in vivo* therapies which limit activity to most commensal bacteria. Notably, however, class IIa bacteriocins with selective activity towards *Listeria* were also identified, which could further reduce activity to commensal enterococci *in vivo*. For future work developing proteolytically stable class IIa bacteriocins for *in vivo* efficacy, evaluations should focus on those containing C-terminal disulfides which reduce accessibility of common protease cleavage sites, and we have identified divercin V41 as one compelling molecule with increased potency and stability over other tested class IIa bacteriocins.

2.6 Materials and Methods

2.6.1 Bacterial culture and strains

Lactococcus lactis NZ9000 cells were grown in brain heart infusion (BHI) medium (W.W. Grainger Inc.), which contained 1.6% (vol/vol) agar in the case of solid phase growth. Cultures were grown stationary at 30°C. *E. coli* cells were grown in lysogeny broth (LB, Fisher BioReagents) in liquid, which contained 1.6% (vol/vol) agar in the case of solid phase growth. All *E. coli* cultures were grown at 37°C, with liquid cultures shaking at 250 RPM. When specified, LB was supplemented with 5 µg/mL chloramphenicol. All enterococci and *Listeria* strains were grown in liquid BHI medium at 37°C with shaking at 250 RPM. All bacterial strains and plasmids used are listed in Table S2.1.

2.6.2 Class IIa bacteriocin library design

The class II bacteriocin protein family full alignment from the Pfam database [110] was used to seed a search in Jackhmmer [111] to identify class IIa bacteriocin sequences in the UniProt database [88]. There was no taxonomic restriction, and iterations were performed with the hit threshold set to an E value of 0.01 until convergence to identify the maximum number of homologous sequences. The output of this search was parsed to eliminate duplicate protein sequences as well as sequences shorter than 30 amino acids or longer than 50 amino acids. A total of 150 remaining sequences were included in the library of class IIa bacteriocins. Sequence identifiers for natural AMPs are the UniProt entry IDs (Table S2.2). Natural AMPs which have been previously characterized are also identified in the text and Table S2 with their given names. All 150 class IIa bacteriocins from the UniProt database were included in library 1.

Enterocin A, enterocin P, enterocin NKR-5-3C, divercin V41, sakacin P_{K11E}, and pediocin PA-1_{D17N} were selected as seed sequences for the rational and random library designs; the first four AMPs were previously shown in literature to have moderate to high antimicrobial activity and, thus, were believed to be strong starting points for generating improved variants [78,92,93,95,112] whereas sakacin P_{K11E} and pediocin PA-1_{D17N} were chosen as inactive variants of the parental sakacin P and pediocin PA-1 for rational and random mutant library design to serve as negative controls and to evaluate the likelihood of finding mutations capable of restoring antimicrobial activity.

AMP variants were rationally designed to test several hypotheses regarding class IIa bacteriocin structure and function. AMP chimeras were designed by

swapping all unique N-terminal domains of all active seed sequences. A consensus [96–98] interior domain was designed from all 150 class IIa bacteriocins identified from UniProt, and chimeras were constructed with this interior domain and all unique N-terminal and C-terminal domains. Enterocin A variants were designed for improved proteolytic stability. PeptideCutter software (Swiss Institute of Bioinformatics) was used to identify sites in enterocin A with a high susceptibility to trypsin, chymotrypsin, and pepsin [89]. The most frequent susceptible residue identified was tyrosine, so all variants were designed such that all tyrosine residues, individually and combinatorically, were mutated to serine. All rational and chimera variants were included in library 1.

A library of 900 random multi-mutants were generated from the six seed sequences to evaluate the class IIa bacteriocin tolerance to random mutagenesis. The library was composed of 50 mutants containing two, four, or six simultaneous random mutations in the 33-amino acid interior region of each of the six seed sequences. As an example, to generate the group of variants containing two mutations in enterocin A, two positions from 1 to 33 were randomly chosen and mutated to a random amino acid that was not the parental residue. This process was repeated to generate 50 variants containing two amino acid mutations, 50 variants containing four amino acid mutations, and 50 variants containing six amino acid mutations for a total of 150 enterocin A random variants. The process was then repeated for the remaining five seed sequences for a total of 900 random variants. For the inactive sakacin P_{K11E} and pediocin PA-1_{D17N} variants, positions

11 and 17, respectively, were not allowed to be mutated during this process. The group of 900 random multi-mutants was defined as library 2.

2.6.3 Oligonucleotide library construction

AMP libraries 1 and 2 were synthesized as oligonucleotides by Twist Biosciences. The libraries were designed such that each library could be amplified through PCR independent of the other library using specific DNA primers. Following a 16-cycle PCR for initial amplification of each library, DNA encoding the Usp45 signal peptide was ligated upstream of the AMP, as this signal peptide has been shown to enable high secretion efficiency in *L. lactis* [100]. This construct was then amplified and assembled (HiFi, New England Biolabs) into the chloride-inducible pNZC expression vector [99]. Following assembly, the final DNA constructs were transformed into *E. coli* (5-alpha competent cells, NEB), and cells were grown overnight on LB-agar plates with 5 µg/L chloramphenicol. Following overnight growth, all cells were collected from agar plates and the DNA was extracted and stored at -20°C until further use. All primers used for library construction are included in Table S2.3.

2.6.4 96-well stock plate preparation

Final DNA constructs for libraries 1 and 2 were transformed into electrocompetent *L. lactis* prepared according to published procedure [113]. Following transformation, cells required roughly 48 hours of growth at 30°C for colonies to be visible on agar plates. Individual colonies were plucked from agar plates and inoculated into wells of deep 96-well plates containing 1 mL BHI. All plates contained two control wells that were inoculated with freezer stocks of *L. lactis*

expressing enterocin P (positive control) and empty pNZC vector (negative control), respectively. Plates were incubated for 18-24 hours at 30°C. Following growth to saturation, 100 µL of culture from each well was added to 100 µL of 60% glycerol in sterile 96-well plates to create a 30% glycerol stock plate. The plates were covered and stored at -80°C until further use. Five plates were prepared containing library 1 constructs and 10 plates were prepared containing library 2 constructs.

2.6.5 Illumina sequencing and well identification

Following 30% glycerol plate preparation, whole-cell PCR was conducted on 1 µL of culture from each well with primers that appended row, column, and plate indices adjacent to the coding region of each construct to identify AMP sequences through high-throughput sequencing. Nextera N501-N508 and N701-N712 index adaptors were used as row and column indices, respectively. Combinations of unique 5-base-pair and 4-base-pair sequences on the 5' and 3' ends of the coding region, respectively, were used as plate indices. PCRs were conducted in multiple 96-well PCR plates to allow for independent amplification of all wells using Q5 High-Fidelity DNA Polymerase (New England Biolabs). PCR products for all wells of a plate were mixed, purified (QIAquick PCR Purification Kit, Qiagen, Hilden, Germany) to reduce sample volume, and gel extracted. DNA pools were then sequenced on two Illumina iSeq 100 runs to identify AMP constructs present in each well of all glycerol stock plates. Sequencing on the Illumina iSeq 100 was conducted at the University of Minnesota Genomics Center. All primers used for DNA sequencing are included in Table S2.3.

Illumina iSeq sequencing generated approximately 7 million reads specific to libraries 1 and 2. Sequences were processed using Usearch by filtering for a max error rate of 0.001 and denoised using the *unoise3* command to correct single-base pair errors which may have occurred during sequencing [114,115]. A second processing step was conducted to identify the individual AMPs in each well of the glycerol stock plates.

For each sequencing run, the distribution of reads of AMP constructs per well was calculated, as some wells were poorly amplified during whole-cell PCR and, thus, were poorly represented in sequencing data. Thresholds of 90 reads and 10 reads for the first and second sequencing runs, respectively, were used to distinguish true AMP identification from noise (Figure S2.1). Unique sequences with reads above this threshold were isolated and analyzed via a custom MATLAB script. Sequencing primers annealed to the plasmid backbone before the signal peptide, so reads of plasmid containing no gene were ~40 nucleotides. Reads containing true AMPs would include this backbone region, the Usp45 signal peptide, and the AMP of interest, so a 75-nucleotide cutoff was selected to distinguish empty vector from true constructs. All wells where the most frequent unique sequence was ≤75 nucleotides and accounted for ≥60% of total reads were discarded, as these wells contained empty vector. Wells with two or more unique sequences of length ≥75 nucleotides but neither accounting for ≥50% of total reads in the well were also discarded as multi-construct wells. A well was identified as a single AMP construct if either of the following criteria were met:

i. If the most frequent unique sequence was ≥75 nucleotides and accounted for ≥50% of the total well reads, the well was identified as the most frequent unique sequence; or

ii. If the most frequent unique sequence was ≤75 nucleotides (empty vector), and the second most frequent unique sequence was ≥75 nucleotides and accounted for 40% of the total well reads, the well was identified as the second most frequent sequence.

Following identification of wells containing single constructs, DNA sequences were trimmed to remove the Usp45 signal peptide and were translated into amino acid sequences for further analysis. AMP sequences ≤10 amino acids were excluded from analysis, but all other sequences were retained, including out-of-library constructs. Following the individual AMP identification process, whole-cell PCR was conducted on 40 random wells, and the products were Sanger sequenced to validate the sequence identification process.

2.6.6 Agar diffusion activity assays to measure growth inhibition

Agar diffusion activity assays were conducted to test the inhibitory activity of every well of each glycerol stock plate against *E. faecium* 8-E9, *E. faecium* 6-E6, *E. faecalis* V583, *E. faecalis* CH116, *E. faecalis* Pan7, and *E. faecalis* Com1 (Table S2.1). One mL of BHI was added to each well of sterile, deep 96-well plates. Each well was then inoculated with cells from the corresponding well of a glycerol stock plate. Deep 96-well plates were incubated stationary for 18-24 hours at 30°C.

For making bacterial agar plates, 5 mL of BHI was inoculated with cells from 30% glycerol stocks of enterococci cultures to create starter cultures. Enterococci starter cultures were incubated for 18-24 hours at 37°C with shaking at 250 RPM. After overnight growth, BHI-agar mixture was prepared by adding 1.6% (wt/vol) agar to BHI medium and autoclaving. The mixture was allowed to cool to ~45°C and then enterococci culture was added at 0.05% (vol/vol) and mixed by inversion. Approximately 17 mL of the mixture was spread to a thin layer on fresh Petri dishes and allowed to cool for 30-60 minutes at room temperature. Once solidified, 3 µL of overnight culture from each deep 96-well plate was deposited onto each pathogen plate, allowed to dry, and incubated stationary for 18-20 hours at 37°C. Due to the size of agar plates, one 96-well plate had culture from wells split between 3 agar plates, as shown in Figure S2.4. Following overnight growth, colonies were washed off each plate with 5 mL sterile PBS for improved resolution of halo formation and size. Halo formation and size was then recorded relative to halo formation from the enterocin P positive control wells to give AMPs activity scores. Non-halo forming wells were scored as a 0, halos which were noticeably smaller than the positive control wells were scored as a 0.5, and halos which were comparable or larger than the positive control were scored as a 1 (Figure S2.2). The scoring of halo size relative to an enterocin P positive control, which was included on all plates, allowed for inter-plate comparisons.

2.6.7 Individual AMP Production and Ammonium Sulfate Precipitation

Unmodified AMPs were produced in *L. lactis* cultures and ammonium sulfate (AS) precipitated to generate more concentrated samples rather than through use of

purification tags which may affect structure or activity of small class IIa bacteriocins. Additionally, this method allowed for the activity testing of many class IIa bacteriocins with relatively low cost compared to chemical AMP synthesis. To produce individual AMP solutions, 40 mL of BHI was inoculated from glycerol stocks of AMP-producing or negative control, empty pNZC-containing *L. lactis* and incubated stationary overnight at 30°C. The culture was centrifuged at 3500x*g* for 5 minutes, the supernatant was discarded, and the pellet was resuspended in equal volume fresh BHI. The culture was then incubated stationary for 4 hours at 30°C. Following incubation, the culture was centrifuged at 3500x*g* for 5 minutes, and the supernatant was sterile filtered. AS was added to supernatant at 45% (wt/vol) to achieve a 70% saturated AS solution, which was rotated for 18 hours at 4°C. The AS solution was centrifuged for 10 minutes at 11000x*g*, and the pellet was resuspended in 1 mL ultrapure milliQ water and heat sterilized at 98°C for 10 minutes. The resulting AS precipitation solutions were stored at -20°C until further use.

2.6.8 Agar Diffusion Activity Assays to Determine Total AMP Inhibitory Activity

Agar diffusion assays were conducted to determine the total inhibitory activity of AMP AS precipitate solutions against eight enterococci strains and four *Listeria* strains (Table S2.1). Agar plates containing the indicator cultures were created as described previously. Fresh aliquots of AMP and pNZC precipitate solutions were thawed overnight at 4°C prior to being used. After thawing, a threefold dilution series of AMP precipitate solution in pNZC precipitate solution was prepared in a

deep 96-well plate, and 5 µL of each dilution was plated on indicator agar plates in triplicate. Agar plates were incubated stationary for 18-24 hours at 37 °C, after which halo formation was identified and recorded to determine AMP inhibitory activity. To quantify total inhibitory activity, we defined the minimum inhibitory dilution (MID) as the lowest dilution of resuspended AS precipitation solution that inhibited growth. This unitless metric of total activity is the product of an AMP's ability to be produced by *L. lactis* and inhibit growth, which is directly meaningful to intended applications with probiotic delivery. Following determination of the lowest fraction of AMP solution that resulted in halo formation for all replicates, the MID was statistically calculated given the true MID is between the observed MID and the next lowest tested AMP concentration, as described previously for statistically determining minimum inhibitory concentrations [39,116].

2.6.9 Proteolytic Stability Assay

AMP AS precipitate solutions were treated with varying concentrations of trypsin, chymotrypsin, and pepsin (Sigma-Aldrich, T1426, C1429, P6887) to determine the AMP proteolytic susceptibility to relevant proteases. Trypsin and chymotrypsin dilutions were prepared in 0.1 mM HCl solutions and were incubated with AMP AS precipitate solutions at 25 °C. Pepsin dilutions were prepared in 3 mM HCl solutions and were incubated with AMP AS precipitate solutions at 37 °C. AMP AS precipitate solutions were incubated at the specified temperatures for 5 minutes as a 1:1:1 mixture with the protease solutions at various concentrations and a pH buffer to achieve a final solution pH within the activity range for the specified proteases. Trypsin and chymotrypsin samples were mixed with 1 part of 51 mM

NaOH, 80 mM glycine at pH 10 to achieve a final mixture pH of ~7.5. Pepsin samples were mixed with 1 part of 20 mM HCl, 100 mM KCl at pH 2 to achieve a final mixture pH of ~3.5. The samples were heated to 98 °C for 20 minutes to inactivate the proteases, and remaining AMP activity was determined using the agar diffusion assays as described previously using *L. monocytogenes* ATCC 51775 as the indicator strain. Samples containing no protease were also included as a positive control.

To quantify activity of all tested samples, ImageJ software was used to analyze images of the agar diffusion plates. The average brightness of each halo, excluding the interior point made by the pipette tip, was measured using ImageJ. Brightness values were locally normalized to the mean brightness of four points at the corners of each halo (Figure S2.3). The inverse of locally normalized brightness values was then globally fit to Equation 1, which is derived from the Michaelis-Menten equation assuming low substrate (Supplemental derivation), using the *fitnlm* function on Matlab. The inverse of normalized brightness values was used as darker halos with lower brightness values correlate with high activity. Global fitting of triplicate data was used to calculate standard error in the parameter estimates.

$$B = B_{min} + (B_{max} - B_{min}) exp\left(-\frac{k_{cat}'[E]_0 \Delta t}{K_m}\right) \quad (1)$$

B in equation 1 represents brightness. Reported k_{cat}'/K_m values are normalized to initial substrate concentration and, thus, include a [mM substrate]$^{-1}$ term.

2.6.10 Bacteriocin IIa AMP and ManPTS Sequence Analysis and Modeling

ManPTS EIIC and EIID sequences were extracted from genomic sequences of strains *E. faecium* 6E6 and *E. faecalis* V583, CH116, Com1, and Pan7 from NCBI[117]. Primers were designed from the sequence of the *E. faecium* 6E6 manPTS (Table S2.3) and used to amplify the gene from *E. faecium* 8E9 using colony PCR. The gene fragment was then Sanger sequenced. Independently, class IIa bacteriocin and manPTS sequences were aligned using the *multialign* function on Matlab with default settings. All 210 bacteriocin IIa sequences identified during evaluation of libraries 1 and 2 were included. For sequence modeling, sequences were one-hot encoded and all positions that were conserved across all observations were eliminated to minimize the size of the final one-hot encoded matrix.

Sequence-activity data was modeled using the *lassoglm* function on Matlab with five-fold cross-validation, lambda values ranging from 10^{-3} to 10^{3}, and assuming activity scores were drawn from a binary distribution. Alpha values were set to 0.01 to use ridge regression due to overfitting concerns. Activity data was either the binary result of whether an AMP had activity towards at least one strain (210 data points) or the binary result of whether an AMP had activity to a particular indicator strain (210 AMPs and 6 indicator strains for a total of 1260 data points). The datasets were randomly split into 10 equal partitions. Models were iteratively trained on nine partitions and evaluated on the tenth until all partitions had been evaluated (10-fold cross validation). Reported RMSE and positive prediction values are based on the predictions for all 10 evaluation data sets for a given model. Models were trained and evaluated on the following datasets:

i. The reduced one-hot encoded AMP sequences fit to their binary activity towards all indicator strains, *i.e.* if an AMP is active towards at least one indicator strain, it is deemed active (210 data points);

ii. The reduced one-hot encoded AMP sequences fit to their binary activity towards each individual indicator strain (6 models, each with 210 data points)

iii. The reduced one-hot encoded AMP sequences and a 6-column, strain-indicator matrix fit to their binary activity towards each individual strain (1260 data points); and

iv. The reduced one-hot encoded AMP and manPTS sequences fit to their binary activity towards each individual strain (1260 data points).

ManPTS sequence similarities between species were calculated between the manPTS EIIC and EIID sequences from *E. faecium* NRRL B2354, *E. faecalis* V583, and *L. monocytogenes* ATCC 51775. *E. faecium* NRRL B2354 and *E. faecalis* V583 manPTS genes are identical or nearly identical to all other available manPTS genes from strains used in this work of the same species. Given *L. monocytogenes* ATCC 51775 is the only *Listeria* strain with available manPTS genes, it was assumed this is representative of all four strains tested, given the high homology observed in *E. faecium* and *E. faecalis*.

2.7 Acknowledgements

This work was supported by a grant from the National Institute of Health (R01 GM121777). We thank Prof. Patricia Ferrieri and Prof. Gary Dunny of the

University of Minnesota, and the NRRL ARS Culture Collection for their donations of *Enterococci* strains used in this study. We also thank Prof. Francisco Diez-Gonzalez of the University of Georgia for his donation of *Listeria* strains used in this study. We thank Alex Golinski and Dr. Seth Ritter for their suggestions regarding experimental methods and analysis. Support from the University of Minnesota Genomics Center and the Minnesota Supercomputing Institute at the University of Minnesota is gratefully acknowledged.

2.8 Supplemental Information

2.8.1 Derivation of equations for fitting protease data

Begin with Michaelis-Menten kinetics assuming low substrate (AMP) concentration:

$$v = \frac{\partial S}{\partial t} = -\frac{k_{cat}[E]_0[S]}{K_m} \tag{S2.1}$$

Solve for substrate concentration:

$$\frac{[S]}{[S]_0} = f_i = exp\left(-\frac{k_{cat}[E]_0\Delta t}{K_m}\right) \tag{S2.2}$$

Assume culture 'brightness' (growth) is inversely proportional to AMP concentration; *i.e.*, a more concentrated AMP will cause a darker zone of inhibition. As an AMP is degraded by a given protease, full-length AMP concentration will decrease, causing an increasing brightness of the observed zone of inhibition. The normalized inverse brightness values are then fit to Equation 2.1:

$$B = B_{min} + (B_{max} - B_{min})exp\left(-\frac{k_{cat}[E]_0\Delta t}{K_m}\right) \tag{2.1}$$

B is the normalized inverse brightness value, and B_{min} and B_{max} are the minimum and maximum fit values. Given the known $[E_0]$ and Δt, k_{cat}/K_m can be determined. Standard error of model fit values is also reported.

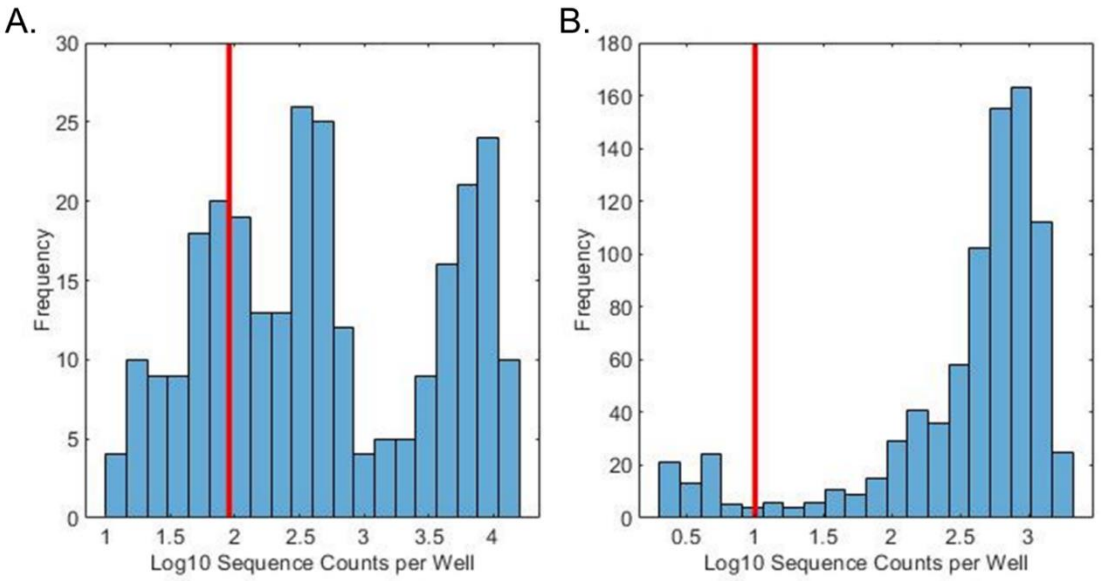


Figure S2.1 Distribution of AMP constructs per well.

Histograms showing frequency and distribution of AMP constructs (≥75 nucleotides) per well from iSeq sequencing runs 1 (A) and 2 (B). Red lines identify noise thresholds of 90 and 10 sequences per well for sequencing runs 1 and 2, respectively.

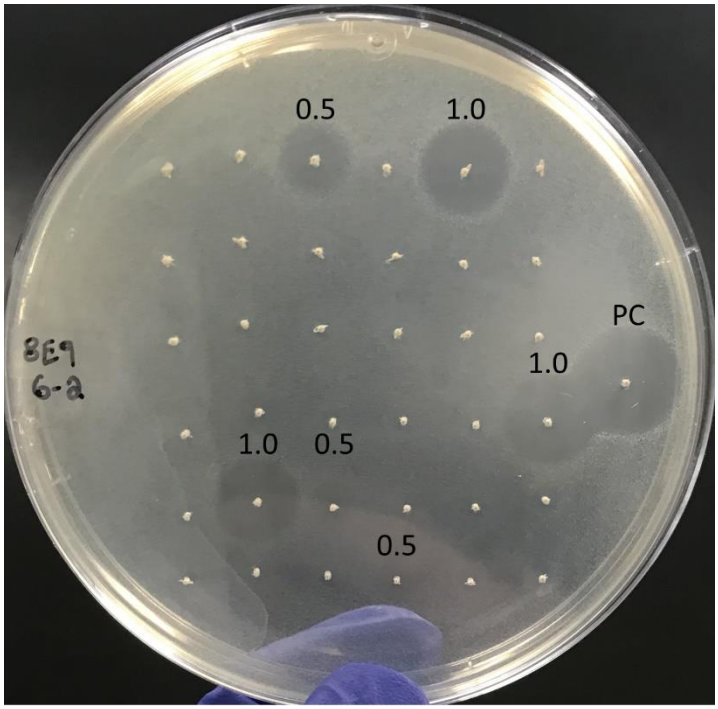


Figure S2.2 Activity scoring of zones of inhibition based on size relative to enterocin P positive control.

All identified zones of inhibition from the example plate have activity scores shown above the zone. PC is the enterocin P positive control. Scores of 1.0 are determined to be comparable or larger than the positive control. Scores of 0.5 are determined to be smaller than the positive control and include faint formation of zones of inhibition as well.

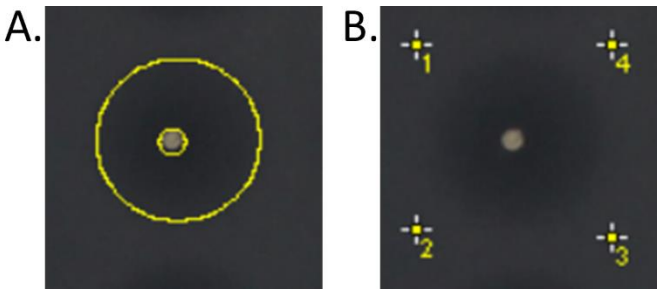
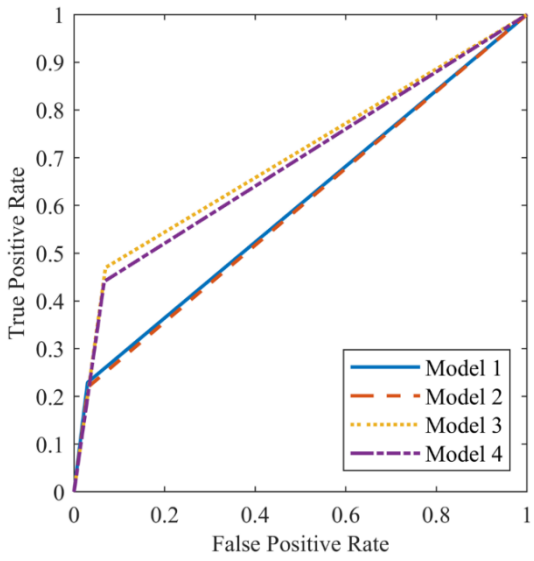


Figure S2.3 Location of image brightness measurements for protease assay AMP loss-of-activity quantification.

(A) Brightness of halos was measured, excluding the interior circle made by the pipette tip, with a circle of the same area on all halos on a given plate. (B) Brightness of halos was normalized to the brightness of four point-measurements at the corners of each halo.



	Model 1	Model 2	Model 3	Model 4
AUC	0.60±0.02	0.59±0.01	0.70±0.04	0.69±0.04
RMSE	0.38±0.01	0.350±0.008	0.35±0.02	0.35±0.02

Figure S2.4 Class IIa bacteriocin sequence-activity models show mild success. (A)Models 1 and 2 have limited success predicting activity towards at least one indicator strain or each indicator strain individually, with AUCs of 0.60±0.02 and 0.59±0.01, respectively. Models 3 (strain identity) and 4 (manPTS sequence), which both include strain-specific information in the model input, perform better than Models 1 and 2, with AUCs of 0.70±0.04 and 0.69±0.04.

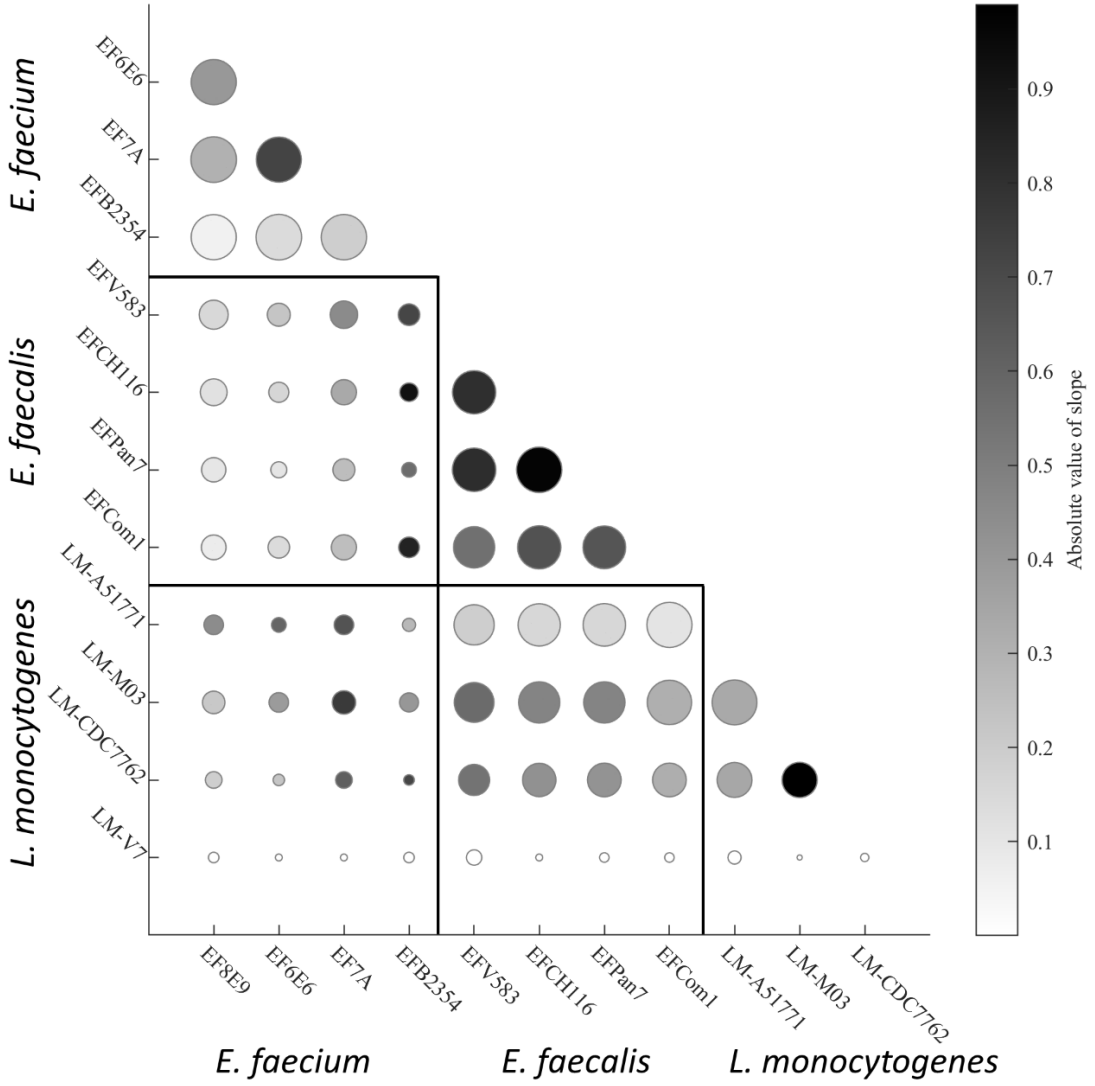


Figure S2.5 Correlation and slope of class IIa bacteriocin MIDs between species identify trends in susceptibility.

Size and color of the circle represents the magnitude of the correlation coefficient and the absolute value of the slope of a linear fit between the AMP MIDs towards strains *i* and *j*, respectively. The inverse of slopes greater than one were used for color plots. Larger correlations and slopes near 1 between strains of the same species can be identified near the diagonal. Large circles between *E. faecalis* and *L. monocytogenes* strains suggest additional correlation between their susceptibilities to class IIa bacteriocins. Nearly no correlation is identified between *L. monocytogenes* V7 and other strains due to its low susceptibility to bacteriocin IIa AMPs.

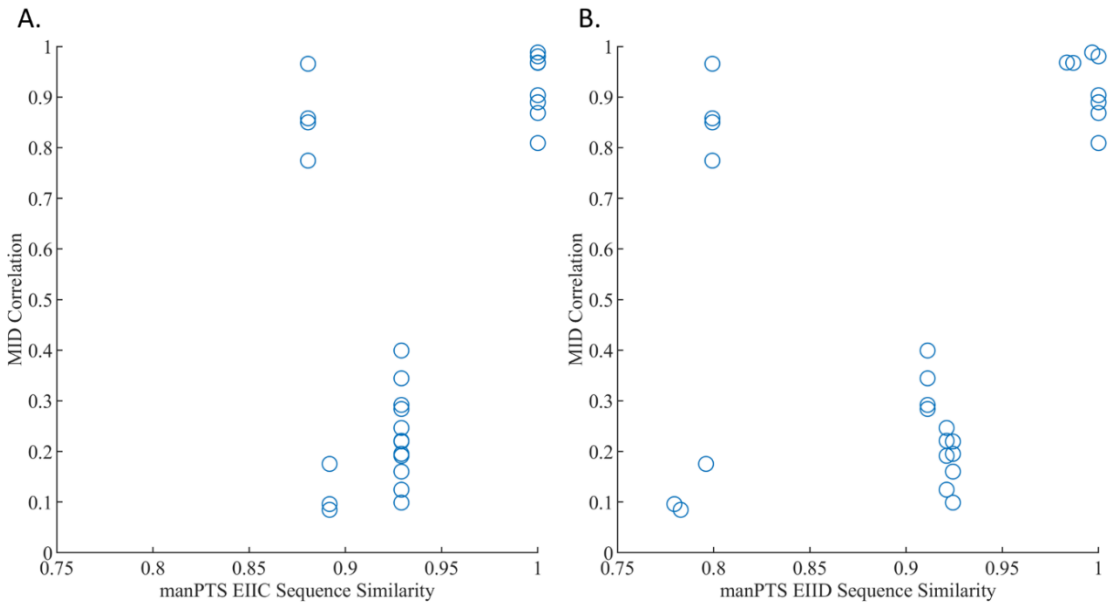


Figure S2.6 Correlation in susceptibility to class IIa bacteriocins vs manPTS sequence similarity.

Susceptibility is plotted against manPTS EIIC (A) and EIID (B) sequence similarity for *E. faecium* 8E9, *E. faecium* 6E6, *E. faecium* NRRL B2354, *E. faecalis* V583, *E. faecalis* CH116, *E. faecalis* Com1, *E. faecalis* Pan7, and *L. monocytogenes* ATCC 51775. The remaining 4 strains were not included because manPTS sequence information was unavailable.

Table S2.1 Strains and plasmids used in this study.

Bacterial Strain or Plasmid	**Description**	**Source**
Production strains		
Lactococcus lactis NZ9000	Plasmid-free strain, derivative of MG1363; pepN::nisRK; nonbacteriocin producer	MoBiTec
NEB 5-alpha *E. coli*	Chemically competent, high efficiency derivative of DH5α E. coli cells; T1 phage resistant and endA deficient	New England Biolabs
***Enterococcus faecium* strains**		
E. faecium 6E6	Ampicillin/vancomycin/linezolid resistant	Prof. Patricia Ferrieri, University of Minnesota
E. faecium 7A	Ampicillin/linezolid resistant	Prof. Patricia Ferrieri, University of Minnesota
E. faecium 8E9	Ampicillin/vancomycin/linezolid resistant	Prof. Patricia Ferrieri, University of Minnesota
E. faecium NRRL B2354	Non-pathogenic commensal strain	NRRL Agricultural Research Service Culture Collection
***Enterococcus faecalis* strains**		
E. faecalis Com1	Fecal sample from healthy volunteer	Prof. Gary Dunny, University of Minnesota
E. faecalis CH116	Gentamicin/kanamycin/streptomycin/tetracycline/erythromycin/penicillin-resistant, β-lactamase-producing isolate	Prof. Gary Dunny, University of Minnesota
E. faecalis Pan7	Panose 7; fecal sample from healthy volunteer	Prof. Gary Dunny, University of Minnesota
E. faecalis V583	ATCC 700802; vancomycin resistant	Prof. Gary Dunny, University of Minnesota
***Listeria monocytogenes* strains**		
L. monocytogenes ATCC 51775	--	Prof. Francisco Diez-Gonzalez
L. monocytogenes M-03-1213B-1	--	Prof. Francisco Diez-Gonzalez
L. monocytogenes CDC 7762	--	Prof. Francisco Diez-Gonzalez
L. monocytogenes V7	--	Prof. Francisco Diez-Gonzalez
Plasmid		
pNZC	Chloride-inducible L. lactis expression vector	University of Minnesota

Table S2.2 AMPs observed in library evaluation

UniProt ID[b]	AMP Identifier[c]	Sequence	Observed Activity Score[a]						Obs.
			EFs6E6	EFs8E9	EFmCh116	EFmV583	EFmCom1	EFmPan7	
SAKA_LACCU	Curvacin A	ARSYGNGVYCNNKKCWVNRGEATQSIIGGMISGWASGLAGM	0.25	0.75	0.25	0.25	0.5	0.5	2
I7H7S7_ENTFC_4M_39	Enterocin NKR-5-3C_Y3C, S12K, G20V, L35W	ATCYGNGLYCNKKKCWVEWVITGGCLAQYAIGGWWGGAVPGKC	0	0	0	0	0	0	1
I7H7S7_ENTFC_2M_18	Enterocin NKR-5-3C_Y3C, C15E	ATCYGNGLYCNSKKEWVEWGITGGCLAQYAIGGWLGGAVPGKC	0	0	0	0	0	0	1
UNID_AMP_P12_36	UNID_AMP_P12_36	ATCYGVGLYCNAKKCWVEWGITGGCIADVAIGGWLGGAVPDKC	0	0	0	0	0	0	1
UNID_AMP_P12_5	UNID_AMP_P12_5	ATCYGVGLYCNAKKCWVEWGITGGCIADVAIGGWLGGAVPG	0	0	0	0	0	0	2
I7H7S7_ENTFC_6M_36	Enterocin NKR-5-3C_Y3C, N6V, S12A, L26I, Q28D, Y29V	ATCYGVGLYCNAKKCWVEWGITGGCIADVAIGGWLGGAVPGKC	0	0	0	0	0	0	1
UNID_AMP_P6_75	UNID_AMP_P6_75	ATDYGNGLYCNSNKCWVVWGIFGGCLHQYLIGGWLGGAVPGKC	0	0	0	0	0	0	2
UNID_AMP_P8_69	UNID_AMP_P8_69	ATDYGNGLYCNSNKCWVVWGIFGGCLHQYRIGGWLGGAVPGKY	0	0	0	0	0	0	1
UNID_AMP_P8_12	UNID_AMP_P8_12	ATDYGNGLYCNSNKCWVVWGIFGGCLHQYRIGSWLGGAVPGKC	0	0	0	0	0	0	3
A0A1V8X506_ENTHR_6M_31	Enterocin P_S4C, S14N, V18R, A23F, S33V, A36D	ATRCYGNGVYCNNNKCWRNWGEFKENIAGIVIVGWDSGLAGMGH	0	0	0	0	0	0	7
A0A1V8X506_ENTHR_2M_28	Enterocin P_S4C, S33D	ATRCYGNGVYCNNSKCWVNWGEAKENIAGIVIDGWASGLAGMGH	0.5	0.5	0.125	0.375	0.125	0.375	4
A0A1V8X506_ENTHR_4M_30	Enterocin P_S4D, G8C, C11Y, G29D	ATRDYGNCVYYNNSKCWVNWGEAKENIADIVISGWASGLAGMGH	0	0	0	0	0	0	1
A0A1V8X506_ENTHR_6M_21	Enterocin P_S4D, N7V, S14T, N19D, E22T, K24A	ATRDYGVGVYCNNTKCWVDWGTAAENIAGIVISGWASGLAGMGH	0	0	0	0	0	0	3
UNID_AMP_P6_78	UNID_AMP_P6_78	ATRRHGNDIYCNNNKCWVNWGEAAEQIAGIVYSGWASGLAGMGH	0	0	0	0	0	0	3
A0A1V8X506_ENTHR_6M_6	Enterocin P_S4R, Y5H, N13C, V18H, I30N, G34Q	ATRRHGNGVYCNCSKCWHNWGEAKENIAGNVISQWASGLAGMGH	0	0	0	0	0	0	2
A0A1V8X506_ENTHR_4M_3	Enterocin P_S4R, Y5H, G8V, K24S	ATRRHGNVVYCNNSKCWVNWGEASENIAGIVISGWASGLAGMGH	0	0	0	0	0	0	1
A0A1V8X506_ENTHR_2M_44	Enterocin P_Y5C, K15H	ATRSCGNGVYCNNSHCWVNWGEAKENIAGIVISGWASGLAGMGH	0.5	0.5	0.429	0.5	0.071	0.429	7
A0A1V8X506_ENTHR_6M_12	Enterocin P_Y5E, G8C, C11Q, G21S, I27G, I30N	ATRSEGNCVYQNNSKCWVNWSEAKENGAGNVISGWASGLAGMGH	0	0	0	0	0	0	2
A0A1V8X506_ENTHR_4M_18	Enterocin P_Y5E, V9D, N12D, A28R	ATRSEGNGDYCDNSKCWVNWGEAKENIRGIVISGWASGLAGMGH	0	0	0	0	0	0	3
UNID_AMP_P10_76	UNID_AMP_P10_76	ATRSEWNGVYCNTSKCWVNWEEAKENIAGGVISGWASGLAGMGH	0	0	0	0	0	0	1
A0A1V8X506_ENTHR_4M_24	Enterocin P_Y5E, G6W, N13T, I30G	ATRSEWNGVYCNTSKCWVNWGEAKENIAGGVISGWASGLAGMGH	0	0	0	0	0	0	7
A0A1V8X506_ENTHR_2M_10	Enterocin P_G6D, K15G	ATRSYDNGVYCNNSGCWVNWGEAKENIAGIVISGWASGLAGMGH	1	1	1	1	0.5	1	1
A0A1V8X506_ENTHR_2M_39	Enterocin P_G8D, N13A	ATRSYGNDVYCNASKCWVNWGEAKENIAGIVISGWASGLAGMGH	0	0	0	0	0	0	4
A0A1V8X506_ENTHR_2M_30	Enterocin P_V9N, A28R	ATRSYGNGNYCNNSKCWVNWGEAKENIRGIVISGWASGLAGMGH	0	0	0	0	0	0	3
A0A1V8X506_ENTHR_4M_9	Enterocin P_N13C, K24S, G29N, V31S	ATRSYGNGVYCNCSKCWVNWGEASENIANISISGWASGLAGMGH	0	0	0	0	0	0	3
A0A1V8X506_ENTHR_4M_36	Enterocin P_V18T, N19C, N26S, W35V	ATRSYGNGVYCNNSKCWTCWGEAKESIAGIVISGVASGLAGMGH	0	0	0	0	0	0	3
A0A1V8X506_ENTHR_2M_50	Enterocin P_E22C, G29E	ATRSYGNGVYCNNSKCWVNWGCAKENIAEIVISGWASGLAGMGH	0	0	0	0	0	0	2
A0A1V8X506_ENTHR_2M_1	Enterocin P_G29E, A36W	ATRSYGNGVYCNNSKCWVNWGEAKENIAEIVISGWWSGLAGMGH	0	0	0	0	0	0	1
UNID_AMP_P8_66	UNID_AMP_P8_66	ATRSYGNGVYCNNSKCWVNWGEAKENIAEIVISGWWSGLAGMRH	0	0	0	0	0	0	1
A0A1V8X506_ENTHR	Enterocin P	ATRSYGNGVYCNNSKCWVNWGEAKENIAGIVISGWASGLAGMGH	1	1	1	1	1	1	13
A0A1V8X506_ENTHR_4M_19	Enterocin P_G29N, I30N, I32M, W35F	ATRSYGNGVYCNNSKCWVNWGEAKENIANNVMSGFASGLAGMGH	0	0	0	0	0	0	1
A0A1V8X506_ENTHR_4M_13	Enterocin P_G21P, I30V, W35F, A36N	ATRSYGNGVYCNNSKCWVNWPEAKENIAGVVISGFNSGLAGMGH	0	0	0	0	0	0	2
A0A1V8X506_ENTHR_2M_38	Enterocin P_K15R, V31S	ATRSYGNGVYCNNSRCWVNWGEAKENIAGISISGWASGLAGMGH	1	1	1	1	1	1	1
A0A1V8X506_ENTHR_4M_39	Enterocin P_V9W, C11Y, N13T, S14T	ATRSYGNGWYYNTTKCWVNWGEAKENIAGIVISGWASGLAGMGH	0	0	0	0	0	0	1
UNID_AMP_P12_17	UNID_AMP_P12_17	ATRSYGNRIYCNNSKCWVNHGEAATNIAGIVISGINSGLAGMGH	0	0	0	0	0	0	2
A0A1V8X506_ENTHR_6M_3	Enterocin P_G8R, K15H, W20P, K24S, I32Y, S33D	ATRSYGNRVYCNNSHCWVNPGEASENIAGIVYDGWASGLAGMGH	0	0	0	0	0	0	2
A0A1V8X506_ENTHR_4M_32	Enterocin P_N7R, G8R, N19C, I27G	ATRSYGRRVYCNNSKCWVCWGEAKENGAGIVISGWASGLAGMGH	0	0	0	0	0	0	1
A0A1V8X506_ENTHR_2M_21	Enterocin P_N7V, Y10E	ATRSYGVGVECNNSKCWVNWGEAKENIAGIVISGWASGLAGMGH	0	0.5	0	0	0	0	3
A0A1V8X506_ENTHR_2M_6	Enterocin P_G6N, N12K	ATRSYNNGVYCKNSKCWVNWGEAKENIAGIVISGWASGLAGMGH	0	0	0	0	0	0	1
A0A1V8X506_ENTHR_6M_33	Enterocin P_G6N, K15G, C16H, W17N, N26S, G29M	ATRSYNNGVYCNNSGHNVNWGEAKESIAMIVISGWASGLAGMGH	0	0	0	0	0	0	15
A0A1V8X506_ENTHR_2M_19	Enterocin P_G6S, V18S	ATRSYSNGVYCNNSKCWSNWGEAKENIAGIVISGWASGLAGMGH	0	0	0	0	0	0	1
A0A1V8X506_ENTHR_6M_34	Enterocin P_G6S, E22Q, K24S, V31S, S33T, W35V	ATRSYSNGVYCNNSKCWVNWGQASENIAGISITGVASGLAGMGH	0	0	0	0	0	0	2
I7H7S7_ENTFC_6M_18	Enterocin NKR-5-3C_Y3R, E18D, W19H, I21C, C25Q, G32S	ATRYGNGLYCNSKKCWVDHGCTGGQLAQYAISGWLGGAVPGKC	0	0	0	0	0	0	2
A0A1V8X506_ENTHR_6M_14	Enterocin P_S4Y, N19V, E22C, I27D, A28H, I30G	ATRYYGNGVYCNNSKCWVVWGCAKENDHGGVISGWASGLAGMGH	0	0	0	0	0	0	11
I7H7S7_ENTFC_6M_14	Enterocin NKR-5-3C_Y3V, G5W, K13Q, G20V, A30G, G32N	ATVYWNGLYCNSQKCWVEWVITGGCLAQYGINGWLGGAVPGKC	0	0	0	0	0	0	3
I7H7S7_ENTFC_6M_4	Enterocin NKR-5-3C_Y4E, N6V, S12T, W19S, Q28E, G33Q	ATYEGVGLYCNTKKCWVESGITGGCLAEYAIGQWLGGAVPGKC	0	0	0	0	0	0	7

I7H7S7_ENTFC_4M_16	Enterocin NKR-5-3C_Y4H, C15H, W19S, Q28D	ATYHGNGLYCNSKKHWVESGITGGCLADYAIGGW LGGAVPGKC	0	0	0	0	0	0	3
I7H7S7_ENTFC_6M_40	Enterocin NKR-5-3C_Y4H, K13V, V17H, T22L, G33I, W34I	ATYHGNGLYCNSVKCWHEWGILGGCLAQYAIGIIL GGAVPGKC	0	0	0	0	0	0	2
I7H7S7_ENTFC_4M_47	Enterocin NKR-5-3C_Y4Q, V17R, A27H, Q28N	ATYQGNGLYCNSKKCWREWGITGGCLHNYAIGGW LGGAVPGKC	0	0	0	0	0	0	3
UNID_AMP_P5_62	UNID_AMP_P5_62	ATYYDNGLYCGAKKCWVEWGITSGCLAEYAITQW LGGAVPGKC	0	0	0	0	0	0	3
I7H7S7_ENTFC_4M_15	Enterocin NKR-5-3C_N6C, G23W, I31A, G32T	ATYYGCGLYCNSKKCWVEWGITWGCLAQYAATG WLGGAVPGKC	0	0	0	0	0	0	4
I7H7S7_ENTFC_6M_26	Enterocin NKR-5-3C_N6C, C15S, I21T, C25E, A30G, G32N	ATYYGCGLYCNSKKSWVEWGTTGGELAQYGINGW LGGAVPGKC	0	0	0	0	0	0	1
I7H7S7_ENTFC_4M_48	Enterocin NKR-5-3C_N6C, C10Y, I21C, L26I	ATYYGCGLYYNSKKCWVEWGCTGGCIAQYAIGGW LGGAVPGKC	0	0	0	0	0	0	1
I7H7S7_ENTFC_4M_42	Enterocin NKR-5-3C_N6C, L8W, W19P, A27H	ATYYGCGWYCNSKKCWVEPGITGGCLHQYAIGGW LGGAVPGKC	0	0	0	0	0	0	5
I7H7S7_ENTFC_6M_2	Enterocin NKR-5-3C_N6D, W19S, I21C, T22F, G24N, A27G	ATYYGDGLYCNSKKCWVESGCFGNCLGQYAIGGW LGGAVPGKC	0	0	0	0	0	0	4
I7H7S7_ENTFC_6M_23	Enterocin NKR-5-3C_G7C, Y9S, S12T, K14C, G20V, G32N	ATYYGNCLSCNTKCCWVEWVITGGCLAQYAINGW LGGAVPGKC	0	0	0	0	0	0	1
I7H7S7_ENTFC_6M_12	Enterocin NKR-5-3C_G7C, E18C, W19S, C25Q, Q28M, G32S	ATYYGNCLYCNSKKCWVCSGITGGQLAMYAISGW LGGAVPGKC	0	0	0	0	0	0	4
I7H7S7_ENTFC_4M_38	Enterocin NKR-5-3C_G7C, L8V, Y9M, C10I	ATYYGNCVMINSKKCWVEWGITGGCLAQYAIGGW LGGAVPGKC	0	0	0	0	0	0	2
I7H7S7_ENTFC_6M_28	Enterocin NKR-5-3C_G7D, S12A, W16S, I21C, G24N, I31A	ATYYGNDLYCNAKKCSVEWGCTGNCLAQYAAGG WLGGAVPGKC	0	0	0	0	0	0	4
UNID_AMP_P5_18	UNID_AMP_P5_18	ATYYGNDLYCNAKKCSVEWGCTGNCLAQYAAGG WLGGSVPGKC	0	0	0	0	0	0	11
I7H7S7_ENTFC_4M_12	Enterocin NKR-5-3C_Y9M, N11I, W16G, G32N	ATYYGNGLMCISKKCGVEWGITGGCLAQYAINGW LGGAVPGKC	0	0	0	0	0	0	1
UNID_AMP_P6_56	UNID_AMP_P6_56	ATYYGNGLSCNSKKCWVDWGITGGCLAQYGINGW LGGAVPGKC	0	0	0	0	0	0	1
I7H7S7_ENTFC_4M_14	Enterocin NKR-5-3C_Y9S, E18D, A27H, G32S	ATYYGNGLSCNSKKCWVDWGITGGCLHQYAISGW LGGAVPGKC	0	0	0	0	0	0	12
I7H7S7_ENTFC_4M_11	Enterocin NKR-5-3C_N11I, S12A, G20H, G24N	ATYYGNGLYCIAKKCWVEWHITGNCLAQYAIGGW LGGAVPGKC	0	0	0	0	0	0	15
UNID_AMP_P9_84	UNID_AMP_P9_84	ATYYGNGLYCIAKKCWVEWHITGNCLAQYAIGSW LGGAVPGKC	0	0	0	0	0	0	1
I7H7S7_ENTFC_2M_7	Enterocin NKR-5-3C_N11I, A30M	ATYYGNGLYCISKKCWVEWGITGGCLAQYMIGGW LGGAVPGKC	0	0	0	0	0	0	1
I7H7S7_ENTFC_2M_27	Enterocin NKR-5-3C_S12A, E18D	ATYYGNGLYCNAKKCWVDWGITGGCLAQYAIGG WLGGAVPGKC	0	0.708	0.75	0.917	0.583	0.667	12
I7H7S7_ENTFC_4M_32	Enterocin NKR-5-3C_K14H, C15E, E18H, W19H	ATYYGNGLYCNSKHEWVHHGITGGCLAQYAIGGW LGGAVPGKC	0	0	0	0	0	0	4
I7H7S7_ENTFC_4M_3	Enterocin NKR-5-3C_W16H, W19S, T22C, W34V	ATYYGNGLYCNSKKCHVESGICGGCLAQYAIGGVL GGAVPGKC	0	0.036	0	0	0	0	14
I7H7S7_ENTFC_2M_5	Enterocin NKR-5-3C_E18C, G20H	ATYYGNGLYCNSKKCWVCWHITGGCLAQYAIGG WLGGAVPGKC	0	0	0	0	0	0	5
I7H7S7_ENTFC_2M_22	Enterocin NKR-5-3C_I21C, A27G	ATYYGNGLYCNSKKCWVEWGCTGGCLGQYAIGG WLGGAVPGKC	0	0.5	0.5	0.5	0	0.5	1
I7H7S7_ENTFC	Enterocin NKR-5-3C	ATYYGNGLYCNSKKCWVEWGITGGCLAQYAIGGW LGGAVPGKC	0.4	0.9	1	1	1	1	5
I7H7S7_ENTFC_2M_4	Enterocin NKR-5-3C_A30G, G33R	ATYYGNGLYCNSKKCWVEWGITGGCLAQYGIGRW LGGAVPGKC	0	0	0	0	0	0	7
I7H7S7_ENTFC_4M_33	Enterocin NKR-5-3C_G20S, A27G, G32S, G33R	ATYYGNGLYCNSKKCWVEWSITGGCLGQYAISRW LGGAVPGKC	0	0	0	0	0	0	1.000
UNID_AMP_P7_65	UNID_AMP_P7_65	ATYYGNGLYCNSKKCWVEWSITGGCLGQYAISRW LSGAVPGKC	0	0	0	0	0	0	1
I7H7S7_ENTFC_2M_25	Enterocin NKR-5-3C_G20V, I31M	ATYYGNGLYCNSKKCWVEWVITGGCLAQYAMGG WLGGAVPGKC	0	0.429	0	0	0	0	7
UNID_AMP_P9_88	UNID_AMP_P9_88	ATYYGNGLYCNSKKCWVEWVITGGCLAQYAMVG WLGGAVPGKC	0	0	0	0	0	0	2
I7H7S7_ENTFC_2M_2	Enterocin NKR-5-3C_E18V, Y29C	ATYYGNGLYCNSKKCWVVWGITGGCLAQCAIGG WLGGAVPGKC	0	0	0	0	0	0	2
I7H7S7_ENTFC_2M_17	Enterocin NKR-5-3C_K14R, A27R	ATYYGNGLYCNSKRCWVEWGITGGCLRQYAIGGW LGGAVPGKC	0	0	0	0	0	0	3
I7H7S7_ENTFC_6M_16	Enterocin NKR-5-3C_K13V, G20V, G23W, C25S, L26R, Y29C	ATYYGNGLYCNSVKCWVEWVITWGSRAQCAIGG WLGGAVPGKC	0	0	0	0	0	0	2
I7H7S7_ENTFC_2M_19	Enterocin NKR-5-3C_S12T, T22A	ATYYGNGLYCNTKKCWVEWGIAGGCLAQYAIGG WLGGAVPGKC	0	0.75	0.75	1	0.5	0.75	2
I7H7S7_ENTFC_2M_43	Enterocin NKR-5-3C_C10Q, G24T	ATYYGNGLYQNSKKCWVEWGITGTCLAQYAIGGW LGGAVPGKC	0	0	0	0	0	0	7
I7H7S7_ENTFC_2M_21	Enterocin NKR-5-3C_C10Y, K14H	ATYYGNGLYYNSKHCWVEWGITGGCLAQYAIGGW LGGAVPGKC	0	0	0	0	0	0	10
I7H7S7_ENTFC_6M_17	Enterocin NKR-5-3C_L8V, E18V, G20H, T22C, A30M, L35D	ATYYGNGVYCNSKKCWVVWHICGGCLAQYMIGG WDGGAVPGKC	0	0	0	0	0	0	1
I7H7S7_ENTFC_4M_26	Enterocin NKR-5-3C_L8V, K13V, A30R, W34F	ATYYGNGVYCNSVKCWVEWGITGGCLAQYRIGGF LGGAVPGKC	0	0	0	0	0	0	2
UNID_AMP_P10_35	UNID_AMP_P10_35	ATYYGNGWYCDSKKCWVEWGITGTCLAQNAIGGG WLGGAVPGKC	0	0	0	0	0	0	2
I7H7S7_ENTFC_4M_43	Enterocin NKR-5-3C_L8W, N11D, G24T, Y29N	ATYYGNGWYCDSKKCWVEWGITGTCLAQNAIGG WLGGAVPGKC	0	0	0	0	0	0	12
I7H7S7_ENTFC_6M_43	Enterocin NKR-5-3C_G7V, S12K, K14C, W19H, Q28E, G32T	ATYYGNVLYCNKKCCWVEHGITGGCLAEYAITGW LGGAVPGKC	0	0	0	0	0	0	1

UNID_AMP_P8_29	UNID_AMP_P8_29	ATYYGRGLNCNSKKCWVCSGILGGCLAQNAIGGL GGAVPGKC	0	0	0	0	0	0	1
I7H7S7_ENTFC_6M_6	Enterocin NKR-5-3C_N6R, Y9N, E18C, W19S, T22L, Y29N	ATYYGRGLNCNSKKCWVCSGILGGCLAQNAIGGW LGGAVPGKC	0	0	0	0	0	0	2
UNID_AMP_P8_89	UNID_AMP_P8_89	ATYYGVGLYCNSKKCWGRVGITGGCLAQYAIVGFL GGAVPGKC	0	0	0	0	0	0	1
I7H7S7_ENTFC_2M_29	Enterocin NKR-5-3C_N6V, W34F	ATYYGVGLYCNSKKCWVEWGITGGCLAQYAIGGF LGGAVPGKC	0	0	0	0	0	0	9
I7H7S7_ENTFC_4M_40	Enterocin NKR-5-3C_N6V, L8N, W34V, L35A	ATYYGVGNYCNSKKCWVEWGITGGCLAQYAIGGV AGGAVPGKC	0	0	0	0	0	0	6
UNID_AMP_P7_89	UNID_AMP_P7_89	ATYYGYGWYCNSKKCWVEPGITGGCLHQYAIGGW LGGAVPGKC	0	0	0	0	0	0	1
I7H7S7_ENTFC_6M_8	Enterocin NKR-5-3C_G5S, N6C, E18V, W19S, G24T, G32N	ATYYSCGLYCNSKKCWVVSGITGTCLAQYAINGWL GGAVPGKC	0	0	0	0	0	0	1
PPA1_PEDAC_6M_13	Pediocin PA-1_T8E, C9M, K11N, W18H, I26H, N28I	KYYGNGVEMGNHSCSVNHGKATTCIHNIGAMAWA TGGHQGNHKC	0	0	0	0	0	0	7
UNID_AMP_P12_7	UNID_AMP_P12_7	KCEGNGVTCGKHSCSVNWGKATACIINNGATAIWT GGHRGNHKC	0	0	0	0	0	0	5
UNID_AMP_P12_35	UNID_AMP_P12_35	KCYGNGVHCGEHSCHSDLGIAIGNIGGNAAANWAT GGNAGSNK	0	0	0	0	0	0	1
SAKP_LACSK_4M_48	Sakacin P_Y2C, H12K, V16R, A29G	KCYGNGVHCGEKSCTRDWGTAIGNIGNNGAANWA TGGNAGWNK	0	0	0	0	0	0	2
SAKP_LACSK_6M_7	Sakacin P_Y2D, G4N, T15G, V16S, G19V, G26A	KDYNNGVHCGEHSCGSDWVTAIGNIANNAAANWA TGGNAGWNK	0	0	0	0	0	0	2
SAKP_LACSK_6M_38	Sakacin P_Y2V, G4D, N5V, N24C, I25G, N27E	KVYDVGVHCGEHSCTVDWGTAIGCGGENAAANW ATGGNAGWNK	0	0.25	0	0	0	0	2
PPA1_PEDAC_2M_29	Pediocin PA-1_Y2V, C9Y	KVYGNGVTYGKHSCSVNWGKATTCIINNGAMAWA TGGHQGNHKC	0	0	0	0	0	0	1
SAKP_LACSK_6M_35	Sakacin P_Y3H, V7W, C9I, H12T, C14H, W18S	KYHGNGWHIGETSHTVDSGTAIGNIGNNAAANWA TGGNAGWNK	0	0	0	0	0	0	2
SAKP_LACSK_6M_2	Sakacin P_Y3Q, G4S, V7W, S13K, D17V, A31V	KYQSNGWHCGEHKCTVVWGTAIGNIGNNAAVNW ATGGNAGWNK	0	0	0	0	0	0	4
SAKP_LACSK_2M_4	Sakacin P_N5C, N28V	KYYGCGVHCGEHSCTVDWGTAIGNIGNVAAANWA TGGNAGWNK	0	0	0	0	0	0	3
SAKP_LACSK_6M_37	Sakacin P_N5C, G10I, D17N, G19H, N27D, A31V	KYYGCGVHCIEHSCTVNWHTAIGNIGDNAAVNWA TGGNAGWNK	0	0	0	0	0	0	1
PPA1_PEDAC_4M_30	Pediocin PA-1_G6C, G10I, M31N, W33R	KYYGNCVTCIKHSCSVNWGKATTCIINNGANARAT GGHQGNHKC	0	0	0	0	0	0	1
I7H7S7_ENTFC_LM4	I7H7S7_ENTFC_LM4	KYYGNGLYCNSKKCWVEWGITGGCLAQYAIGGWLGGA VPGKC	0.5	0.5	0.5	1	0.5	0.5	1
SAKP_LACSK_2M_10	Sakacin P_H8E, C14S	KYYGNGVECGEHSSTVDWGTAIGNIGNNAAANWATGG NAGWNK	0	0	0	0	0	0	2
SAKP_LACSK_6M_4	Sakacin P_H8E, G10K, T15W, G26L, N28C, A34W	KYYGNGVECKEHSCWVDWGTAIGNILNCAAANWWTGG NAGWNK	0	0	0	0	0	0	18
SAKP_LACSK_4M_31	Sakacin P_V16S, A31S, N32Q, A34W	KYYGNGVHCGEHSCTSDWGTAIGNIGNNAASQWWTGG NAGWNK	0	0	0	0	0	0	2
SAKP_LACSK_2M_25	Sakacin P_G23H, I25T	KYYGNGVHCGEHSCTVDWGTAIHNTGNNAAANWATGG NAGWNK	0	0	0	0	0	0	1
SAKP_LACSK_4M_26	Sakacin P_A21F, N28G, A29M, W33F	KYYGNGVHCGEHSCTVDWGTFIGNIGNGMAANFATGGN AGWNK	0	0	0	0	0	0	2
SAKP_LACSK_2M_37	Sakacin P_G19P, T20C	KYYGNGVHCGEHSCTVDWPCAIGNIGNNAAANWATGG NAGWNK	0	0	0	0	0	0	1
UNID_AMP_P8_28	UNID_AMP_P8_28	KYYGNGVHCGEHSCWVDCGTAIGNIHNNAAANWATGG NAGWNK	0	0	0	0	0	0	2
Q4U1B4_LACCU	Lactocin DT2	KYYGNGVHCGKYSCTVDWGTAIGNIGNNAAANWATGG NAGWNK	0	0.5	0.5	0	1	1	1
UNID_AMP_P3_7	UNID_AMP_P3_7	KYYGNGVHCTKSGCRVNWGEAFSAGVHRLANGGNGFW	0	0.5	0	0.5	0.5	1	1
UNID_AMP_P4_68	UNID_AMP_P4_68	KYYGNGVHCTKSGCSVNWGEAASAEIHRLANGGNGFW	0	0	0	0	0	0	1
MTCY_LEUME	Mesentericin Y105	KYYGNGVHCTKSGCSVNWGEAASAGIHRLANGGNGFW	0	0	0	0	0	0.167	3
LCCA_LEUGE	Leucocin A	KYYGNGVHCTKSGCSVNWGEAFSAGVHRLANGGNGFW	0.167	0.111	0.5	0.167	0.222	0.5	9
UNID_AMP_P2_39	UNID_AMP_P2_39	KYYGNGVHCTKSGCSVNWGEAKENIAGIVISGWASGLA GMGH	0	0	0	0	0	0	1
UNID_AMP_P13_3	UNID_AMP_P13_3	KYYGNGVHMREHSSTVDWGTAIGQTGGGAAANWATGG NAGWNK	0	0	0	0	0	0	1
UNID_AMP_P5_32	UNID_AMP_P5_32	KYYGNGVHYKEHSCTVDWGTTIGNIGMNAAVNWATGG NAGWNK	0	0	0	0	0	0	3
PPA1_PEDAC_4M_40	Pediocin PA-1_T8M, I25R, N28G, M31N	KYYGNGVMCGKHSCSVNWGKATTCRINGGANAWAT GGHQGNHKC	0	0	0	0	0	0	2
SAKP_LACSK_2M_36	Sakacin P_H8N, V16S	KYYGNGVNCGEHSCTSDWGTAIGNIGNNAAANWATGG NAGWNK	0	0	0	0	0	0	7
UNID_AMP_P3_74	UNID_AMP_P3_74	KYYGNGVSCTKKHGCKVNWGQDFTCSVNRFANFGHGN C	0	0	0	0	0	0	1
UNID_AMP_P11_33	UNID_AMP_P11_33	KYYGNGVTCDKHHSSVNWGKATTCIHNNWAIGRRNRG HQGNHKC	0	0	0	0	0	0	1
UNID_AMP_P9_77	UNID_AMP_P9_77	KYYGNGVTCDNHRCSGKLGQNNHPNYQQLGHGLGHW WTPGQP	0	0	0	0	0	0	1
UNID_AMP_P10_16	UNID_AMP_P10_16	KYYGNGVTCGKHSCSVDWGKATTYIINNGAMAWATGG HQGTHKC	0	0	0	0	0	0	1
PPA1_PEDAC	Pediocin PA-1	KYYGNGVTCGKHSCSVNWGKATTCIINNGAMAWATGG HQGNHKC	0	0	0	0	0	0	1
PPA1_PEDAC_2M_36	Pediocin PA-1_A21T, A30E	KYYGNGVTCGKHSCSVNWGKTTTCIINNGEMAWATGG HQGNHKC	0	0	0	0	0	0	1
Q93FV7_LACPN	Plantaracin 423	KYYGNGVTCGKHSCSVNWGQAFSCSVSHLANFGHGKC	0.25	0.5	0.625	0.5	0.5	0.625	4
PPA1_PEDAC_2M_19	Pediocin PA-1_K20Q, T22L	KYYGNGVTCGKHSCSVNWGQALTCIINNGAMAWATGG HQGNHKC	0	0	0	0	0	0	2
PPA1_PEDAC_6M_12	Pediocin PA-1_H12K, W18L, T23H, N27D, M31S, W33I	KYYGNGVTCGKKSCSVNLGKATHCIIDNGASAIATGGHQ GNHKC	0	0	0	0	0	0	4
PPA1_PEDAC_6M_22	Pediocin PA-1_G10N, V16S, W18P, T22W, G29R, M31N	KYYGNGVTCNKHSCSSNPGKAWTCIINNRANAWATGGH QGNHKC	0	0	0	0	0	0	1
PPA1_PEDAC_6M_21	Pediocin PA-1_C9Q, H12K, C14E, V16T, I26G, N28V	KYYGNGVTQGKKSESTNWGKATTCIGNVGAMAWATGG HQGNHKC	0	0	0	0	0	0	1
UNID_AMP_P12_61	UNID_AMP_P12_61	KYYGNGVTQGKKSESTNWGKATTCIGNVGAMGLGHW WTPGQP	0	0	0	0	0	0	4

CONS_E3_L5	Consensus_E3_L5	KYYGNGVYCNKKKCWVDWGQAWTCIGNNSANGWASG LAGMGH	0	0	0	0	0	0	2
A0A0M1XYG5_ENTFC	Enterocin DT6	KYYGNGVYCTKNKCTVDWAKATTCIAGMSIGGFLGGAI PG	0	0	0	0	0	0	3
A0A076H018_ENTFC	Enterocin DT5	KYYGNGVYCTKNKCTVDWAKATTCIAGMSIGGFLGGAI PGK	0	0	0	0	0	0	1
SAKP_LACSK_2M_29	Sakacin P_G6R, A29M	KYYGNRVHCGEHSCTVDWGTAIGNIGNNMAANWATGG NAGWNK	0	0	0	0	0	0	2
PPA1_PEDAC_6M_33	Pediocin PA-1_G6R, C9M, N28G, M31T, A32R, A34W	KYYGNRVTMGKHSCSVNWGKATTCIINGGATRWWTGG HQGNHKC	0	0	0	0	0	0	3
SAKP_LACSK_6M_49	Sakacin P_N5R, G10N, G19H, G23T, N27G, A30M	KYYGRGVHCNEHSCTVDWHTAITNIGGNAMANWATGG NAGWNK	0	0	0	0	0	0	3
SAKP_LACSK_4M_45	Sakacin P_G4N, H12K, W18S, N28V	KYYNNGVHCGEKSCTVDSGTAIGNIGNVAAANWATGGN AGWNK	0	0	0	0	0	0	2
SAKP_LACSK_6M_50	Sakacin P_G4N, G6V, C9I, V16R, D17V, I25R	KYYNNVVHIGEHSCTRVWGTAIGNRGNNAAANWATGG NAGWNK	0	0	0	0	0	0	1
A0A109DGD3_9LACO_3	Lactocin DT4	RYHYYGNGVSCNRYHCRVDWSRSWYCIVNRAGGAYA TGGQATIGNC	0	0	0	0	0	0	3
A0A0R2AIP5_9LACO	Lactocin DT3	SRYYGNGITCGKHKCTVNWGQAWTCGVNRLANFGHGN C	0.25	0.25	0.375	0.25	0.25	0.375	4
E7FPZ4_9LACO	Lactocin DT1	SRYYGNGVTCGKHKCTVNWGQAWTCGVNRLANFGHG NC	0.333	0.5	0.667	0.667	0.5	0.833	3
A0A1H0XID5_9LACT	Carnocin DT1	STYYGNGVSCTKKKCSVNWGQSWTEGVQRWGDHLFG	0.167	0.5	0	0	0.167	0.167	3
Q9Z4J1_CARDV_4M_22	Divercin V41_Y3C, G20S, C25S, W34V	TKCYGNGVYCNSKKCWVDWSQASGSIGQTVVGGVLGG AIPGKC	0	0	0	0	0	0	6
Q9Z4J1_CARDV_4M_38	Divercin V41_Y3D, N6C, G7V, W34V	TKDYGCVVYCNSKKCWVDWGQASGCIGQTVVGGVLGG AIPGKC	0	0	0	0	0	0	3
A0A1V8X506_ENTHR_LM2	Enterocin P_LM2	TKSYGNGVYCNNSKCWVNWGEAKENIAGIVISGWASGL AGMGH	0.9	0.8	0.9	0.6	0.6	0.8	5
Q9Z4J1_CARDV_6M_24	Divercin V41_Y4C, G5S, W19L, G20H, V31Y, L35D	TKYCSNGVYCNSKKCWVDLHQASGCIGQTVYGGWDGG AIPGKC	0	0	0	0	0	0	3
Q9Z4J1_CARDV_4M_27	Divercin V41_Y4E, K14G, V17R, G27A	TKYEGNGVYCNSKGCWRDWGQASGCIAQTVVGGWLGG AIPGKC	0	0	0	0	0	0	1
Q9Z4J1_CARDV_4M_25	Divercin V41_Y4E, W16G, V17S, W19L	TKYEGNGVYCNSKKCGSDLGQASGCIGQTVVGGWLGG AIPGKC	0	0	0	0	0	0	5
UNID_AMP_P7_34	UNID_AMP_P7_34	TKYHGNGVYCNSKKCWVDWGYAVGCIGQTVVGGILSG AIPGKC	0	0	0	0	0	0	1
Q9Z4J1_CARDV_4M_47	Divercin V41_Y4Q, S12A, W16H, C25M	TKYQGNGVYCNAKKCHVDWGQASGMIGQTVVGGWLG GAIPGKC	0	0	0	0	0	0	1
UNID_AMP_P10_37	UNID_AMP_P10_37	TKYYDNGVYCNSKKPWRCWGQAAGSIGQTVVGGILGG AIPGKC	0	0	0	0	0	0	2
Q9Z4J1_CARDV_6M_16	Divercin V41_G5D, C10M, D18N, V30R, V31Y, G32S	TKYYDNGVYMNSKKCWVNWGQASGCIGQTRYSGWLG GAIPGKC	0	0	0	0	0	0	2
Q9Z4J1_CARDV_2M_44	Divercin V41_N6C, W16G	TKYYGCGVYCNSKKCGVDWGQASGCIGQTVVGGWLGG AIPGKC	0	0	0	0	0	0	1
Q9Z4J1_CARDV_2M_21	Divercin V41_N6D, W34F	TKYYGDGVYCNSKKCWVDWGQASGCIGQTVVGGFLGG AIPGKC	0	0	0	0	0	0	1
Q9Z4J1_CARDV_4M_42	Divercin V41_N6D, C10Y, A22C, T29G	TKYYGDGVYYNSKKCWVDWGQCSGCIGQGVVGGWLG GAIPGKC	0	0	0	0	0	0	2
Q9Z4J1_CARDV_4M_18	Divercin V41_G7C, V17T, Q28D, G32T	TKYYGNCVYCNSKKCWTDWGQASGCIGDTVVTGWLGG AIPGKC	0	0	0	0	0	0	10
UNID_AMP_P11_65	UNID_AMP_P11_65	TKYYGNGIYYNSKKSWVCWGQASGQIGQTVVGGWLGG AIPGKC	0	0	0	0	0	0	1
I7H7S7_ENTFC_LM2	Enterocin NKR-5-3C_LM2	TKYYGNGLYCNSKKCWVEWGITGGCLAQYAIGGWLGG AVPGKC	0.313	0.5	0.5	0.813	0.563	0.563	8
Q9Z4J1_CARDV_6M_2	Divercin V41_Y9E, N11K, W16S, V17T, D18C, I26G	TKYYGNGVECKSKKCSTCWGQASGCGGQTVVGGWLGG AIPGKC	0	0	0	0	0	0	16
Q9Z4J1_CARDV_4M_15	Divercin V41_Y9N, C25Q, V30G, L35A	TKYYGNGVNCNSKKCWVDWGQASGQIGQTGVGGWAG GAIPGKC	0	0	0	0	0	0	1
Q9Z4J1_CARDV_2M_4	Divercin V41_Y9S, K13T	TKYYGNGVSCNSTKCWVDWGQASGCIGQTVVGGWLGG AIPGKC	0.1	0.6	0.1	0.1	0	0.1	5
Q9Z4J1_CARDV_6M_10	Divercin V41_N11G, S12C, Q21K, G27H, G32V, L35N	TKYYGNGVYCGCKKCWVDWGKASGCIHQTVVVGWNG GAIPGKC	0	0	0	0	0	0	1
CONS_E1_L4	Consensus_E1_L4	TKYYGNGVYCNKKKCWVDWGQAWTCIGNNSANGWAG GAIPGKC	0	0	0	0	0	0	1
UNID_AMP_P8_92	UNID_AMP_P8_92	TKYYGNGVYCNSKKCWVDWGQASGCIGQTVVGGWLG GAIPDKC	0	0	0	0	0	0	4
Q9Z4J1_CARDV	Divercin V41	TKYYGNGVYCNSKKCWVDWGQASGCIGQTVVGGWLG GAIPGKC	0.485	0.939	0.833	0.864	0.727	0.833	33
UNID_AMP_P4_95	UNID_AMP_P4_95	TKYYGNGVYCNSKKCWVDWGQASGCIGQTVVVGWLG GAIPGKC	0	0	0	0	0	0	1
Q9Z4J1_CARDV_2M_30	Divercin V41_I26R, G32S	TKYYGNGVYCNSKKCWVDWGQASGCRGQTVVSGWLG GAIPGKC	0	0	0	0	0	0	5
UNID_AMP_P1_39	UNID_AMP_P1_39	TKYYGNGVYCNSKKCWVDWGQAWTCIGNNSANGWAT GRHQGNHKC	0	0	0	0	0	0	5
Q9Z4J1_CARDV_2M_40	Divercin V41_A22C, I26T	TKYYGNGVYCNSKKCWVDWGQCSGCTGQTVVGGWLG GAIPGKC	0	0	0	0	0	0	6
Q9Z4J1_CARDV_2M_17	Divercin V41_G20H, L35W	TKYYGNGVYCNSKKCWVDWHQASGCIGQTVVGGWWG GAIPGKC	0	0	0	0	0	0	6
Q9Z4J1_CARDV_6M_6	Divercin V41_C15P, V17R, D18C, S23A, C25S, W34I	TKYYGNGVYCNSKKPWRCWGQAAGSIGQTVVGGILGG AIPGKC	0	0	0	0	0	0	4
Q9Z4J1_CARDV_4M_39	Divercin V41_S12T, V17R, Q21C, G27H	TKYYGNGVYCNTKKCWRDWGCASGCIHQTVVGGWLG GAIPGKC	0	0	0	0	0	0	4
Q9Z4J1_CARDV_6M_12	Divercin V41_C10I, S12T, W16H, A22L, I26T, Q28N	TKYYGNGVYINTKKCHVDWGQLSGCTGNTVVGGWLGG AIPGKC	0	0	0	0	0	0	2
Q9Z4J1_CARDV_4M_13	Divercin V41_C10M, Q28D, T29C, G33R	TKYYGNGVYMNSKKCWVDWGQASGCIGDCVVGRWLG GAIPGKC	0	0	0	0	0	0	6
Q9Z4J1_CARDV_4M_14	Divercin V41_C10Q, K13N, A22C, V30M	TKYYGNGVYQNSNKCWVDWGQCSGCIGQTMVGGWLG GAIPGKC	0	0	0	0	0	0	1
Q9Z4J1_CARDV_4M_4	Divercin V41_C10Y, C15S, D18C, C25Q	TKYYGNGVYYNSKKSWVCWGQASGQIGQTVVGGWLG GAIPGKC	0	0	0	0	0	0	4

Q9Z4J1_CARDV_4M_37	Divercin V41_N6R, I26T, G27R, V31E	TKYYGRGVYCNSKKCWVDWGQASGCTRQTVEGGWLG GAIPGKC	0	0	0	0	0	0	5
Q9Z4J1_CARDV_6M_50	Divercin V41_G5S, N6C, W16N, Q21Y, V31A, W34V	TKYYSCGVYCNSKKCNVDWGYASGCIGQTVAGGVLGG AIPGKC	0	0	0	0	0	0	7
Q9Z4J1_CARDV_4M_1	Divercin V41_G5S, W19P, T29C, W34R	TKYYSNGVYCNSKKCWVDPGQASGCIGQCVVGGRLGG AIPGKC	0	0	0	0	0	0	7
UNID_AMP_P6_28	UNID_AMP_P6_28	TKYYSNGVYCNSKKHWHDWGQASGCIGQTVVGGWLG GAIPGKC	0	0	0	0	0	0	1
Q9Z4J1_CARDV_6M_3	Divercin V41_G5S, G7V, V17H, C25E, G27A, V30M	TKYYSNVVYCNSKKCWHDWGQASGEIAQTMVGGWLG GAIPGKC	0	0	0	0	0	0	1
Q9Z4J1_CARDV_6M_32	Divercin V41_G5S, G7V, K13V, C15E, S23V, V31M	TKYYSNVVYCNSVKEWVDWGQAVGCIGQTVMGGWLG GAIPGKC	0	0	0	0	0	0	2
Q9Z4J1_CARDV_4M_17	Divercin V41_G5W, A22F, S23V, W34F	TKYYWNGVYCNSKKCWVDWGQFVGCIGQTVVGGFLGG AIPGKC	0	0	0	0	0	0	11
E2JE35_ENTFC_S8	Enterocin A_S8	TTHSGKSSGNGVSCTKNKCTVDWAKATTCIAGMSIGGFL GGAIPGKC	0	0	0	0	0	0	6
E2JE35_ENTFC_6M_48	Enterocin A_Y7V, N10D, C14I, K25T, T27S, M33G	TTHSGKVYGDGVYITKNKCTVDWATASTCIAGGSIGGFL GGAIPGKC	0	0	0	0	0	0	1
E2JE35_ENTFC_6M_43	Enterocin A_Y8E, V21T, A24P, I30R, I35A, G36N	TTHSGKYEGNGVYCTKNKCTTDWPKATTCRAGMSAN GFLGGAIPGKC	0	0	0	0	0	0	2
E2JE35_ENTFC_2M_24	Enterocin A_V12N, M33G	TTHSGKYYGNGNYCTKNKCTVDWAKATTCIAGGSIGGF LGGAIPGKC	0	0	0	0	0	0	3
E2JE35_ENTFC_4M_29	Enterocin A_Y13N, T15I, N17V, V21T	TTHSGKYYGNGVNCIKVKCTTDWAKATTCIAGMSIGGFL GGAIPGKC	0	0	0	0	0	0	2
E2JE35_ENTFC_4M_2	Enterocin A_Y13S, N17V, V21T, G37R	TTHSGKYYGNGVSCTKVKCTTDWAKATTCIAGMSIGRFL GGAIPGKC	0	0	0	0	0	0	6
E2JE35_ENTFC_4M_37	Enterocin A_T15K, A26F, T28A, A31G	TTHSGKYYGNGVYCKKNKCTVDWAKFTACIGGMSIGGF LGGAIPGKC	0	0	0	0	0	0	1
CONS_1	Consensus_1	TTHSGKYYGNGVYCNKKKCWVDWGQAWTCIGNNSAN GWAGGAIPGKC	0	0	0	0	0	0	4
E2JE35_ENTFC_4M_20	Enterocin A_D22H, A24G, I35M, F38V	TTHSGKYYGNGVYCTKNKCTVHWGKATTCIAGMSMG GVLGGAIPGKC	0	0	0	0	0	0	1
UNID_AMP_P5_75	UNID_AMP_P5_75	TTHSGKYYGNGWYCTKNKCTVDPAKFTTCIAGMSISGFL GGAIPDKC	0	0	0	0	0	0	3
E2JE35_ENTFC_4M_16	Enterocin A_N10R, Y13E, V21S, D22V	TTHSGKYYGRGVECTKNKCTSVWAKATTCIAGMSIGGFL GGAIPGKC	0	0	0	0	0	0	3
UNID_AMP_P11_10	UNID_AMP_P11_10	TTYYGVGLYCNSKKCWVEWGITGGCLAQYAIGGFLGGA VPGKC	0	0	0	0	0	0	1
R2SRR0_9ENTE	Enterocin DT3	TYYGNGVSCGKTTCTVDWAAAGTCIAQISIGGFLGGAIPG KC	0.667	0.917	0.5	0.167	0	0.333	6
D2DXK5_ENTAV	Avicin A	TYYGNGVSCNKKGCSVDWGKAISIIGNNSAANLATGGA AGWKS	0.5	0.5	0	0	0.5	0.5	1
A0A229NE30_ENTFC_2	Enterocin DT2	TYYGNGVYCNTQKCWVDWSRARSEIVDRGVKAYVNGF TKVLGGVGGR	0.5	0.5	0.5	0	0	0.5	1
UNID_AMP_P1_21	UNID_AMP_P1_21	TYYGNGVYCNTQKCWVDWSRARSETVDRCVKAYVNG FTKVLGGVGGR	0	0	0	0	0	0	1
S0RBF1_9ENTE_2	Enterocin DT1	TYYGNGVYCNTQKCWVDWSRARSETVDRGVKAYVNG FTKVLGGVGGR	0.429	0.643	0.286	0.214	0	0.143	7
UNID_AMP_P1_53	UNID_AMP_P1_53	TYYGNGVYCNTQKCWVWSRARSETVDRGVKAYVNGF TKVLGGVGGR	0	0	0	0	0	0	1
UNID_AMP_P1_7	UNID_AMP_P1_7	YDNGIYCNNSKCWVNWGEAIGIIGNNSAANLATGGAAG WKS	0	0	0	0	0	0	3
Q8L0Y0_ENTFC	Enterocin DT4	YDNGIYCNNSKCWVNWGEAKENIAGIVISGWASGLAGM GH	0.025	0.025	0.025	0.025	0	0.025	20

[a]Activity score is averaged over all observations; [b]Where relevant, UniProt ID of the sequence is included. For mutant sequences included in the oligopool, it is the seed sequence UniProt entry name followed by a library identifier. [c]AMP identifier is the identifier used commonly in previous literature to describe a particular AMP. Identifiers for random or rational variants of seed sequences include the relevant positional mutations or descriptions of mutations (LM# or L# for N-terminal chimeras, E# for C-terminal chimeras, S# for stability mutations). For pediocin PA-1_{D17N} and sakacin P_{K11E}, the mutants included in the seed sequence are eliminated from the name. In instances of a novel natural AMP, the identifier is modified to describe the species of origin (Lactocin for *Lactobacillus*, enterocin for *Enterococcus*, etc.) and an index (DT#). For mutant sequences not included in the oligopool, it is UNID_AMP followed by a plate location identifier; EFm: *E. faecium*; EFs: *E. faecalis*; n: number of observations.

Table S2.3 DNA constructs and primers used in this study

Primer	Sequence (5' → 3')
High-throughput sequencing primers	
Forward Set 1 - CGTGA[a, b]	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAG(N)$^{1-3}$CGTGAATAAGGAGGTATGCCATGG
Forward Set 2 - ACATC[a, b]	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAG(N)$^{1-3}$ACATCATAAGGAGGTATGCCATGG
Forward Set 3 - GCCTA[a, b]	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAG(N)$^{1-3}$GCCTAATAAGGAGGTATGCCATGG
Forward Set 4 - TGGTC[a, b]	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAG(N)$^{1-3}$TGGTCATAAGGAGGTATGCCATGG
Forward Set 5 - CACTG[a, b]	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAG(N)$^{1-3}$CACTGATAAGGAGGTATGCCATGG
Reverse Set 1 - ATTG[a, b]	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAG(N)$^{1-3}$ATTGTGAGCTCTCTAGAACTAGT
Reverse Set 2 - GATC[a, b]	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAG(N)$^{1-3}$GATCTGAGCTCTCTAGAACTAGT
Reverse Set 3 - TCAA[a, b]	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAG(N)$^{1-3}$TCAATGAGCTCTCTAGAACTAGT
Reverse Set 4 - CTGA[a, b]	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAG(N)$^{1-3}$CTGATGAGCTCTCTAGAACTAGT
Reverse Set 5 - AAGC[a, b]	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAG(N)$^{1-3}$AAGCTGAGCTCTCTAGAACTAGT
Ni5N501	AATGATACGGCGACCACCGAGATCTACACTAGATCGCTCGTCGGCAGCGTC
Ni5N502	AATGATACGGCGACCACCGAGATCTACACCTCTCTATTCGTCGGCAGCGTC
Ni5N503	AATGATACGGCGACCACCGAGATCTACACTATCCTCTTCGTCGGCAGCGTC
Ni5N504	AATGATACGGCGACCACCGAGATCTACACAGAGTAGATCGTCGGCAGCGTC
Ni5N505	AATGATACGGCGACCACCGAGATCTACACGTAAGGAGTCGTCGGCAGCGTC
Ni5N506	AATGATACGGCGACCACCGAGATCTACACACTGCATATCGTCGGCAGCGTC
Ni5N507	AATGATACGGCGACCACCGAGATCTACACAAGGAGTATCGTCGGCAGCGTC
Ni5N508	AATGATACGGCGACCACCGAGATCTACACCTAAGCCTTCGTCGGCAGCGTC
Ni7N701	CAAGCAGAAGACGGCATACGAGATTCGCCTTAGTCTCGTGGGCTCGG
Ni7N702	CAAGCAGAAGACGGCATACGAGATCTAGTACGGTCTCGTGGGCTCGG
Ni7N703	CAAGCAGAAGACGGCATACGAGATTTCTGCCTGTCTCGTGGGCTCGG
Ni7N704	CAAGCAGAAGACGGCATACGAGATGCTCAGGAGTCTCGTGGGCTCGG

Ni7N705	CAAGCAGAAGACGGCATACGAGATAGGAGTCCGTCTCGTGGGCTCGG
Ni7N706	CAAGCAGAAGACGGCATACGAGATCATGCCTAGTCTCGTGGGCTCGG
Ni7N707	CAAGCAGAAGACGGCATACGAGATGTAGAGAGGTCTCGTGGGCTCGG
Ni7N708	CAAGCAGAAGACGGCATACGAGATCCTCTCTGGTCTCGTGGGCTCGG
Ni7N709	CAAGCAGAAGACGGCATACGAGATAGCGTAGCGTCTCGTGGGCTCGG
Ni7N710	CAAGCAGAAGACGGCATACGAGATCAGCCTCGGTCTCGTGGGCTCGG
Ni7N711	CAAGCAGAAGACGGCATACGAGATTGCCTCTTGTCTCGTGGGCTCGG
Ni7N712	CAAGCAGAAGACGGCATACGAGATTCCTCTACGTCTCGTGGGCTCGG
***E. faecium* 8E9 manPTS sequencing primers[c]**	
manPTS EIIC forward primer	ATGTCTATTATTTCAATAATTTTAG
manPTS EIIC reverse primer	TAATAGTCATTCAAAATGTCGCCTA
manPTS EIID forward primer	ATGGCAGAAGAAAAAATCAAATTAT
manPTS EIID reverse primer	TTATAAAAGTCCGATGACGTGTCCA
pNZC-AMP construct sequencing primers	
pNZC-AMP forward primer	GGTGAAGATAGGTTGTCTGAAGC
pNZC-AMP reverse primer	GGCTATCAATCAAAGCAACACG
Library Construction primers	
pNZC-Usp45 forward primer	TATAAGGAGGTATGCCATGGATGAAAAAAAAGATTATCTCAGCTATTTTAATGTCTACAGTGATACTTTCTGCTGCAGCCCCGTT
AMP-pNZC reverse primer	TTGAGCTCTCTAGAACTAGTTTGAGCTCTCTAGAACTAGTTTA
Usp45-AMP ligation forward primer	AGTGATACTTTCTGCTGCAGCCCCGTTGTCAGGTGTTTACGCTN
Usp45-AMP ligation reverse primer	AGCGTAAACACCTGACAACGGGGCTGCAGCAGAAAGTATCACT
AMP-pNZC ligation forward primer	NTGATAAACTAGTTCTAGAGAGCTCAA
AMP-pNZC ligation reverse primer	TTGAGCTCTCTAGAACTAGTTTATCA
Library 1 amplification forward primer	GACAATGTCGTGGCGTATCCTCGGG
Library 1 amplification reverse primer	CGGAGTAAGATCCACTCCATCGCGC
Library 2 amplification forward primer	CCTTGACTCACGAGGTATCCCCTGG
Library 2 amplification reverse primer	GCGTCAGGTACTCGTGCCATCCTTC

[a]Each primer name includes the four or five nucleotides used as a plate identifier index; [b](N)1-3 identifies a group of three primers containing one, two, and three N nucleotides, respectively, encoding for any nucleotide; [c]these primers were designed from the *E. faecium* 6E6 manPTS sequences.

Chapter 3 - Computationally-aided discovery of LysEFm5 variants with improved catalytic activity and stability

Adapted from “Baryakova, T. H., Ritter, S. C., Tresnak, D. T. & Hackel, B. J. 2020. ‘Computationally Aided Discovery of LysEFm5 Variants with Improved Catalytic Activity and Stability.’ *Appl. Environ. Microbiol.* **86**, 1–21.”

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, S.C.R., D.T.T, and B.J.H. In particular, D.T.T. contributed to sequencing, production, purification, and testing of isolated variants and, in collaboration with other authors, manuscript preparation and revisions.

3.1 Abstract

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 1.8 $\pm$ 0.4 – fold improvement in specific activity and an 11.5 $\pm$ 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.

3.2 Statement of Importance

Broad-spectrum antibiotics can indiscriminately kill most bacteria, including commensal species that are a part of the normal human flora. This can potentially lead to the proliferation of drug-resistant bacteria upon elimination of competing species, and in unwanted autoimmune effects in patients. Bacteriophage-derived lysin proteins are an alternative to conventional antibiotics that have co-evolved alongside a specific bacterial host. Lysins are capable of targeting conserved substrates in the bacterial cell wall essential for its viability. To engineer these proteins to exhibit improved therapeutically-relevant properties, homology-guided statistical approaches can be used to identify compelling sites for mutation and

quantify the functional constraints acting on these sites to direct mutagenic library creation. The platform described herein couples this informed approach with a visual plate assay that can be used to simultaneously screen hundreds of mutants for catalytic activity, allowing for the streamlined identification of improved lysin variants.

3.3 Introduction

3.3.1 Antimicrobial lysin proteins

The misuse of antibiotics is a growing problem in the twenty-first century[118]. 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[65]. This reduction in diversity has been correlated with the onset of multiple health issues, including several inflammatory and autoimmune diseases[62]. 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[119]. 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[120–123]. Additionally, many lysins are specific, enabling them

to kill pathogens without exhibiting significant activity against commensal bacteria species[124,125]. 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[120,126]. 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[19]. 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[123]. *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[127]. The structure of each major type of catalytic domain is well-conserved between lysins derived from different phage species[128], as is generally observed for functional sites in enzymes[129]. 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[120]. 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[130,131].

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) [132]. *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[133–135].

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[132]. 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[136,137].

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[138]).

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 Fig. 3.1.

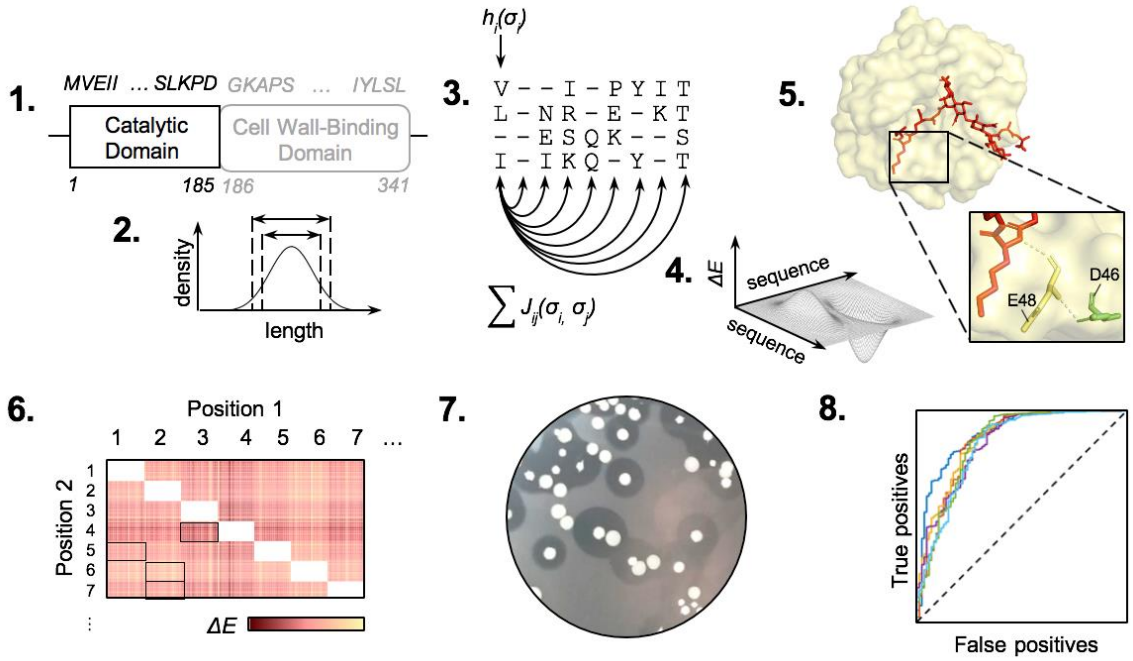


Figure 3.1 Research methodology.

(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.

3.3.2 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[128,139,140]. 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[141]. 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[141–145]. 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[142,146].

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[147]. Thus, the probability $P(\sigma)$ of observing any full-length amino acid sequence, σ, in the system can be computed.

$$P(\sigma) = \frac{e^{E(\sigma)}}{Z} \tag{3.1}$$

It is further assumed that the energy function $E(\sigma)$ in Eq. 3.1 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)[148]:

$$E(\sigma) = \sum_i h_i(\sigma_i) + \sum_i \sum_{j>i} J_{i,j}(\sigma_i, \sigma_j) \tag{3.2}$$

where $E(\sigma)$ is the statistical fitness, h_i 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 20^L 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[141]. 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[141,143,148,149].

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[142]. 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)[142]. 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.

3.4 Results

3.4.1 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[150], 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[151], 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 (Fig. 3.2A, B). Eleven structurally-analogous secondary residues (N32, S33, T34, A35, E38, T40, M45, N47, A74, I87, and V91) were selected as candidates for mutation (Fig. 3.2C). 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.

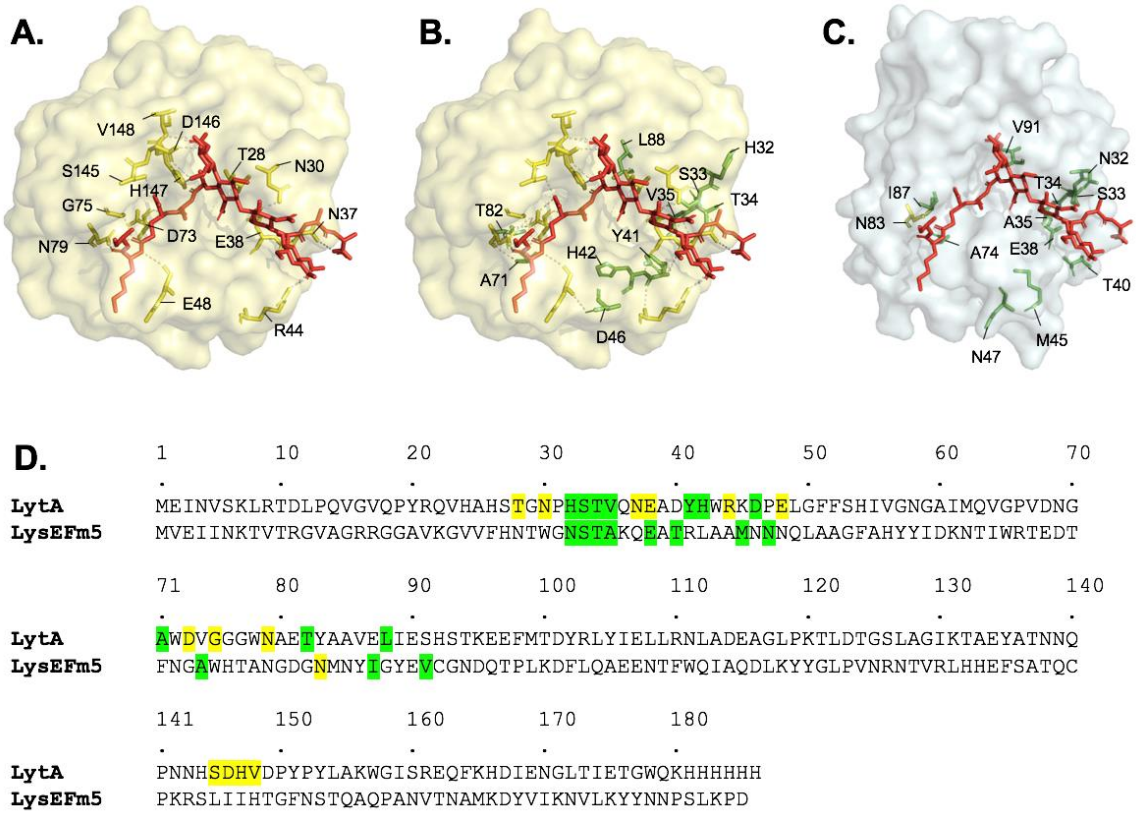


Figure 3.2 Primary and secondary amino acids in LytA and their putative structural analogs in LysEFm5.

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[152]. 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 3.1). Each set was independently input into PROMALS3D, an alignment tool that incorporates both sequence and structure information[153], to create an MSA.

Table 3.1. Predictive performance of the statistical fitness when different groups of homologous sequences are used in the starting MSA

Designation	Length[a]	No. of Seq-uences	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$_{lax\text{-}27k}$	137 – 200	26,950	5,565	0.868	0.888
Bacteria$_{stringent\text{-}23k}$	149 – 188	23,194	4,595	0.851	0.871
Bacteria$_{lax\text{-}3k}$		3,037	1,309 ± 39	0.839 ± 0.006	0.782 ± 0.013
Firmicutes$_{lax\text{-}3k}$	133 – 201	3,037	940	0.852	0.795
Firmicutes$_{stringent\text{-}2k}$	163 – 192	2,007	600	0.840	0.830
Firmicutes+All$_{lax\text{-}3k}$		3,037	1,344 ± 34	0.856 ± 0.005	0.818 ± 0.014
Firmicutes+Bacteria$_{lax\text{-}3k}$		3,037	1,317 ± 30	0.851 ± 0.004	0.814 ± 0.014
Firmicutes+Non-Bacteria$_{lax\text{-}2k}$		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*$_{stringent\text{-}2k}$) 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 (Fig. 3.3). 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 $\pm$ 2 observed for Library 1 to the lowest

value of -12 ± 3 observed for Library 8. (Fig. 3.3, 3.4). Library 8 contained the putative primary residue, N83.

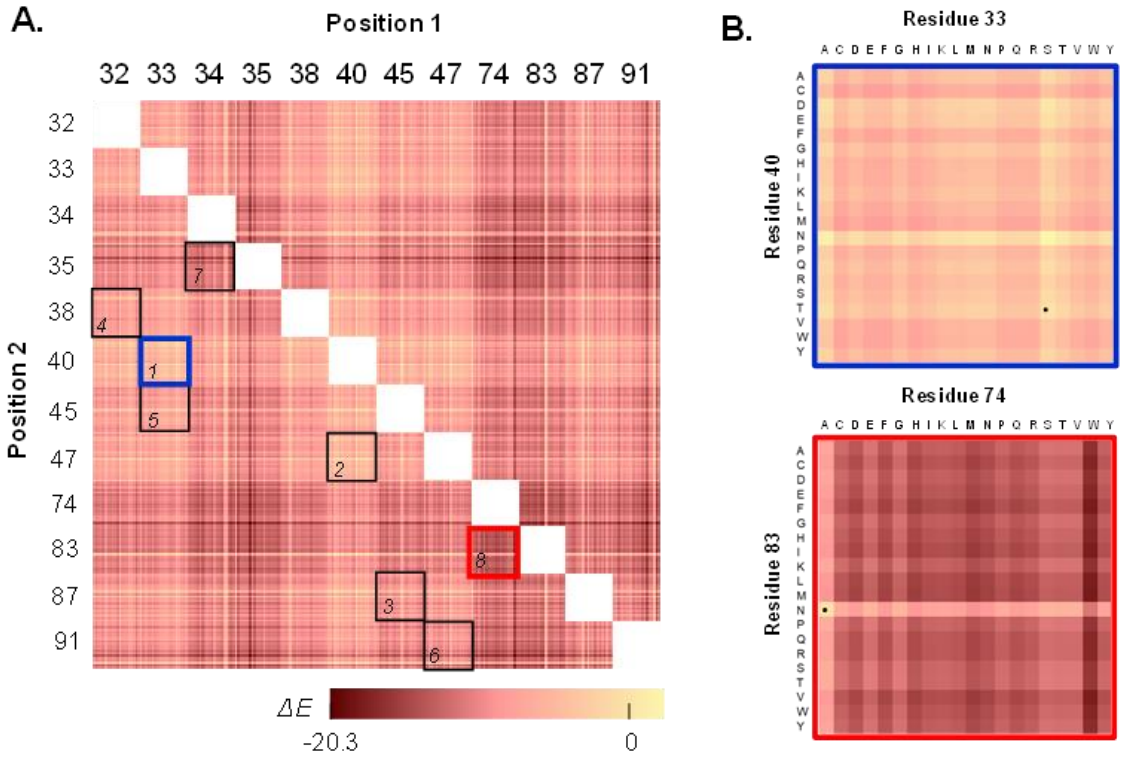


Figure 3.3 The predicted changes in statistical fitness for double mutants. The change in statistical fitness compared to the WT, *ΔE*, for all double mutants with diversity at positions 32, 33, 34, 35, 38, 40, 45, 47, 74, 83, 87, and/or 91 was computed using PLMC with inputs from MSA group *Firmicutes*$_{stringent\text{-}2k}$ (Table 3.1). (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.

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*)[154]. The resulting optimal library (Library 9) diversified the same two positions as Library 1 (33 and 40), 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).

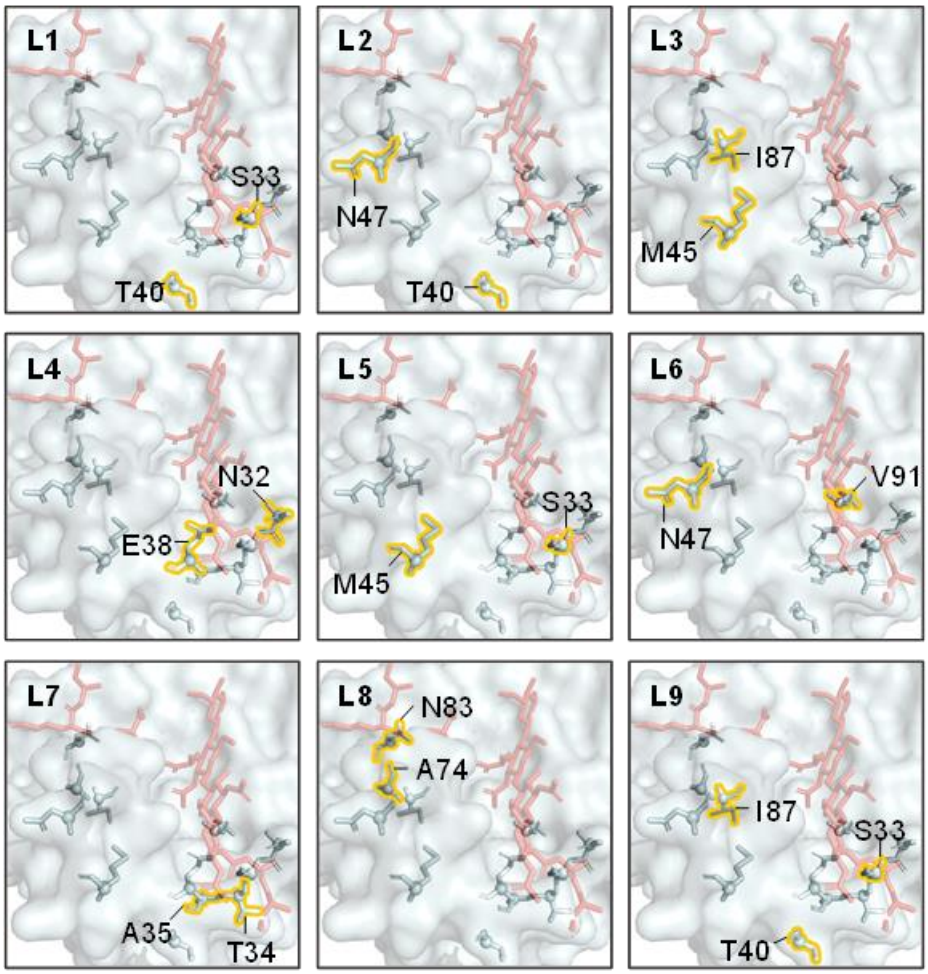


Figure 3.4 Location of residues in each library (L) relative to superimposed ligand in red

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[155]. 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.

3.4.2 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 (Fig. 3.5). Halo formation did not occur when *E. coli* transformed with a plasmid encoding a phage lysin with specific activity against *Clostridium perfringens*[50] were plated in an identical manner (Fig. S3.2). 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.[156]. 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[157], 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*[158]. 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*.

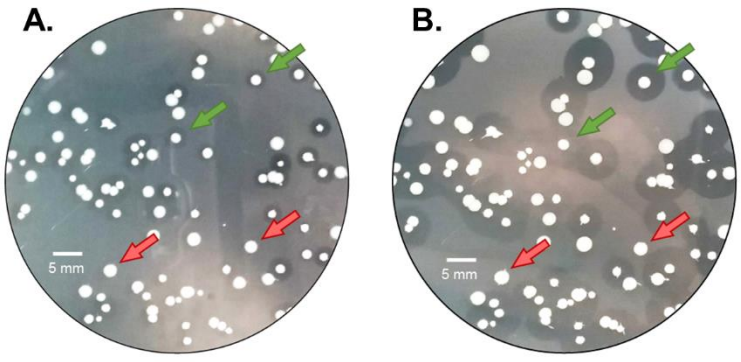


Figure 3.5 Halo formation over time.
This image is representative of the appearance of halo-forming and non-halo-forming variants in general. Libraries were plated on top of LB + kan/VRE/IPTG plates. (A) At approximately 16 – 18 hr following incubation at 37°C, discernable halos appeared around catalytically 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 (Fig. S3.5).

3.4.3 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 3.2). There is no general trend in the experimental retention of catalytic activity and the average statistical fitness of these libraries (Fig. 3.6). 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[151]. 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[137,159]. 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 $\pm$ 2 for both, compared to ΔE = -3 $\pm$ 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 3.2 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.

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.

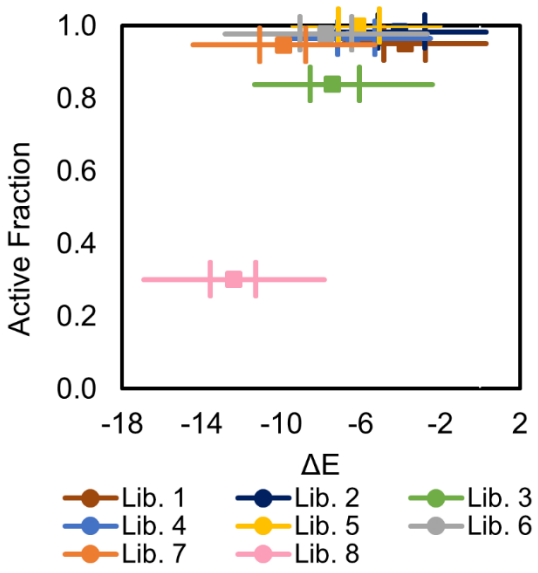


Figure 3.6 Active fraction of variants in library as a function of average ΔE. The difference between the statistical fitness of each variant (using MSA group *Firmicutes*$_{stringent\text{-}2k}$) and the WT, *ΔE*, is plotted against the active fraction of classified mutants in a library. The central 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.

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 3.3). 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 3.3 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.

Fig. 3.7 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).

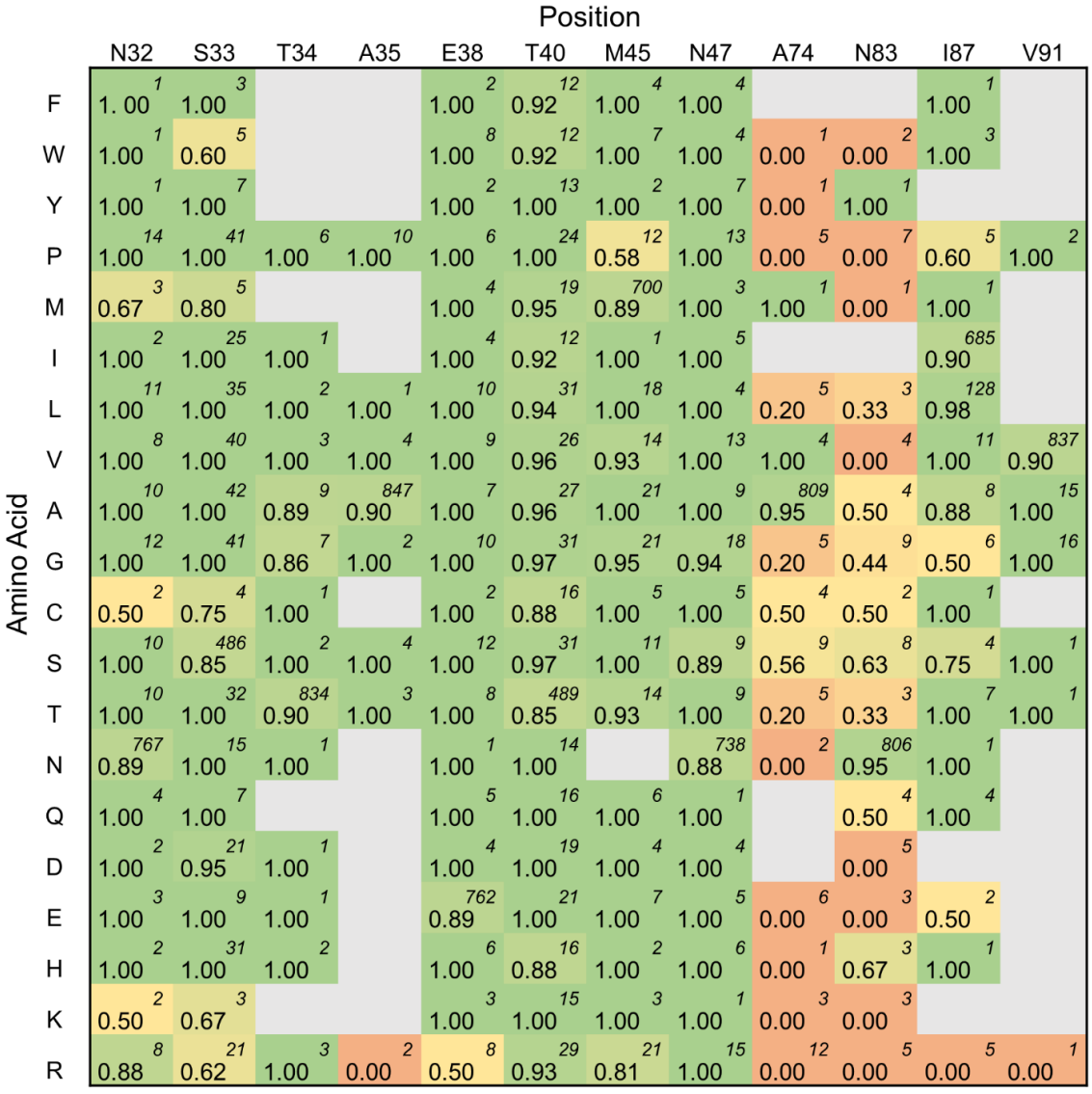


Figure 3.7 Average active fraction for a sequence based on the amino acid at the specified position.

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 1008 of independent mutation outcomes.

3.4.4 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.

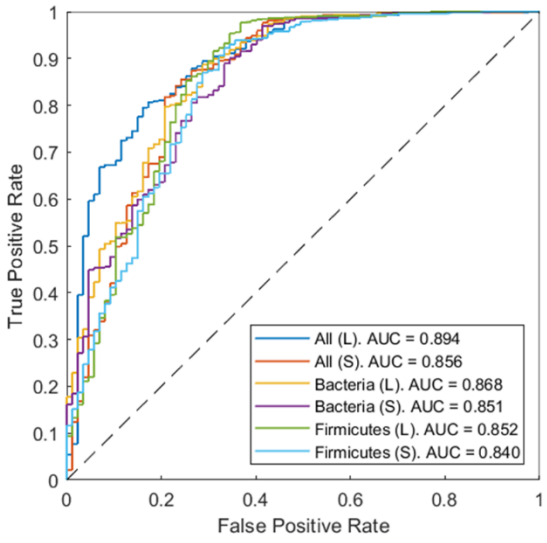


Figure 3.8 ROC curves demonstrating that the statistical fitness is indicative of variant activity.

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 3.1). 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.

All six ROC curves for the initial homology sequence sets yielded area under the ROC curve (AUC) values between 0.840 (*Firmicutes*$_{stringent\text{-}2k}$) and 0.894 (*All*$_{lax\text{-}29k}$), demonstrating the ability of the statistical fitness to discriminate between active and inactive variants (Fig. 3.8). 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 (Fig. 3.9A). 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[142].

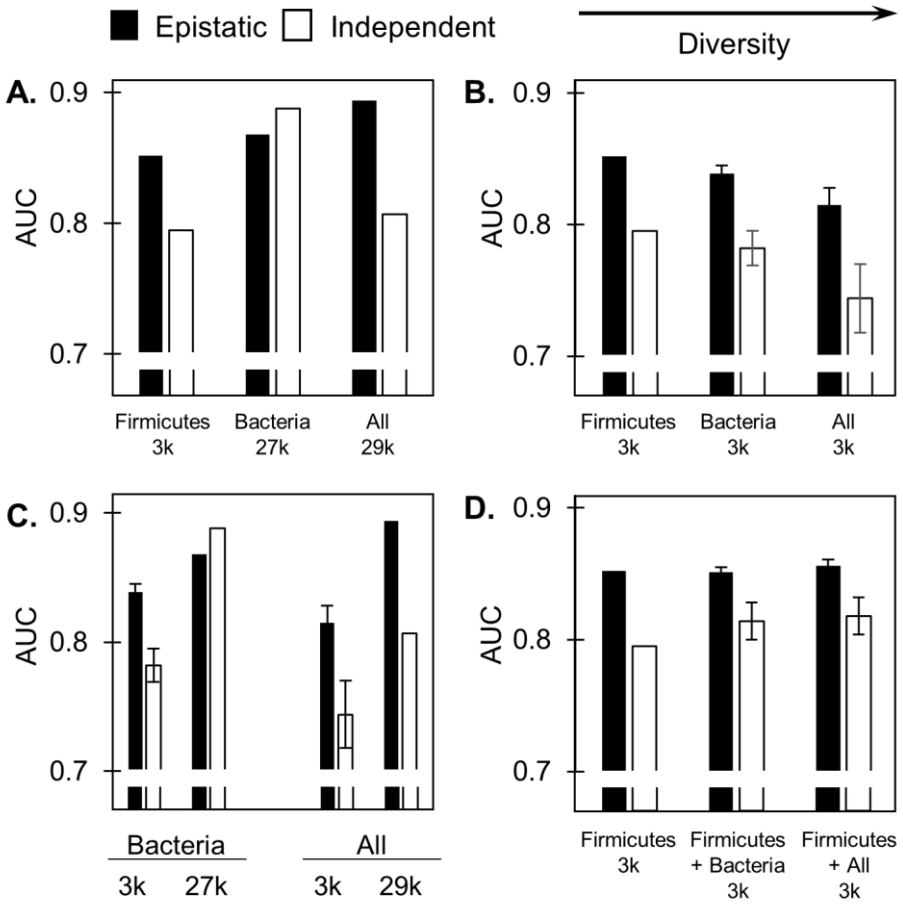


Figure 3.9 Effects of varying sequence diversity and depth on the predictive performance of the statistical fitness.

(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 1, 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*$_{lax\text{-}29k}$ and *Bacteria*$_{lax\text{-}27k}$ were randomly sampled twenty times each to create subgroups containing the same number of sequences as in *Firmicutes*$_{lax\text{-}3k}$ (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 3.1).

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 (Fig. 3.9B). Conversely, when the acceptable diversity was fixed, including more sequences in the starting MSA improved the predictive performance of both models (Fig. 3.9C).

Notably, 91% of the 2.9 x 10^4 sequences in the *All*$_{lax\text{-}29k}$ 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*$_{lax\text{-}27k}$) led to a +0.016 improvement in AUC. Supplementing further with only 2,548 non-bacterial sequences (to yield the group *All*$_{lax\text{-}29k}$) led to a further +0.026 improvement in AUC. The same was not true in the independent model: supplementing the group *Firmicutes*$_{lax\text{-}3k}$ to yield *Bacteria*$_{lax\text{-}27k}$ improved the AUC from 0.795 to 0.888

(+0.093), but further supplementing the group *Bacteria*$_{lax\text{-}27k}$ to yield *All*$_{lax\text{-}29k}$ 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*$_{lax\text{-}3k}$ 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*$_{lax\text{-}27k}$, 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.

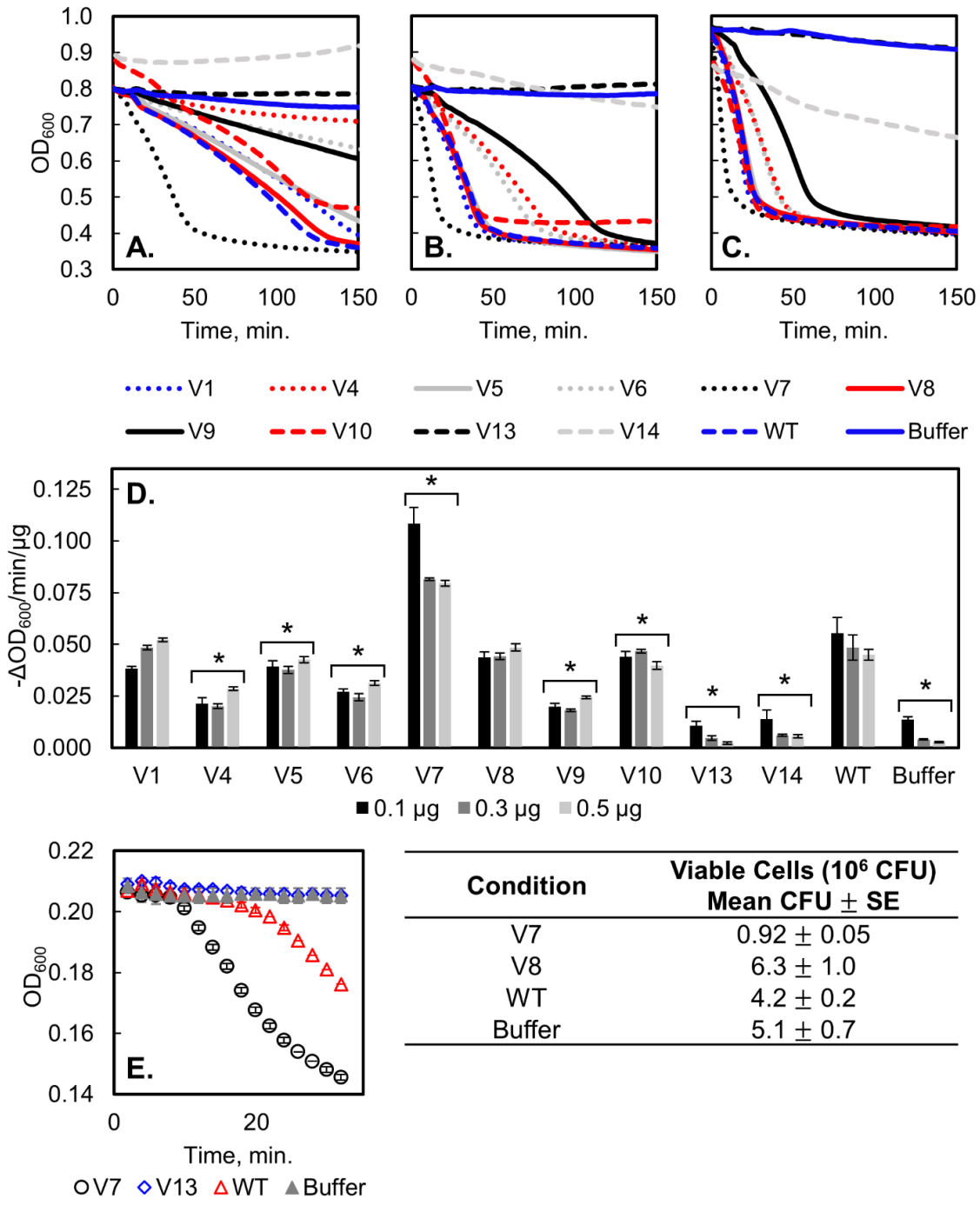
Thus, although using a small set of sequences that were phylogenetically similar to the WT led to a superior predictive performance as compared to using a small set of diverse sequences (AUC *Fimicutes*$_{lax\text{-}3k}$ > AUC *Bacteria*$_{lax\text{-}3k}$ > AUC *All*$_{lax\text{-}3k}$), once the MSA was seeded with a collection of close homologs, including more diverse sequences further improved the predictive performance of the epistatic model (AUC *All*$_{lax\text{-}29k}$ > AUC *Bacteria*$_{lax\text{-}27k}$). 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*$_{lax\text{-}29k}$ (to yield the subgroup *Firmicutes+All*$_{lax\text{-}3k}$) or *Bacteria*$_{lax\text{-}27k}$ (to yield *Firmicutes+Bacteria*$_{lax\text{-}3k}$), once again resulting in a total of 3,037

sequences per subgroup. The performance of both subgroups was compared to *Firmicutes*$_{lax\text{-}3k}$. 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 (Fig. 3.9D).

3.4.5 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 halo-forming variants from libraries 1 – 9 were randomly selected during the halo plate assay to more sensitively quantify their specific activity. 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%) (Fig S3.4). Variant 3 (M45E/I87D, Library 3) was unable to be produced in sufficient quantities and excluded from further testing. Activity was assayed using the turbidity reduction method. 0.1 µg, 0.3 µg, or 0.5 µg of each variant or the WT were co-incubated with crude VRE cell wall material. The OD_{600} of each sample was monitored over the course of four hours (Fig. 3.10A – C). The largest slope over a 10-minute timeframe was calculated for each replicate as a proxy for specific catalytic activity (Fig. 3.10D). Across all starting conditions, the WT exhibited a normalized change in OD_{600} of 0.048 $\pm$ 0.007 -ΔOD_{600}/min/µg (n = 32). Five variants (4, 5, 6, 9, and 10) exhibited activities that were moderately diminished as compared to the WT. Two variants (1 and 8)

exhibited activity that was statistically indistinguishable from the WT. Variant 7 exhibited a markedly improved activity of 0.09 $\pm$ 0.01 -ΔOD_{600}/min/μg ($p < 0.001$). To further validate the halo assay, we also tested eight random non-halo-forming variants (Table 3.4). Six of the eight yielded negligible recombinant production (four of which were revealed via sequencing to encode for undigested pET-24 vector), which validates their inability to generate halos. Variants 13 and 14, which were non-halo-forming, had activities of 0.005 $\pm$ 0.004 and 0.008 $\pm$ 0.004 -ΔOD_{600}/min/μg, respectively, which were statistically indistinguishable from the performance of the buffer negative control, 0.004 $\pm$ 0.004 -ΔOD_{600}/min/μg ($n = 32$), further supporting the validity of the halo assay. See Table 4 for a summary of variant mutations, libraries of origin, and -ΔOD_{600}/min/μg values.



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

Figure 3.10 LysEFm5 variant and WT activity against VRE.
(A) Turbidity reduction assay results where 0.1 µg (*n* = 4 for variants; *n* = 8 for the WT), (B) 0.3 µg (*n* = 4 for variants; *n* = 8 for the WT) or (C) 0.5 µg (*n* = 8 for variants; *n* = 16 for the WT) of each lysin was combined with VRE cell wall fragments in 200

μL of PBS (total). Additionally, a buffer negative control was included where no lysin was added (n = 32). OD_{600} was monitored over time, with data collected every two minutes. Error bars are not shown for visual clarity; a measure of the uncertainty between replicates is given in Fig. 10D as the standard deviations in the slope of the linear regions. (D) Quantified activity for variants, WT, and the buffer negative control. The maximum change in OD_{600} over a 10-minute period was calculated from the turbidity assay results for each replicate. *p < 0.004 (compared to the WT) for a two-tailed, two-sample heteroscedastic Student's t-test with a Bonferroni correction applied (η = 12). The activity of variants 13 and 14 were additionally statistically insignificant from the buffer negative control (p = 0.54 and p = 0.009, respectively). (E) Bacteriolytic activity of variants 7 and 13, 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.

Additionally, variant 7 (most active), variant 13 (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 OD_{600} of cultured VRE cells was found to agree well with the results of the turbidity reduction assay (Fig. 3.10E).

Table 3.4 Information for purified variants and the WT

Variant	Halo-forming?	Library	Mutations	-ΔOD$_{600}$/min/μg
WT	Y	N/A	N/A	0.048 ± 0.007
1	Y	2	T40P, N47V	0.048 ± 0.006
2	Y	N/A	R17L, T40L	N/A
3	Y	3	M45E, I87D	N/A
4	Y	4	N32G, E38T	0.025 ± 0.004
5	Y	5	S33P, M45W	0.041 ± 0.003
6	Y	6	N47A, V91P	0.028 ± 0.003
7	Y	7	T34S, A35V	0.087 ± 0.013
8	Y	1	T40A	0.046 ± 0.003
9	Y	1	S33G, T40S	0.022 ± 0.003
10	Y	9	S33G,T40E,I87L	0.043 ± 0.003

-	N	N/A	uncut vector	N/A
11	N	3	M45P,I87G	N/A
-	N	N/A	uncut vector	N/A
-	N	N/A	uncut vector	N/A
12	N	7	A35G,T34T	N/A
-	N	N/A	uncut	N/A
13	N	8	A74R, N83M	0.005 $\pm$ 0.004
14	N	8	A74T, N83Y	0.008 $\pm$ 0.004

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 – 9, 13, and the WT was assessed by Sypro Orange thermal denaturation assay (Fig. 3.11). Variant 1 was unable to be produced in sufficient quantities to use in this assay and was not tested. The melting temperature, T_m, of the WT was 43.4 ± 0.5 °C. Variant 4 did not exhibit a signal consistent with unfolding, perhaps resulting from a low T_m or increased disorder in the molecule. Variants 8 and 9 exhibited T_m that were statistically indistinguishable from the WT. Variants 5, 6, 7, and 13 exhibited improved T_m at or above 46.5 $\pm$ 0.4 °C, with variant 7 exhibiting the highest value, 54.9 $\pm$ 0.6 °C, an 11.5 $\pm$ 0.8 °C improvement over the WT. This variant comprises chemically homologous mutations – T34S and A35V – at adjacent sites (Fig. 4). The retention of amino acid characteristics at these two sites may be key to the observed improvements in both activity and stability.

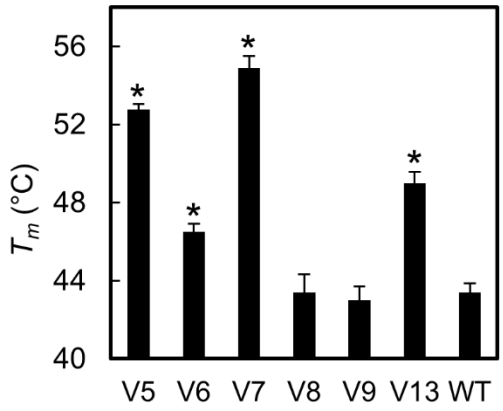


Figure 3.11 Lysin thermal stability.
The midpoint of thermal denaturation was measured for lysin variants 5 – 9, 13 and the WT by Sypro Orange assay. *p < 0.001 for a two-tailed, two-sample heteroscedastic Student's t-test.

3.5 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*[160]. 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[161]. Improvements in one or both characteristics can contribute to heightened bioavailability in an infected host, which can increase treatment efficacy. Of the eight randomly-selected, halo-forming lysin variants that were assayed for catalytic activity, three exhibited activity that was indistinguishable from the WT (0.048 ± 0.007 -ΔOD_{600}/min/µg) or improved. Six of the seven variants assayed for thermal stability exhibited a T_m 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 -ΔOD_{600}/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 tested 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 (ΔE = -12 ± 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[50]. 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[142,162,163]. 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[164].

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.

3.6 Materials and Methods

3.6.1 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.

3.6.2 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) (Fig. 3.2D) 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 1.0×10^{-5}.

A two-component Gaussian mixture model was constructed to describe the distribution of sequence lengths in the *Bacteria* and *All* groupings (Fig. S3.1). 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 3.1.

PROMALS3D[153] was used to generate an MSA for each of the resulting sets of protein sequences. The PLMC algorithm[142] was run using the recommended regularization parameters for the single-site and pairwise coupling constraints, without the inclusion of gap states. The strength of l_2-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.

3.6.3 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[151]. 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 *Firmicutes*$_{stringent\text{-}2k}$) 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 (Fig. 3.3). This heat map was used to select eight sets of two discrete sites for NNK library creation (Table 3.2), sampling a range of average ΔE values.

Library 9 was designed using the SwifLib tool[154]. The matrix of predicted change in statistical fitness was discretized into an integer matrix using the following criteria:

$-10 < \Delta E < -8,\ f(\Delta E) = -5$ $\quad -3 < \Delta E < 0\ f(\Delta E) = 1$ $\quad \Delta E > 1,\ f(\Delta E) = 10$

$-8 < \Delta E < -6,\ f(\Delta E) = -2$ $\quad \Delta E == 0,\ f(\Delta E) = 2$

$-6 < \Delta E < -3,\ f(\Delta E) = 0$ $\quad 0 < \Delta E < 1,\ f(\Delta E) = 5$

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.

3.6.4 Plasmid creation

A gBlock gene fragment (Integrated DNA Technologies) encoding the entire 341 AA LysEFm5 gene[132] was amplified via Q5 PCR (New England Biolabs, NEB), visualized on an agarose gel, and purified. The product was Gibson assembled[155] (NEBuilder® HiFi DNA Assembly, NEB) into a pET-24 vector modified to include a C-terminal 6xHIS tag[165] 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.

3.6.5 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 Tables S3.1 and S3.2. 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.

3.6.6 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 (Fig. S3.3). 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*[50] was plated alongside a positive control. At no time did halos form on the plate containing the *C. perfringens*-specific lysin (Fig. S3.2).

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.

3.6.7 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/I^q* 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 (7.90×10^8 colony forming units/OD_{600}/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.

3.6.8 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 S3.3). 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 NanoDrop™ 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.

3.6.9 Production of variants

During the agar plate assay, eight halo-forming variants were successfully isolated from libraries 1 (x2), 2, 3, 4, 5, 6, 7, and 9 (out of ten plucked), and two non-halo-forming variants were successfully isolated from library 8 (x2) (out of eight plucked). These 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 OD_{600} 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 Na_2HPO_4, 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 $MgSO_4$ 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.

3.6.10 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.

3.6.11 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.

3.6.12 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 OD_{600} 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 OD_{600} of 1. Each sample was tested with replication of 4 – 8 with randomized well positions. Measurements of the Abs_{600} were taken every 2 minutes for multiple hours.

3.6.13 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 (OD_{600} ~ 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 OD_{600} 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.

3.7 Acknowledgements

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

3.8 Supplement

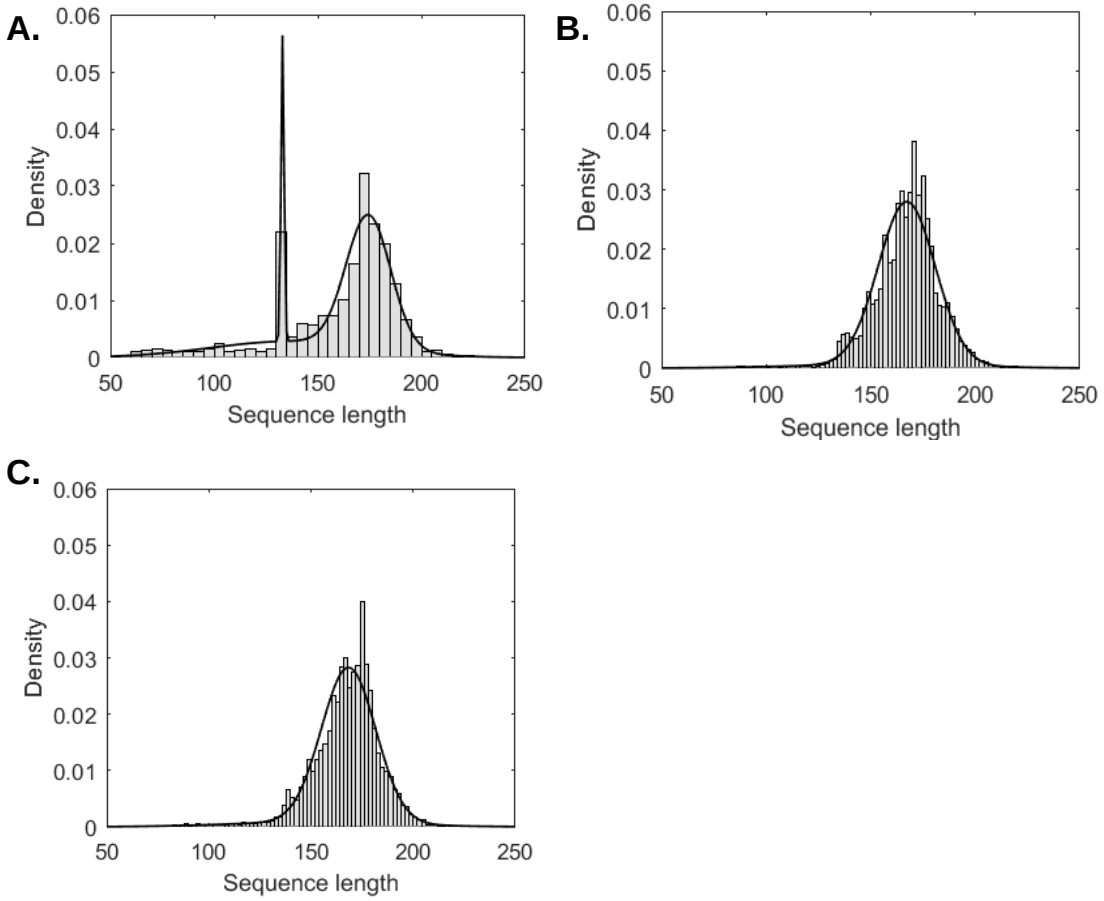


Figure S3.1 Two- or three-component Gaussian mixture models (GMMs) fitted to the distribution of sequence lengths resulting from a jackhmmer search of the UniProtKB database for homologs to the wild-type amidase domain of LysEFm5. (A) The taxonomy in the search was restricted to firmicutes only (three-component 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).

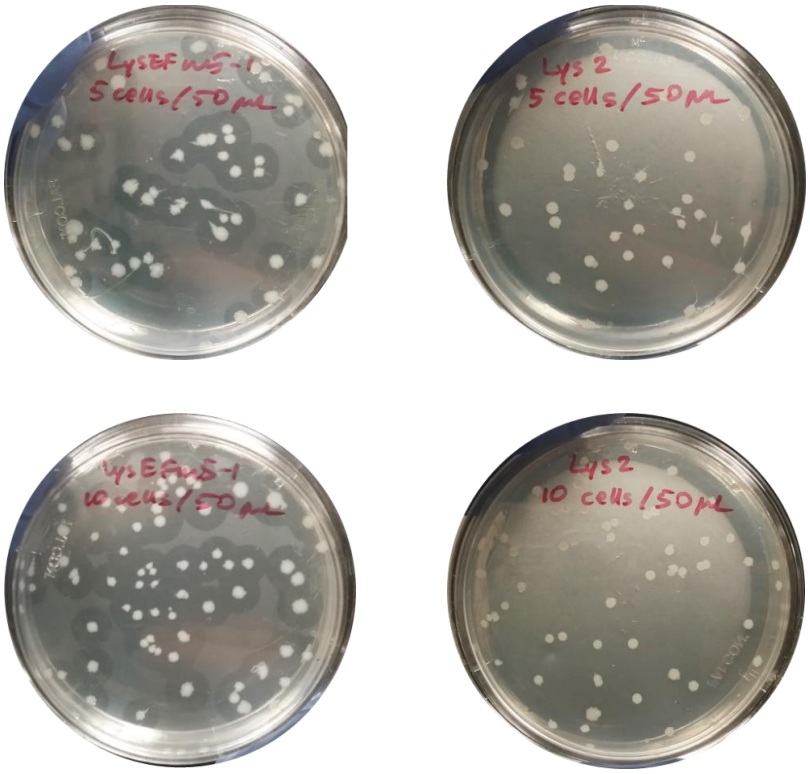


Figure S3.2 E. coli expressing a lysin with alternative specificity do not form halos on LB+kan/VRE/IPTG plates.

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.

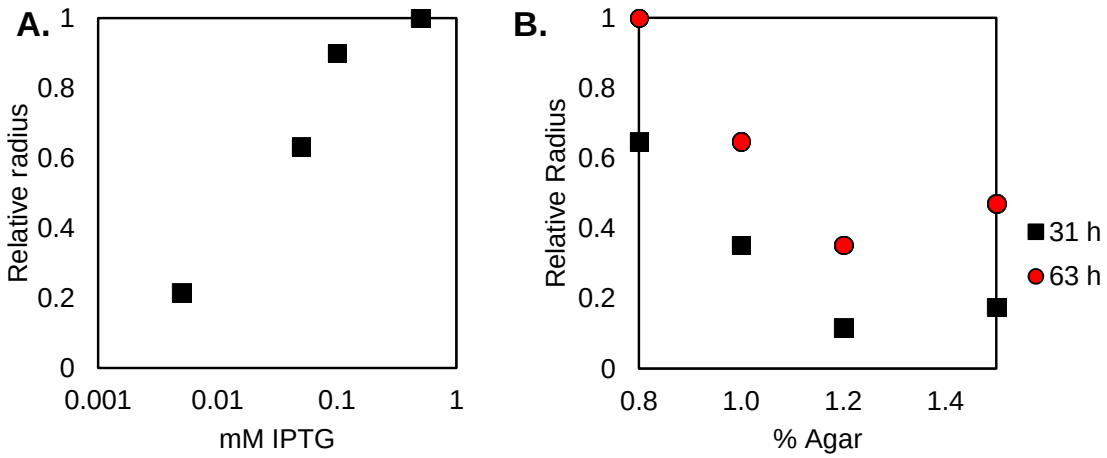
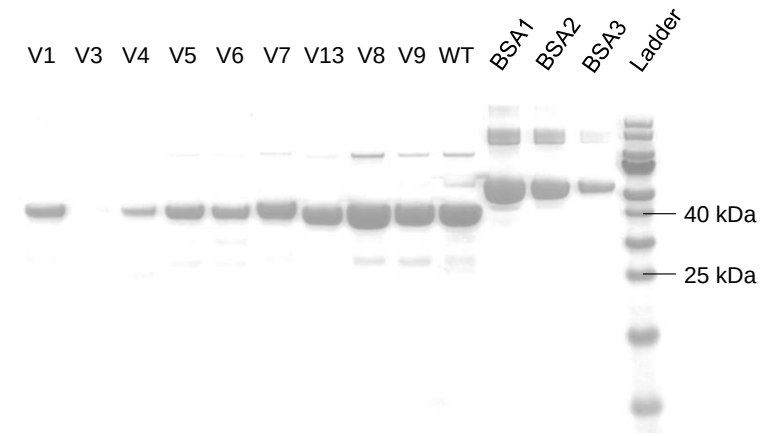
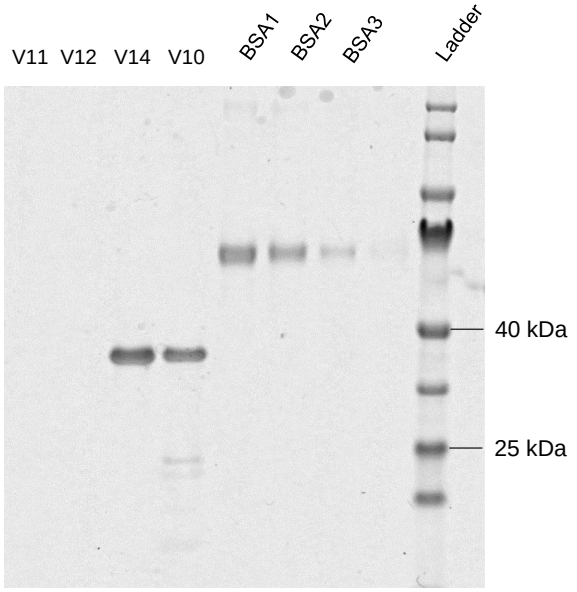


Figure S3.3 Effect of IPTG, agar content, and time on halo radius. Expression-competent *E. coli* containing pET:LysEFm5 were plated on LB + kan agar plates made with 2 w/v% LB, 50 µg/mL kan, 0.05 v/v% of stock autoclaved VRE (~0.3 g/mL cell material suspended in PBS), deionized water, and (A) 0.005, 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.



Variant	Purity
1	0.99
3	N/A
4	0.97
5	0.97
6	0.96
7	0.98
8	0.90
9	0.93
13	0.99
WT	0.91



Variant	Purity
10	0.85
11	N/A
12	N/A
14	1.00

Figure S3.4 SDS-PAGE used to determine variant concentrations.
The bands of variants at ~37 kDa correspond to the expected size of LysEFm5.

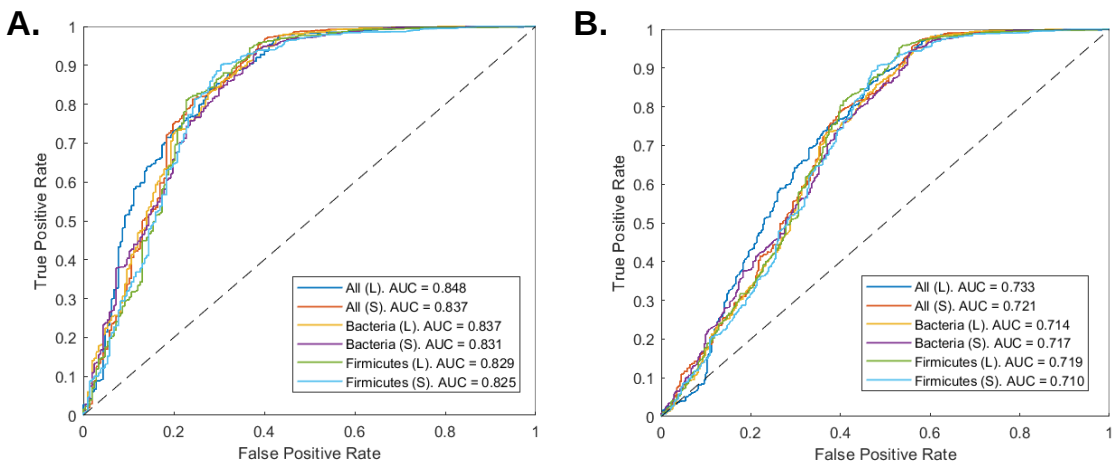


Figure S3.5 Excluding sequences that lack a majority of inactive or active designations improves the classification accuracy of the statistical fitness. ROC curves that quantify the binary classification ability of the statistical fitness parameter are plotted for each starting MSA (Table 3.1), provided 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 (Fig. 3.8).

Table S3.1 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 S3.2 NNK primer sequence identities

Primer Name	Library	Sequence identity (5' → 3')[a]
PRIM01	1	AAGAAGGAGATATACATATGGTTGAG
PRIM02	1	CAGTGATGATGGTGATGGTGGCATCCNNNNNNNNNTTATTAATGGTGGTGATGGTG
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	GATATAGTTCATMNNACCATCGCCATTGGC AGTGTGCCAMNNACCATTGAACGT
PRIM19	8	ACGTTCAATGGTNNKTGGCACACTGCCAAT GGCGATGGTNNKATGAACTATATC
PRIM11	9	CGCCGCAAGGCGSNNTGCTTCTTGTTTTG CAGTGNBATTACCCCAAGT
PRIM20	9	ACTTGGGGTAATVNCACTGCAAAACAAGAA GCANNSCGCCTTGCGGCG
PRIM23	9	GACTTCATATCCTAGATAGTTCATATT
PRIM26	9	AATATGAACTATCTAGGATATGAAGTC

[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.

Table S3.3 High-throughput sequencing primer identities.

Primer Name	Sequence identity (5' → 3')[a]
FA1	TTTCCCTACACGACGCTCTTCCGATCTNNNNGTCGTTTTTCATA ATACTTGGGGT
FA2	TTTCCCTACACGACGCTCTTCCGATCTNNNNNGTCGTTTTTCAT AATACTTGGGGT
FA3	TTTCCCTACACGACGCTCTTCCGATCTNNNNNNGTCGTTTTTCA TAATACTTGGGGT
FA4	TTTCCCTACACGACGCTCTTCCGATCTNNNNNNNGTCGTTTTTC ATAATACTTGGGGT
FA5	TTTCCCTACACGACGCTCTTCCGATCTNNNNNNNNGTCGTTTTT CATAATACTTGGGGT
RA1	GTTCAGACGTGTGCTCTTCCGATCTNNNNAGTCTGATCGTTGC CACA
RA2	GTTCAGACGTGTGCTCTTCCGATCTNNNNNAGTCTGATCGTTG CCACA

RA3	GTTCAGACGTGTGCTCTTCCGATCTNNNNNNAGTCTGATCGTTGCCACA
RA4	GTTCAGACGTGTGCTCTTCCGATCTNNNNNNNAGTCTGATCGTTGCCACA
RA5	GTTCAGACGTGTGCTCTTCCGATCTNNNNNNNNAGTCTGATCGTTGCCACA
FB	AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCT
RB1	CAAGCAGAAGACGGCATACGAGATAAGCTAGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT
RB2	CAAGCAGAAGACGGCATACGAGATGTAGCCGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT
RB3	CAAGCAGAAGACGGCATACGAGATTACAAGGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT
RB4	CAAGCAGAAGACGGCATACGAGATTTGACTGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT
RB5	CAAGCAGAAGACGGCATACGAGATTGACATGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT
RB6	CAAGCAGAAGACGGCATACGAGATGGACGGGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT

[a]N = G, T, A, or C.

Chapter 4 – Sequence-diverse libraries and deep activity and stability analysis inform lysin sequence-function mapping

Daniel T. Tresnak 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 D.T.T and B.J.H. The work presented here is ongoing and will be submitted for publication upon completion of experiments outlined throughout the chapter.

4.1 Abstract

Lysins offer a compelling solution to combat the development of antibiotic resistance given their innate targeting of bacteria. Yet, development of lysins and other antimicrobial proteins is limited both by our inability to evaluate antimicrobial activity in high throughput, due to functional complexity, and the absence of such functional data to inform protein design. We developed a depletion-based assay for screening lysin inhibitory activity and coupled it with a high throughput proteolytic stability assay to assess the activity and stability of ~5×10^4 lysin catalytic domains. The resulting data set was ridge regressed to elucidate sequence-function relationships, and model performance was evaluated across an array of lysin diversification strategies to assess which best informs epistatic statistical protein models. Our work provides an efficient strategy for constructing protein sequence-function landscapes, drastically increases screening throughput for engineering lysins, and yields promising lysins for further development.

4.2 Introduction

Antimicrobial resistance is a critical and worsening public health threat, projected to cause ten million annual deaths globally by 2050[14,118]. One contributing factor to this problem is the limited discovery and development of new antimicrobials, as only one new class of small molecule antibiotic has been discovered in the last two decades and regulatory approval of new antibiotics has remained relatively low in recent years[5,166]. In this antibiotic discovery void, antimicrobial proteins and peptides (AMPs) have garnered interest as potential alternatives or complements to traditional antibiotics because of their abundance, potency, diverse modes of action, and potential for selective targeting of pathogens of interest[18,21,22,167]. Yet AMPs need potency against the pathogen, specificity to avoid detriment to beneficial microbiota, efficient discovery to combat the array of pathogens and outpace resistance, and stability to withstand physiological environments. Thus, protein engineering is needed to advance their functionality for therapeutic use.

While AMPs have attracted much attention as potential next generation antimicrobial therapies, clinical development and FDA approval of AMPs has remained limited[22,168], in part due to the inability to screen antimicrobial activity in high throughput (HT) and model AMP sequence-function relationships. Naturally occurring AMPs offer compelling starting points for developing therapies, but they often require engineering to improve activity, stability, solubility, or other molecular properties to ultimately be translatable[22]. While the field has developed HT assays for screening large (10^6-10^9 variants) protein libraries for stability[28,169], developability[30], and other functions of interest[170–173], screening of AMP

activity has remain mostly limited given the complex and diverse modes of antimicrobial activity[38–41,116,174]. The complex nature of antimicrobial activity and the limited quantity of activity data has further limited development of statistical models to predict AMP activity and inform design of improved AMP variants.

Recently, self-depletion assays have emerged as one HT option for screening AMPs[31,43]. These assays, capable of screening libraries up to 10^5-10^6 variants, express AMPs in a manner such that active AMPs kill the cell that expresses them, while inactive AMPs minimally impact cell growth. Thus, over time, active AMPs are depleted, and inactive AMPs are enriched in the population. This enrichment and depletion can be quantified via HT sequencing of the initial and final populations to quantify AMP activities. These innovative approaches have been shown to differentiate a continuous range of antimicrobial activities[31], providing more precise inhibitory activity quantification and empowering engineering of AMPs.

Towards modeling AMP sequence-function relationships, in the absence of HT functional data, coevolutionary models trained on natural protein sequence diversity have been highly informative. These coevolutionary models calculate sitewise (SW) and pairwise (PW) frequencies of amino acids in large sets of homologous genome-mined sequences to predict the impact of mutations on protein fitness. Coevolutionary models, such as the EVCouplings protein model, have been used to predict protein structure[145,163,175], model protein-protein interactions[149,176–178], inform protein stability[142,179], and guide other protein engineering efforts[180,181]. While these coevolutionary frameworks have

been informative in predicting protein performance for protein families with large genomic datasets, it is not immediately clear how to generate and screen protein libraries from a single lead protein molecule with few homologs to best inform PW models. Additionally, protein fitness predicted by coevolutionary models is a combination of many molecular properties and does not always correlate with functions of interest, such as antimicrobial activity, so we hypothesize that training protein models with experimental data will be more informative towards identifying proteins with desired phenotypes.

Lysins are one particularly compelling family of AMPs. Lysins can selectively kill pathogens via binding and degradation of cell wall peptidoglycan structures[119,182]. These AMPs are produced by bacteriophage to lyse infected cells and release phage progeny, and, because of this role, have evolved high specificity to target species[182]. Natural cell wall binding domains encode lysin specificity, while distinct catalytic domains degrade critical bonds in the peptidoglycan[123,183,184]. The combination of binding and catalytic efficiency yields reasonable potency. Yet, robust therapeutic performance with minimal collateral damage and resistance development would benefit from increased specificity and potency as well as greater physiological stability. Thus, efficient engineering of lysins is needed for this innovative class of antimicrobials to realize their full potential.

To aid in development of AMPs, we developed a platform for mapping the lysin sequence-function landscape. We engineered a HT depletion-based activity assay for screening lysin catalytic domain variants (CDs) and demonstrated its

consistency across multiple controls. We then couple this assay with a yeast display protease stability screen to identify compelling lysins across an array of diverse CD library designs, testing ~5×10^4 CD variants, including several which performed better than a previously engineered *E. faecium*-targeting lysin[40]. Lastly, we conduct ridge regression across library designs to assess how diversification strategy impacts epistatic information within a library, identify sequence-function relationships, and design high-performing variants (Figure 4.1). Our work outlines an efficient approach for sequence-function mapping of proteins and significantly expands the scope of lysin engineering efforts.

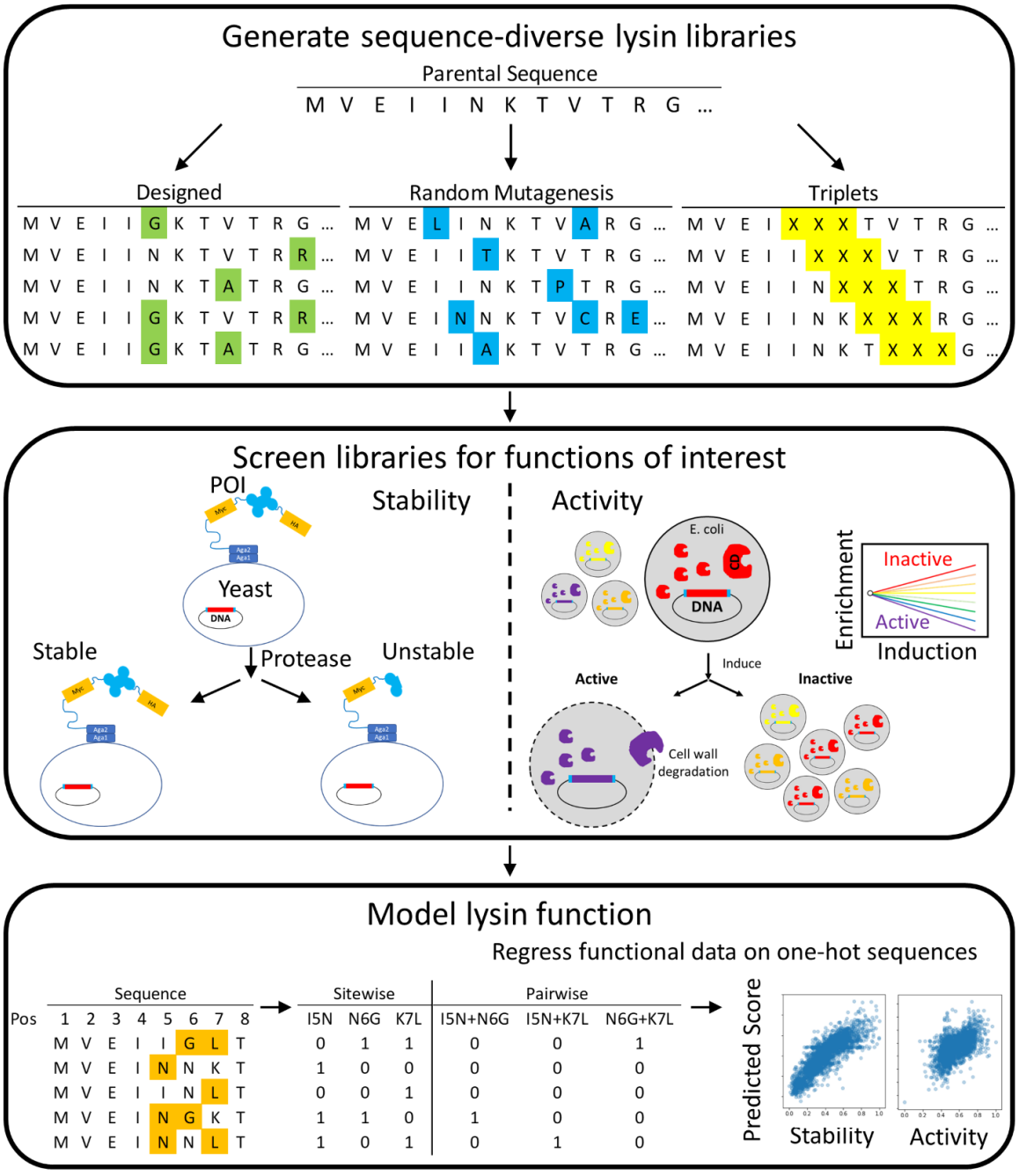


Figure 4.1 Workflow for comparing sitewise and pairwise training information and mapping the CD sequence-function landscape.

Diverse libraries were generated from our lead CD, LysEFm5-V7, via (1) combinatorial insertion of mutations predicted to be beneficial by online protein models (mutations shown in green, defined as Informed sub-library), (2) random mutagenesis (mutations shown in blue), and (3) saturation mutagenesis at all i, i+1, i+2 positions (mutations shown in yellow, defined as Linear Triples sub-library). CD libraries were screened for stability via yeast surface display with protease treatment followed by fluorescence-activity cell sorting (FACS). Libraries

were screened for activity via a novel self-depletion assay where expression of active CDs results in those variants being depleted from the population while inactive CDs are enriched. CD stability and activity scores were then ridge regressed to one-hot-encoded (OH) SW and PW sequence matrices, and model performance was compared across training sets.

4.3 Results

4.3.1 E. coli depletion assay with TorA signal peptide enables differentiation of growth inhibition at clonal and library scales

We hypothesized that a high throughput self-depletion assay for quantifying lysin CD activity could be developed by expressing the CD in the *E. coli* periplasm (Figure 4.2). Several Secretion (Sec) or Twin-arginine translocation (Tat) pathway-targeting signal peptides could enable periplasmic expression to provide access to the *E. coli* peptidoglycan layer (Figure 4.2A)[185,186]. We hypothesized that expression of active CD variants would cause degradation of *E. coli* peptidoglycan, resulting in cell growth inhibition, while expression of inactive CD variants would not impact growth rate (Figure 4.2B). Over time, inactive CD variants would be enriched, and active CD variants would be depleted from the population, which would be quantifiable with Illumina sequencing.

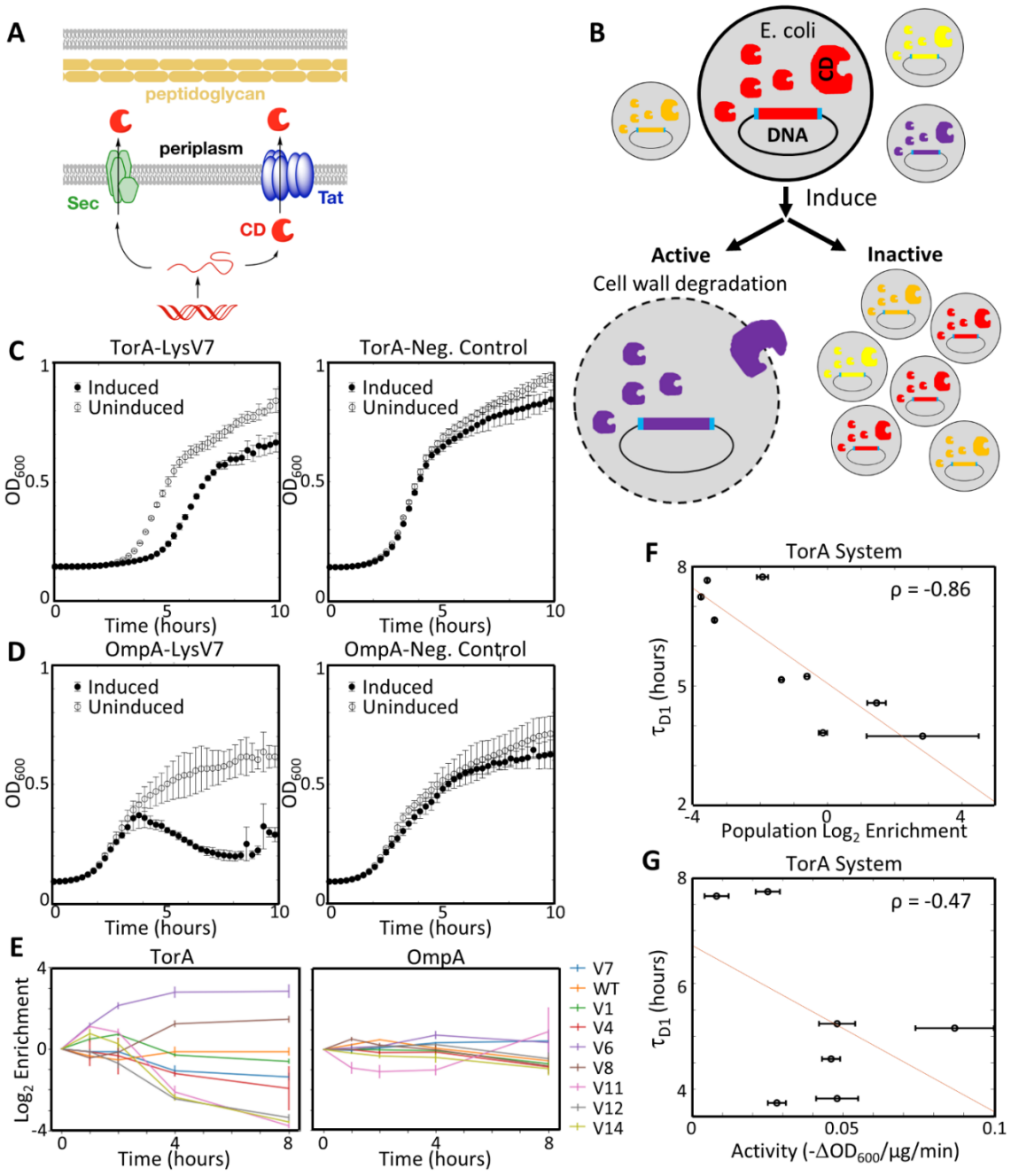


Figure 4.2 Expression of CDs into the periplasm via Tat secretion pathway differentiates growth inhibition at clonal and library scales.

(A) Schematic of Sec and Tat secretion pathways for expression of CDs into the periplasm of *E. coli*. (B) Proposed assay mechanism of screening for CD activity. When expression of a library of CD variants is induced in *E. coli*, active CDs will degrade the cell wall of the cells expressing them, leading to growth inhibition, while inactive CDs will have minimal impact on cell growth. (C and D) Growth curves of *E. coli* expressing LysV7 and a truncated negative control with N-terminal TorA (C) or OmpA (D) signal peptides. Expression was induced with 0.5 mM IPTG.

Error bars show standard deviation across triplicate measurements. (E) Log_2 enrichment of CD controls in proof-of-concept population experiment with either N-terminal TorA or OmpA signal peptide. Data shown are mean values across quadruplicate samples induced at 0.5 mM IPTG. Figures showing 0 mM and 0.1 mM IPTG induction conditions are shown in Supplemental Figure S4.2. (F and G) Comparison of individual growth kinetics to log_2 enrichment in population experiment (F) or previously characterized molecular activity [40] (G) for the TorA expression pathway. Log_2 enrichment values are from 0.5 mM IPTG, 8-hour data point shown in E. Molecular activities were quantified for seven of the nine controls in previous work by Baryakova *et al*. Error bars in enrichment show standard deviation across quadruplicate samples.

To test our hypothesis, we expressed a parental CD variant in the isopropyl-beta-D-thiogalactoside (IPTG)-inducible pET plasmid in *E. coli* with an N-terminal OmpA (Sec pathway) or TorA (Tat pathway) signal peptide (Figure 4.2CD). We hypothesized the OmpA signal peptide would enable secretion into the periplasm given previous successful demonstrations with an array of proteins, including the AMP colicin V[187–189]. However, we also tested the Tat pathway and TorA signal peptide to evaluate whether CDs would be better transported through the cytosol and secreted into the periplasm as folded proteins, which is only achieved via the Tat pathway[190,191]. However, given that neither secretion mechanism is perfectly understood, it was not obvious which pathway would work best for the proposed assay. LysEFm5-V7 (hereafter referred to as LysV7) is a previously engineered CD variant that we selected as our lead molecule in an effort to identify highly functional variants. We acknowledge use of a less optimized lead molecule might aid assessment of how well our workflow improves moderate natural molecules, but we especially aim to advance engineering of highly performant variants. We used an *E. coli* strain expressing the $lacI^q$ gene (New England Biolabs, C3013I) to enable tight control of protein expression, as leaky expression

could result in loss of active AMPs prior to induction[31]. We monitored *E. coli* growth via optical density at 600nm (OD_{600}) at 0 and 0.5 mM IPTG and compared to a truncated inactive negative control. Both expression systems showed a consistent growth inhibition under induced expression of LysV7 compared to uninduced expression and the inactive control (Figure 4.2CD). Induced expression with the TorA signal peptide led to significantly slower *E. coli* growth when compared to the uninduced control, while the OmpA signal peptide displayed similar initial growth rates in the induced and uninduced *E. coli* samples but showed a significant decrease in OD_{600} after 4 hours of growth in induced expression. Both expression systems showed a consistent distinction between the growth of induced and uninduced *E. coli* cells expressing an active lysin CD at long time points, and neither truncated control showed any growth inhibition, supporting further testing of the depletion assay. To further validate these results, an additional three LysEFm5 variants with a range of catalytic activities were constructed and tested in both systems, and similar results were observed (Supplemental Fig. S4.1).

To test whether this depletion assay could differentiate activity within a library of CD variants, we shuttled nine previously evaluated LysEFm5 variants with a range of catalytic activities into the pET expression vector with either N-terminal OmpA or TorA signal peptides (Supplemental Table S4.2). Individual clonal cultures were grown overnight, mixed at an equal cell density, and induced at varying IPTG concentrations. For each system and IPTG concentration, mixed culture aliquots were taken at various time points, DNA was extracted and deep

sequenced via Illumina iSeq100, and the frequencies of all constructs relative to their frequencies in the 0-hour time point were analyzed. The TorA signal peptide expression system showed consistent enrichment and depletion results across all replicates with a 98-fold difference in enrichment from the most- to least-enriched sequence at 8 hours and 0.5 mM IPTG induction condition (Figure 4.2E, Supplemental Figure S4.2). Results were consistent across replicates, with a median relative standard deviation of 14% and 11% at the 4- and 8-hour time points with 0.5 mM IPTG. Alternatively, analysis of the results for the OmpA signal peptide expression system showed only minimal enrichment or depletion across all tested times and induction levels, suggesting that the OmpA expression system cannot successfully distinguish between CDs of varying activity levels (Figure 4.2E, Supplemental Figure S4.2). To evaluate how well the TorA population experiment replicated individual growth behavior, population enrichment scores were compared to the first doubling time (τ_{D1}) and differential first doubling time of induced vs. uninduced samples ($\Delta\tau$) for each of the controls individually. We found that population enrichment strongly inversely correlated with induced τ_{D1} (Figure 4.2F, $\rho = -0.86$) and $\Delta\tau$ (Supplemental Figure S4.3, $\rho = -0.74$), consistent with population experiments accurately representing individual clonal growth behavior.

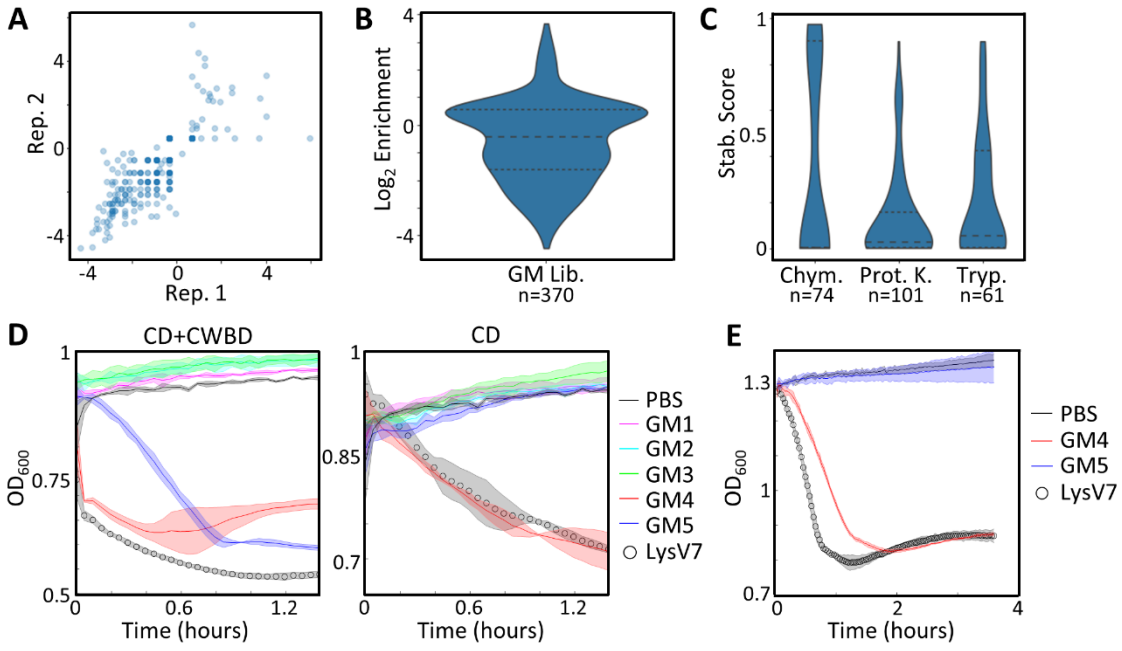
Lastly, we compared τ_{D1} (Figure 4.2G) and $\Delta\tau$ (Supplemental Figure S4.3) to previously reported lysin catalytic activities [40] to assess if the depletion assay correlated with activity. Two modes of performance were observed: very long τ_{D1} and $\Delta\tau$ were observed for two minimally active variants (V4 and V14) whereas a range of moderate times correlated with catalytic activity measurements amongst

the other five variants ($\rho = 0.66$). We suspect the variable outcomes are caused by multiple factors, most significantly that growth inhibition in this assay is related to a combination of a molecule's catalytic activity, producibility, and stability. In our previous work, V4 and V14 were the only tested variants with expression yields <0.2 mg/L (unpublished data). Additionally, while neither of these variants were tested for thermal stability, previously literature has demonstrated a correlation between stability and producibility[30]. Here, correlation is observed between lysin CDs of comparable producibility; so less stable or less producible molecules may inhibit growth when expressed in this assay via mechanisms unrelated to catalytic activity. Thus, we expect enrichment and depletion scores to need to be supplemented with some functional metric of molecular stability and producibility. For further use of this depletion assay here, we coupled our activity screening process with an on-yeast proteolytic stability assay to identify compelling lysins and inform protein sequence models.

4.3.2 Coupling depletion and stability assays identifies compelling lysins in genome-mined library

To test the ability of our proposed method with coupled depletion and stability assays to assess CD performance, we constructed a library of genome-mined CDs (termed GM library). We included the 100 homologs most similar to LysV7 with sequence similarity ranging from 55-99% as well as chimeras recombined via a highly conserved region at positions 87-92 in the LysV7 sequence. To assess inhibitory activity, the GM library was screened in the depletion assay in duplicate, and activity scores were defined as the $\log_2$ enrichment from the initial sample to the final sample taken at eight hours of growth in media with 0.5 mM IPTG (Figure

4.3AB). Strong inhibitors will have enrichment scores less than zero and weak/non-inhibitors will have enrichment scores closer to zero or greater than zero. To enhance rigor, we required variants to be depleted in both replicates, enriched in both replicates, or have a standard deviation of less than 0.5 (to evaluate variants with depletion scores near 0). After filtering, we retained 370 CD sequences with a range of -4.5 to 3.7 (mean: -0.6, median: -0.44) and observed a Pearson's correlation of 0.82 between replicate 1 and 2 scores (Figure 4.3A). 40% of sequences fell between 0-1, as expected for inactive sequences that would be moderately enriched in the population (Figure 4.3B).



F

Variant	Chym.	Tryp.	Prot. K	Aver. Stab.	Depletion	Cat. Activity (-ΔOD$_{600}$/μg/min)	T_m (°C)
GM1	0.006	0.012	0.061	0.026	-2.67		
GM2	0.001	0.207	0.021	0.076	-3.35		
GM3	0.001	0.137	0.274	0.137	-2.23		
GM4		0.742			-2.01	0.111 ± 0.003	49.5 ± 0.1
GM5	0.501	0.255	0.259	0.338		0.018 ± 0.001	45.0 ± 0.1
LysV7	0.814	0.826	0.621	0.753		0.094 ± 0.010	47.2 ± 0.3

Figure 4.3 GM library displays a range of inhibitory activities but poor stability.

(A) Comparison of sequence scores for depletion assay replicates 1 and 2. (B and C) Distributions of GM library performance in depletion assay (B) and stability assays (C). Number of sequences observed across all replicates within each assay are shown below. (D) Turbidity reduction assay results for variants GM1-5 and LysV7 when produced as CD-CWBD fusions or CDs. Turbidity reduction assay was conducted with 3-10 µg of CD-CWBD or CD in 200 µL PBS. (E) Turbidity reduction assay for LysV7, GM-4, and GM-5 as CD-CWBD fusion with 0.36-0.92 µg in 200 µL PBS to quantify CD activity. Standard deviations across triplicate measurements are shown in transparent borders in D and E. (F) Variant performance across all assays. Catalytic activity was quantified from replicate curves shown in E. T_m values were calculated via Sypro denaturation assay. Gray cells were either not present in HT assays or not quantified in catalytic activity and Sypro denaturation assays.

In parallel to depletion assay screening, the GM library was screened for stability to trypsin, proteinase K, and chymotrypsin via yeast surface display. After induction of yeast surface expression, the lysin variants were treated with one of the specified proteases for 15 minutes at 37 °C, then washed and labeled with antibodies specific for epitope tags at the N- and C-termini of the lysin. Stable variants resist proteolysis and exhibit high N-terminal labeling whereas proteolysis of unstable variants reduces N-terminal signal relative to the yeast-anchored C-terminus. Flow cytometric sorting stratified variants: top 5% of N-terminal:C-terminal ratio (score: 0.975), next 10% (0.9), all remaining N-terminal-positive cells (0.425), and N-terminal-negative cells (0) (Supplemental Figure S4.4). The four tiers were deep sequenced, and CD variants were assigned scores based on their average observance across gates. The GM library displayed relatively poor stability across all proteases with median scores below the threshold for intact lysin (Figure 4.3C). This poor stability is expected given the high presence of chimeric CDs in the GM library. Yet several sequences displayed stabilities comparable to LysV7 (Supplemental Figure S4.5), which had an average stability score of 0.75.

To identify compelling lysins, we analyzed protease stabilities in combination with depletion scores (Supplemental Figure S4.6). We identified the three variants with the best depletion scores (termed variants GM1-GM3) among sequences seen across all protease and depletion replicates. To assess the relative importance of depletion and stability scores, we also identified a fourth variant (GM4) which had a moderate depletion score of -2.0 and relatively strong trypsin stability score of 0.74 as well as a fifth variant (GM5) which was not observed in the depletion assay but had moderate stability scores across all proteases (average stability score = 0.34). All five variants as well as the parental LysV7 were produced in *E. coli* as both C-terminal His-tagged CDs and C-terminal fusions to the native LysEFm5 cell wall-binding domain (CWBD). Purified variants were tested for catalytic activity degrading crude *E. faecium* cell wall via turbidity reduction assays[40]. Interestingly, GM1, GM2, and GM3 displayed no catalytic activity as either CDs or fusions to CWBD, while GM4 displayed activity superior to LysV7 in both purified forms (Figure 4.3D-E). GM5 displayed activity at high concentrations as a fusion to the CWBD, but no activity was observed from the GM5 CD. Further, when assessed for thermal stability via a Sypro thermal denaturation assay, GM4 showed an enhanced melting temperature compared to LysV7, while GM5 was slightly less stable, in agreement with on-yeast stability scores (Figure 4.3F). GM4 and GM5 are both chimeras of natural CDs, resulting in highly distinct sequences from LysV7. GM4 and GM5 had 55% and 62% sequence identity (66% and 76% sequence similarity) to LysV7, respectively, and 96% and 83% sequence identity (98% and 88% sequence similarity) to their

closest homolog, respectively. Discovery of the variant GM4 with enhanced activity and stability from a genome-mined chimeric library, despite drastic under-sampling, using a combination of a bacterial self-depletion assay and a yeast-displayed protease resistance assay highlights the utility of this merged platform and the library design.

4.3.3 Lysin libraries efficiently sample catalytic domain sequence space

Coevolutionary models trained on natural sequence data have proven very powerful in predicting protein performance[40,50,142,181]. We sought to expand the utility of such models to protein families lacking large sets of natural homologs by developing a straightforward framework for protein diversification and screening to generate protein sequence-function landscapes. Following confirmation that our screening approach successfully identifies compelling CDs, we wanted to use this approach to screen larger libraries of CDs to inform SW and PW models, testing how varying protein diversification strategies and functional data inform such models, while also identifying compelling CDs. In library design, we aimed to balance sequence diversity with the functional hit rate, acknowledging that we would need diverse sequences and at least a moderate number of functional variants to sufficiently inform PW models. It was not obvious which protein diversification strategy best achieves these aims, so we comparatively evaluated multiple CD libraries.

We generated three libraries each with theoretical diversity of ~10^6-10^7: (1) combination of mutations predicted to be beneficial by EVCouplings and PROSS[181,192] (termed Designed library), (2) saturation mutagenesis at all

three-amino acid sets from position 5 to 180 (termed Triplet library), or (3) random mutagenesis with a 0.35-0.40% error rate resulting in 2.0-2.2 mutations per gene (termed RM library). We hypothesized that combinatorial mutations that were multiple positions away from each other in the Designed and RM libraries would help assess long-range epistatic interactions, while the Triplet library would address potential local interactions. Further, we hypothesized that the moderate mutation rates (and homolog and structural guidance in the Designed library) would result in sufficient functional frequencies to better inform protein models. All three libraries were screened separately, and analysis of depletion and stability assay results for each library followed identical filtering and scoring as described for the GM library.

Within each library, we identified sequences which fit all intended, allowing quantitative comparisons between design performance. For example, within the Designed library, variants with EV-informed mutations performed significantly better than variants within all other designs (p-value = 7.9×10^{-58}, 3.5×10^{-197}, and 1.3×10^{-4} compared to the whole population, PROSS-informed variants, and Triplet variants, respectively, as assessed by the Mann-Whitney U test, Figure 4.4A). Additionally, while the median depletion score for EV-informed variants in the Designed library was comparable to all variants in the full population, EV-informed variants had a higher fraction of total variants with depletion scores < -1, as 17% of EV-informed variants had depletion scores < -1 compared to 14% in PROSS and 12% in the whole population (p-value = 0.02 compared to the whole population, Figure 4.4B). Notably, variants matching the *i*, *i+1*, *i+2* library design

also performed significantly better than the full population in the depletion assay (p-value = 4×10^{-12}). Within the RM library, PROSS-informed variants had a higher median stability score than EV-informed variants (p-value = 0.006), though EV-informed variants had a higher fraction of variants with stability scores > 0.75 (17% of EV-informed compared to 1.4% of PROSS-informed) and a lower median and mean depletion score than the full population, indicating more inhibitory variants (p-value = 0.001, Figure 4.4CD). Only 30 and 9 EV- and PROSS-informed variants were identified within the Triplet library stability results, respectively, and these populations were not distinguishable from the whole population (Figure 4.4E). Only 43 and 17 EV- and PROSS-informed variants were identified within the Triplet library depletion results, and these subsets performed worse than the full library with a higher mean and median depletion score (p-value = 0.001 and 0.01 for the EV- and PROSS-informed mutants compared to the whole population, Figure 4.4F). These results demonstrate that both EVCouplings and PROSS are informative in designing high-functioning libraries.

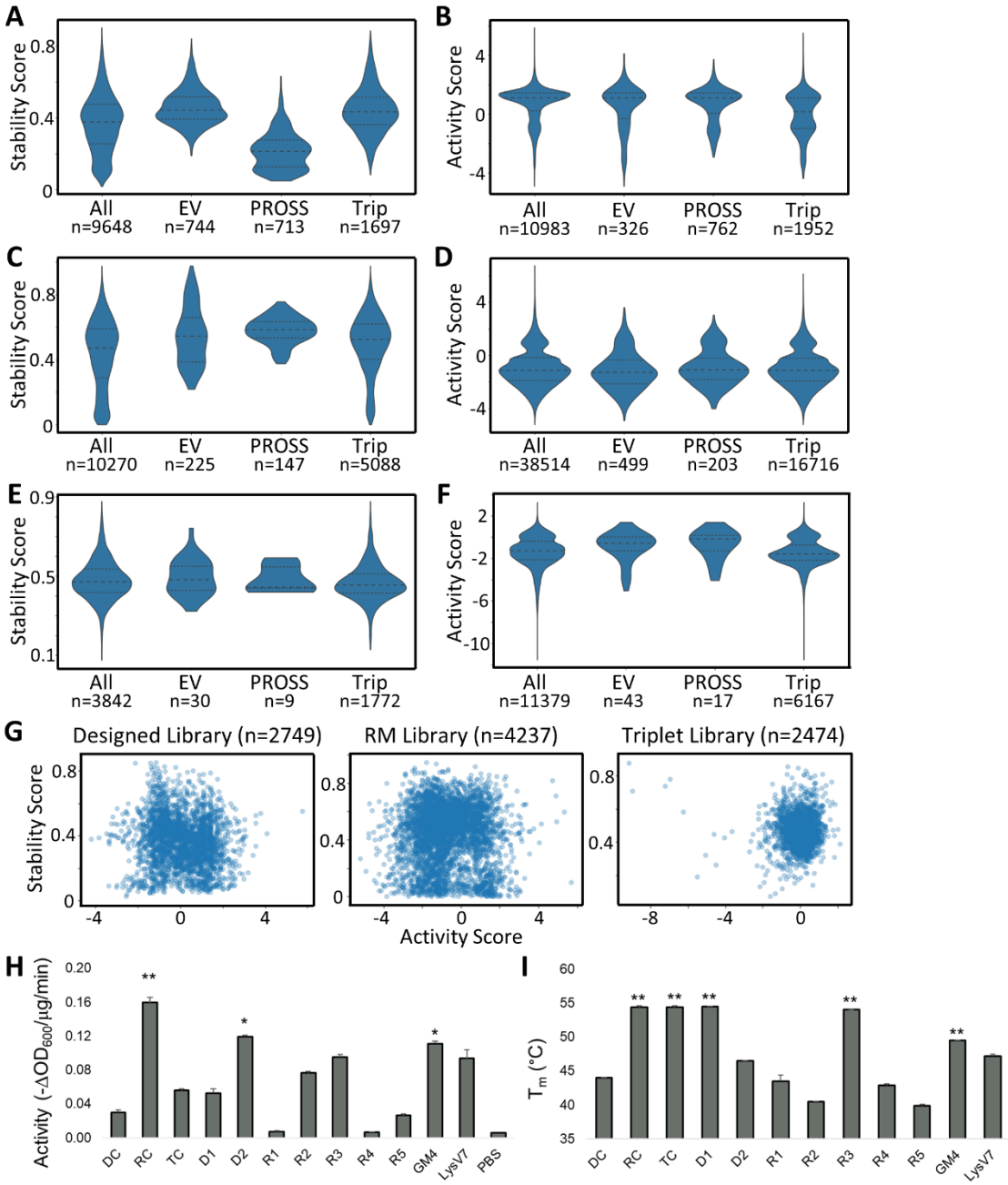


Figure 4.4 Diverse CD libraries display broad performance across all HT assays. (A-F) Distribution of depletion activity and stability scores for Designed (A and B), RM (C and D), and Triplet (E and F) libraries. Stability scores are the average of stability scores across trypsin, chymotrypsin, and proteinase K. 'All' denotes all sequences within a library, 'EV' denotes variants containing EVCouplings-informed mutations, 'PROSS' denotes variants containing PROSS-informed mutations, and 'Trip' denotes variants containing mutations at some combination of positions i, i+1, i+2 (Triplets). (G) Comparison of stability and activity scores for sequences seen in both assays within Designed, RM, or Triplet libraries. Stability

scores shown are the average stability score across trypsin, chymotrypsin, and proteinase K assays. (H and I) Catalytic activity (H) and melting temperature (T_m, I) values of high performing variants from each library. Variant GM4 from the genome-mined library is included for comparison. Catalytic activity values were quantified via turbidity reduction assays with 0.5 µg of CD-CWBD fusions in 200 µL PBS. Melting temperatures were quantified via Sypro thermal denaturation assay. * denotes p-value < 0.05; ** denotes p-value < 10^{-5}.

We coupled stability and depletion results to identify compelling CD variants for individual testing (Figure 4.4G). Only one variant (W30G) in the Designed library and one variant in the Triplet library (Y58R, I59P, D60S) had an average stability score >0.8 and depletion score < -2. However, an additional four variants in the Designed library and six variants in the Triplet library had average stability scores > 0.6 and depletion scores < -1.5. 11 variants in the RM library, 10 of which contained only single-site substitutions, had average stability scores > 0.8 and depletion scores < -2, with 45 total variants with stability scores > 0.7 and depletion scores < -2. Our sequencing strategy prevented identification of LysV7 in the Triplet library, but our parental sequence had activity scores of -0.46 ± 0.27 and -0.26 ± 0.17 and stability scores of 0.47 ± 0.10 and 0.57 ± 0.06 when observed in the Designed and RM libraries, respectively. To assess how well these coupled metrics identify high performing CDs, we selected 11 high performing variants across the three libraries to construct as CD-CWBD fusions, produce, and test. We also identified and constructed the CD variants (denoted as variants DC, RC, and TC) from each library that were observed in both stability and depletion assays and had the highest average stability score, regardless of depletion score, to assess the benefit of coupling stability and depletion assay data (all variants listed in Supplemental Table S4.12).

When assessed for catalytic activity via the turbidity reduction assay, both variants identified from the Designed library (D1 and D2) and three of five variants from the RM library (R2, R3, and R5) showed catalytic activity above the PBS buffer negative control (Figure 4.4H). Additionally, all three variants with the highest average stability scores (DC, RC, and TC) from each library showed catalytic activity above the negative control. Several of these variants also showed melting temperatures superior to that of LysV7 as well, with four variants with $T_m \geq 54$ °C. However, interestingly, all four variants from the Triplet library (T1-T4) were only able to be produced as truncated variants of molecular weight of ~25 kDa as assessed via sodium dodecyl sulfate–polyacrylamide gel electrophoresis (SDS-PAGE) and displayed no catalytic activity. We suspect this truncation is caused by sets of mutations which may lead to unfolding and proteolytic cleavage during production in *E. coli*, though there is not a consensus mutation across all four T1-T4 variants that would cause this. Collectively, variants RC, D2, and R3, in addition to GM4 from the genome-mined library, show promise as variants with enhanced activity and/or stability compared to the parental LysV7. Interestingly, none of these variants share any common mutations, so additional variants with combinations of these mutations (R17L, F71L, S34T, V35A, F54V, K62D) are also compelling for engineering further enhanced catalytic domains.

4.3.4 Deep functional data informs sitewise and pairwise protein models

To evaluate how well our diversified libraries and functional data informed protein sequence models, we trained one-hot-encoded SW and PW ridge regression models on the entirety and subsets of the Designed and RM libraries with stability

and depletion data (Figure 4.5). Modeling focused on only the Designed and RM libraries because they had comparable numbers of sequences in both activity and stability assays, whereas the Triplet library had ~3-fold less sequences in the stability data set. SW models encoded every mutation from the parental sequence while PW models encoded all SW information in addition to every observed pair of mutations from the parental sequence. Ridge regression was used to aid in overfitting, as attempts to use linear, lasso, or elastic net modeling all resulted in overfitting. A simple ridge regression model trained on the Hamming distance for each sequence was also included for comparison as a control model with minimal sequence information (Supplemental Figures S4.7-8). For comparing model performance, we analyzed mean squared error (MSE), Pearson's correlation, and the coefficient of determination (R^2) and found that MSE and Pearson's correlation could be easily skewed by high density of data around the mean, so analysis presented here focuses on R^2 values which tended to be more indicative of model performance. All other metrics and model comparisons are included in the supplemental information (Supplemental Figures S4.7-10).

Both SW and PW models predicted stability significantly better than activity, despite often having more data points in the depletion activity datasets (Figure 4.5). SW and PW stability models trained on the entirety of the Designed library achieved $R^2 = 0.69 \pm 0.02$ and 0.74 ± 0.02, respectively, while those for the RM library has $R^2 = 0.87 \pm 0.01$ for both SW and PW models. To ensure these models were learning sequence trends beyond trypsin and chymotrypsin amino acid cleavage site preferences, we trained regression models of the frequency of such

sites and found these models were not predictive (R^2 = 0.01 ± 0.01 – 0.07 ± 0.02, Supplemental Figure S4.11). Interestingly, we found that, in most cases, SW and PW models performed nearly identically, with the notable exception being in predicting stability results of the Designed library, where PW models were more predictive (Figure 4.5EF, p-value = 3.9×10^{-6}). We hypothesize this improved performance of the PW model amongst the Designed library is caused by the higher rate of combinatorial mutations (mean Hamming distance = 3.14 compared to 1.54 for the RM library, Supplemental Figures S4.12-13), providing more PW sequence information, as well as focus on mutations with predicted benefit. Unfortunately, nearly all activity models trained on sequences observed in the depletion assay yielded poor prediction (R^2 = 0.03 ± 0.04 and 0.00 ± 0.05 for SW and PW in the Designed library and R^2 = 0.24 ± 0.02 and 0.24 ± 0.02 for the RM library). As noted in the commentary for Figure 4.4G above, we expected the addition of stability data to aid prediction of activity data. Indeed, when we analyzed SW and PW models trained on sequences seen in both stability and depletion assays, SW and PW prediction of depletion results improved significantly (Figure 4.5EF). SW and PW activity models yielded R^2 = 0.25 ± 0.06 and 0.29 ± 0.07 for the Designed library and R^2 = 0.37 ± 0.05 and 0.36 ± 0.05 for the RM library, respectively (p-value = 9.5×10^{-9} and 2.6×10^{-9} for Designed library SW and PW models and p-value = 3.7×10^{-7} and 1.1×10^{-6} for the RM library SW and PW models, respectively, when compared to models trained on the full depletion data sets). To assess whether this is the result of reduced sequence diversity or complexity due to the lower number of total sequences, we trained models on a comparable

number of random sequences from the full depletion set, but these models returned to their initial poor predictive performance (SW and PW models R^2 = 0.04 ± 0.07 and -0.02 ± 0.09, respectively, for the Designed library and R^2 = 0.16 ± 0.06 and 0.16 ± 0.05, respectively, for the RM library, Supplemental Figure S4.14). Strong prediction of stability results was retained for both SW and PW models when trained on only sequences observed within both stability and depletion assays. Within this subset of sequences, we also wanted to assess the benefit of including activity information to predict stability or including stability information to predict activity. We tested SW and PW models where OH sequence matrices also included either stability information (when predicting activity) or activity information (when predicting stability). However, we ultimately found no significant predictive benefit from either set of information (Supplemental Figures S4.15-16). Thus, it is not obvious why regression models would better be able to predict activity of sequences that were also observed in our stability data set.

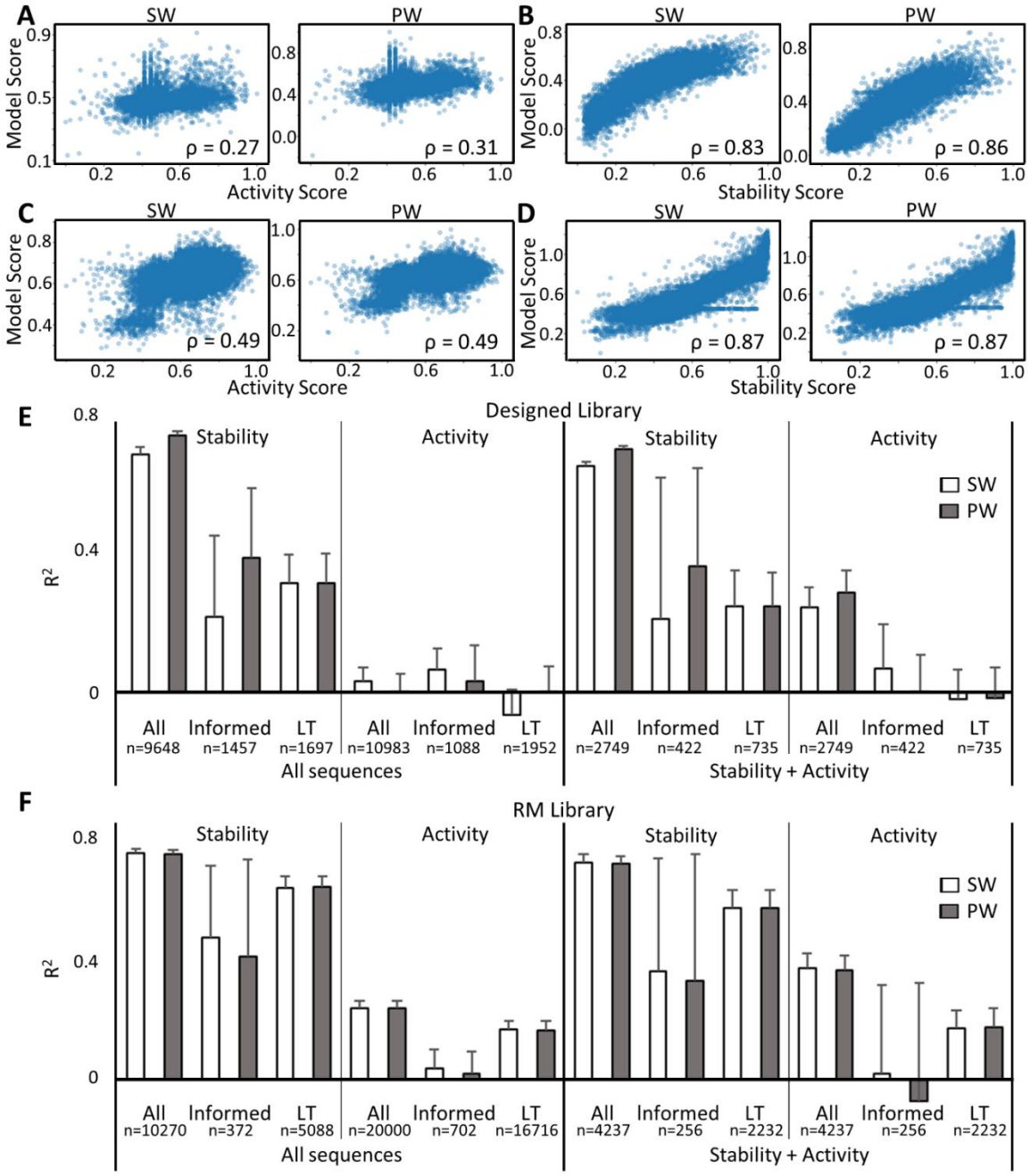


Figure 4.5 SW and PW modeling results across all data sets and sub-libraries of the Designed and RM libraries.

(A-D) Comparison of SW and PW model predicted and experimental values when predicting activity (A and C) or stability (B and D). The Designed library results are shown in A-B, and EP library results are shown in C-D. Pearson's correlation values are shown the bottom right corner for all model results. Activity training for the RM library was conducted with a random sampling of 20000 sequences from the entire data set of 38514 sequences due to increasing size of model parameter space. All assay scores were normalized to [0, 1] for modeling. (E and F) R^2 values

for SW and PW models trained on all and subsets of Designed (E) and RM (F) libraries for stability and activity. Models were trained on all sequences observed within activity or stability assays (left) or on all sequences observed in both activity and stability assays (right). Total number of sequences within each group is included below group titles. S and P denote SW and PW models, respectively.

We also wanted to assess how model performance would change when trained on subsets of sequences which fit specific library designs (Figure 4.5EF). To evaluate this, we retrained SW and PW models for stability and depletion on (1) only sequences which fit the EVCouplings- and PROSS-informed library design (defined as the Informed sub-library) or (2) sequences which fit the *i*, *i+1*, *i+2* library design (defined as the Linear Triples sub-library). Within the Designed library, the R^2 values of SW stability models trained on the Informed and Linear Triples sub-libraries dropped to 0.22 ± 0.24 and 0.32 ± 0.08, respectively (68 and 54% decreases from the models trained on all sequences). R^2 values of PW models dropped similarly to 0.39 ± 0.20 and 0.32 ± 0.08 (48 and 57% decreases). Further, both sets of models performed significantly worse than when trained on a comparable number of random samples from the entire library, as a training set of 1000 random sequences from the entire library yielded R^2 values of 0.65 ± 0.02 and 0.68 ± 0.02 for SW and PW models, respectively (Figure 4.5E, Supplemental Figure S4.17). SW and PW activity models for the entire Designed library, including sub-libraries, were non-predictive ($R^2 < 0.07$ for all models). Within the RM library, stability models trained on the Linear Triples sub-library retained most of the predictive power of those trained on the entire data set, with $R^2 = 0.64 \pm 0.04$ for both SW and PW models (84% of the $R^2 = 0.76$ for SW and PW models trained on all sequences). Models trained on the Informed sub-library offered only moderate

predictive power (R^2 = 0.48 ± 0.24 and 0.41 ± 0.32, Figure 4.5F) However, within the RM library, approximately half of the sequences within the stability data set (5088 of 10270 sequences, 50%) and most sequences within the activity data set (16716 of 20000 sequences, 84%) fit the Triplets library design due to the high presence of single-site substitutions, while only 372 sequences (4%) in the stability data set and 702 sequences (4%) in the activity data set fit the Informed sub-library design. Thus, models trained on the Linear Triples sub-library had significantly more sequence information to inform model predictions. Yet, stability models trained on 500 random sequences from the entire data set still provided better predictive power than the those trained on the Informed sub-library, demonstrating that non-intended variants within libraries add significance value for informing models (SW and PW R^2 = 0.61 ± 0.02 and 0.58 ± 0.02, respectively, Supplemental Figure S4.18).

Lastly, we wanted to assess how well our experimental data correlated with EVCouplings-predicted fitness. We used the EVCouplings model to predict fitness values for all single- and double-mutants observed in the Designed and RM libraries and compared these to activity and stability scores. Within both libraries, we found no correlation between predicted fitness and activity scores (Pearson's correlation $\rho = 0.05$ for both libraries, Supplemental Figures S4.19-20). Among the stability data, correlation was slightly better within the RM library ($\rho = 0.39$), but there was again no correlation within the Designed library ($\rho = -0.03$). We also assessed whether EVCouplings-predicted fitnesses showed better correlation with sequences observed in both stability and activity data sets, since SW and PW

models were better able to predict the activity of these sequences. However, we saw only moderate improvement among depletion data in the RM library (from $\rho = 0.05\ to\ 0.11$) and no significant change in the other data sets. Lastly, we evaluated whether EVCouplings-predicted fitnesses provided predictive information in training of SW or PW models. We appended predicted fitnesses to either OH SW or PW matrices and repeated model training for predicting activity or stability, but we saw no change in correlation or R^2 for either the Designed or RM libraries (Supplemental Figures S4.21-22).

4.3.5 Functional data and pairwise models identify divergence in amino acid preferences for stability and activity

We sought to use our models to identify sequence-function relationships in CDs. We selected the SW and PW models trained on sequences present within both activity and stability data sets for both the Designed and RM libraries. This enabled prediction of both functions and provided the highest predictive power for predicting activity with only a slight decrease in prediction of stability (Figure 4.5E and F). We identified each set of model coefficients and evaluated SW and PW amino acid preferences (Figure 4.6).

Analysis of amino acid preferences across activity and stability models yields some general trends. Regarding stability, we see a high presence of deleterious mutations from positions G23-F26 and N83-E90, as each site, with the exception of G88, has at least 3 mutations predicted to be highly detrimental to stability (Figure 4.6A). Interestingly, even subtle mutations, such as N83S, N83T, N85S, N85T, and N85D, are predicted to be deleterious, suggesting these residues may be critical for CD folding. Conversely, we see a broad tolerance for mutations

between sites W30-T40 and A44-N48, with >1 substitutions being predicted as beneficial at sites W30, N32, S33, V35, Q37, T40, A44, M45, N46, N47, and N48. Analysis of the activity landscape also suggests a fair degree of tolerance to mutagenesis, with multiple individual amino acid substitutions having highly positive model coefficients, such as V24A, V25A, N32D, A52V, A52S, and N108S, to name a few (Figure 4.6B). However, when these observations are considered together, we see a general divergence in amino acid preferences (Figure 4.6C, Supplemental Figure S4.23). This is apparent with substitutions such as A44V, which is beneficial for stability and deleterious for activity, or N72D and G73D, which are predicted to be highly deleterious for stability but tolerated or beneficial for activity, for example.

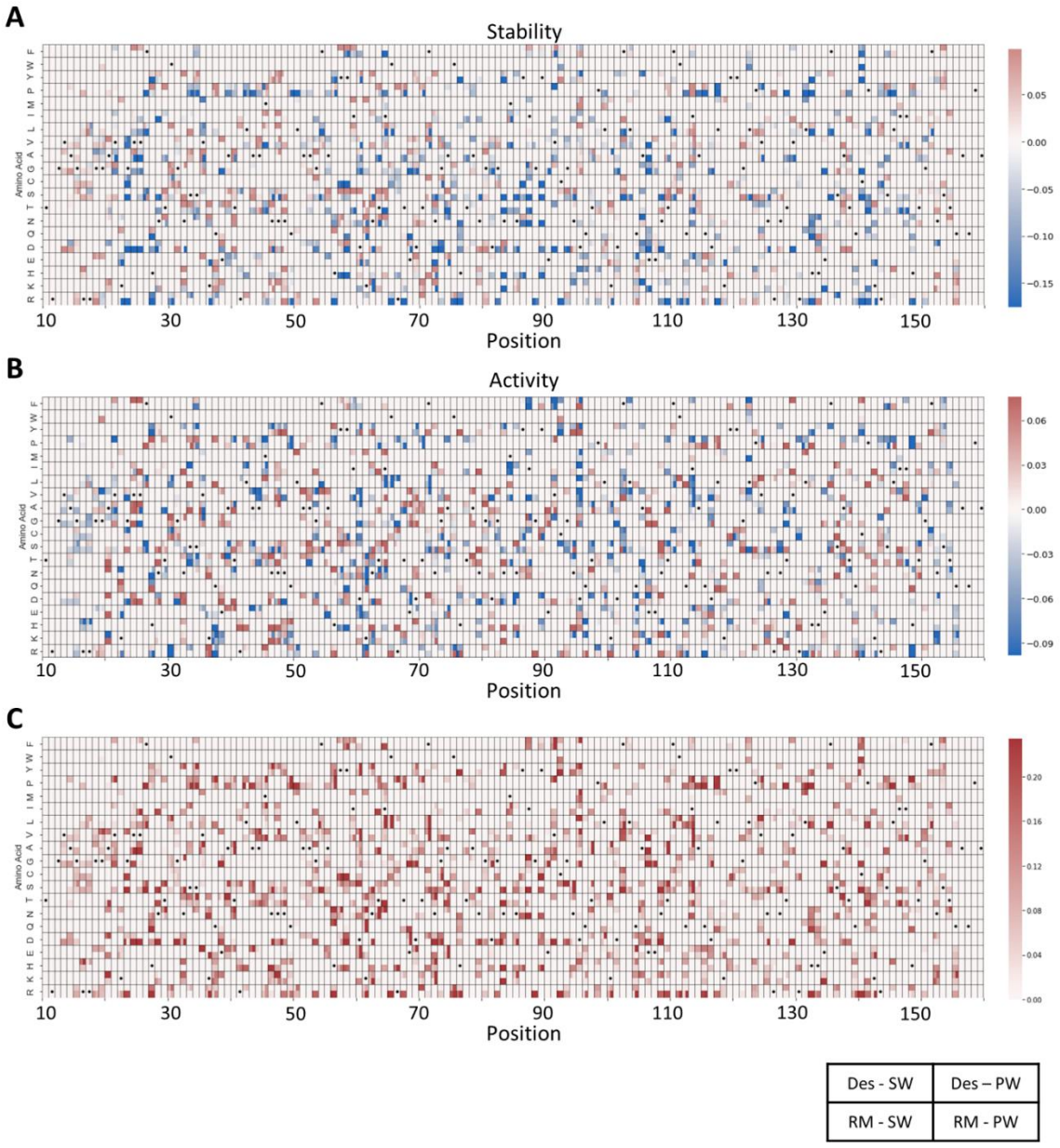


Figure 4.6 Heat maps show divergence in amino acid preferences for activity and stability.

(A, B) Heat map of model coefficient terms for models predicting stability (A) or activity (B). (C) Heat map displaying difference in terms for A and B. For each position/amino acid combination, the top left and top right corners are the coefficients for the SW and PW models trained on the Designed library, respectively, and the bottom left and bottom right corners are the coefficients for the SW and PW models trained on the RM library. Black dots represent the amino acid present in the parental LysV7 sequence. Models used to generate heat maps were trained on the set of sequences observed in both activity and stability assays.

In spite of the functional tradeoff apparent in the stability and activity landscapes, we wanted to assess whether our models could predict variants with high stability and activity that were not observed in our data set. We identified the amino acid substitutions predicted to be most beneficial for stability for PW stability models from either the Designed or RM libraries. For each library model, we then iteratively combined these mutations, identifying the 100-200 variants with the highest predicted stability scores at each i^{th} round to seed the next set of variants with *i+1* mutations, generating variants with up to seven combinatorial mutations. At each round, we compared the rank of our highest performing variants with *i+1* mutations to their highest performing parental clones with *i* mutations. We consistently found that the variants with the highest predicted stability scores arose from parents with the highest predicted stability scores, suggesting that our use of the top 100-200 variants to seed iterative rounds was sufficient to identify highly functional variants (Supplemental Figures S4.24-25). We identified variants with highest predicted stability scores, then compared predicted activity and stability scores to generate high performing variants for individual testing (Supplemental Figure S4.26 and 27). Interestingly, stable variants predicted by the Designed library tended to have more comparable, high activity scores, whereas variants predicted by the RM library exhibited a bimodal distribution of predicted activities. This behavior was caused by the presence of the C92Y substitution, which is predicted to be beneficial for stability but deleterious for activity within the RM library. Our process ultimately yielded two variants containing seven mutations each from each library (DP1-2, RP1-2) predicted to be highly active and stable.

We synthesized oligonucleotides encoding these variants, and constructed, produced, and characterized them as CD-CWBD fusions. Unfortunately, all four variants were only able to be produced as truncated variants of molecular weight of ~25 kDa as assessed via SDS-PAGE and showed no catalytic activity, similar to variants T1-4. While it is not clear which specific mutations likely led to internal unfolding and proteolytic cleavage in *E. coli*, we acknowledge that combining seven model-predicted mutations in these variants is likely to be destabilizing, so this final outcome is not too surprising. Enhanced variants might instead be design via combination of mutations from the characterized variants within the Designed and RM libraries.

4.4 Discussion

Our work here provides the largest mapping of the lysin sequence-function landscape to date while demonstrating a HT approach for continued functional screening of CD sequence space. Our depletion-based assay showed strong agreement between population enrichment and individual growth dynamics for CDs expressed with an N-terminal TorA signal peptide, exhibiting the consistency of this assay and supporting continued development of depletion-based assays in this space. Though, it is surprising that better results were not also observed via periplasmic expression with the OmpA signal peptide, given the extent of its prior use and its ability to differentiate growth inhibition at a clonal level. While the depletion assay with expression with the TorA signal peptide is limited in that it is unable to distinguish between growth inhibition via catalytic activity and other potential mechanisms, such as CD variants which unfold in the cell and disrupt

other cellular processes, we have demonstrated here that coupling this assay with yeast display protease stability information enables HT screening of lysin libraries, identified highly active and stable CD variants, and should enhance future lysin engineering efforts.

Across our array of CD diversification methods, all design strategies implemented here were able to yield compelling variants with activities and stabilities comparable to or greater than LysV7 as shown in our high throughput data. A small genome-mined library yielded identification of the GM4 variant, highlighting the value of this design and chimeric library designs in general. We hypothesize that repeated screening of a similarly designed genome-mined library, allowing for larger library diversity (~10^6 variants), would significantly expand lysin sequence space and yield promising variants. Further, we demonstrated a time- and cost-efficient method for constructing such a diverse library of protein chimeras from a gene oligopool.

Across the Designed, RM, and Triplet libraries, the EVCouplings-informed sub-library consistently performed well across all assays demonstrating that, for proteins with thousands of natural homologs, the EVCouplings framework is an informative approach for generating high performing libraries with combinatorial mutations. Yet, EVCouplings predictions did not significantly improve model performance and only weakly correlated with experimental data. Thus, EVCouplings aids in protein library design but does not effectively predict clonal performance, demonstrating the need for HT assays capable of screening relevant functions of interest. However, we do find it interesting that EVCouplings-predicted

fitnesses do not correlate with experimental stability data. Given potential compounding factors in the depletion assay, we would not expect inhibitory activity to be directly correlated with protein fitness; however, there are several examples in literature showing that experimental stability data and evolutionary fitness are often correlated[142], making the weak correlation observed here surprising. In evaluating other design strategies, we found that the Linear Triples sub-library helped identify high performing variants and achieved moderate overall function but clearly limited our ability to identify longer range epistatic interactions, as shown via PW model performance. It is also noteworthy that, while the PROSS sub-library was altogether not as performative as other design strategies in our experimental data, it offers the benefit of requiring significantly fewer natural homologs than the EVCouplings framework and still yielded moderate function.

Interestingly, we found that both SW and PW models achieve comparable predictive power across most of our data sets. PW models achieved higher predictive performance when trained on data with a higher rate of combinatorial mutations, as expected. The superior model performance seen within PW models trained on the Designed library could also be due, in part, to higher mutation rates at sites with epistatic interactions, due to the incorporation of EVCouplings-informed mutations within the Designed library. However, since EVCouplings predictions did not improve model performance and only weakly correlated with experimental data, it is more likely that increased PW model performance within this library was due to higher combinatorial diversity. Superior model performance when predicting stability compared to activity is also not surprising, as previous

results have demonstrated that protein folding and stability is a more predictable protein function than other functions of interest, such as enzymatic activity[48,193]. However, it is notable that we saw very poor model performance when predicting activity of our full data set with significantly improved performance when modeling the activity of sequences observed within both depletion and stability assays. Yet, inclusion of stability data to predict activity, or vice versa, did not significantly impact model performance, so it is not immediately obvious why restricting modeling to this smaller data set yields superior prediction. It is possible that CDs observed in yeast display have some inherit basal stability or fitness necessary to be displayed on the yeast surface, and that CDs which do not meet this basal limit are inconsistent in the depletion assay due to low stability. While literature suggests most proteins are expressed comparably on the surface of yeast, regardless of stability, there are examples of proteins with lower thermal stability levels where yeast display expression levels correlates with thermal stability[194]. Thus, it is possible that, given the low thermal stability of LysEFm5, observance of a CD in the yeast stability assay filters out highly unstable sequences that perform inconsistently in the depletion assay. Regardless, coupling stability and activity data yields comparable model prediction of stability and improved prediction of activity, demonstrating the benefit of using both assays.

Our work also drastically expanded previously explored lysin sequence space, as this is the largest set of lysin variants screened for activity and stability to our knowledge. Our screening of ~$5*10^4$ CDs identified multiple compelling variants, demonstrated a moderate CD tolerance for mutagenesis with regards to

both activity and stability, and identified interesting sequence-function relationships. Further, we observed a general tradeoff between activity and stability within our data. Yet, we did successfully identify variants with both enhanced activity and stability compared to LysV7. Moving forward, we expect lysin engineering efforts could expand to coupling strategies shown here to CDs capable of targeting particularly problematic Gram-negative pathogens. Further, we suspect applying similar diversification and stability screening strategies, coupled with affinity maturation of lysin CWBDs will enable engineering of highly targeted, enhanced stability lysins.

Antimicrobial resistance is an increasingly critical global health threat and lysins are a particularly compelling solution. However, lysin development and translation has been limited in part by lack of HT methods for screening antimicrobial activity and limited data sets for informing AMP engineering efforts. Our work here expands the screenable throughput of lysin libraries and provides a straightforward, efficient approach for protein diversification and testing to inform epistatic models and guide AMP design. We hypothesize that continued development of HT assays for screening antimicrobial activity, coupled with stability information, will enable better training of protein models and enhance development of translatable AMPs for therapeutic applications.

4.5 Methods

4.5.1 Bacterial and yeast cultures

E. coli cells were grown in lysogeny broth (LB; Fisher BioReagents) which contained 1.6% (vol/vol) agar in the case of solid-phase growth at 37°C, with

shaking of liquid cultures at 250 rpm. When specified, LB cultures were supplemented with 100 µg/mL ampicillin or 50 µg/mL kanamycin. *E. faecium* strain 8E9 (generously provided by Prof. Patricia Ferrieri of the University of Minnesota) was grown in liquid brain heart infusion (BHI, include brand supplier) medium at 37 °C with shaking at 250 rpm.

Saccharomyces cerevisiae strain EBY100 was used for all yeast experiments. Nonselective growth was done in liquid YPD (10 g/liter yeast extract, 20 g/liter Bacto peptone, and 20 g/liter D-glucose) or on solid YPD plates which contained 1.6% (vol/vol) agar. Selective growth was done in liquid SD-CAA (16.8 g/liter sodium citrate dihydrate, 3.9 g/liter citric acid, 20 g/liter D-glucose, 6.7 g/liter yeast nitrogen base, 5 g/liter Casamino Acids) or on solid SD-CAA plates which contained 1.6% (vol/vol) agar. Induction of yeast display constructs was done in SG-CAA (10.2 g/liter $Na_2HPO_4 \cdot 7H_2O$, 8.6 g/liter $NaH_2\ PO_4 \cdot H_2O$, 19 g/liter D-galactose, 1 g/liter D-glucose, 6.7 g/liter yeast nitrogen base, 5 g/liter Casamino Acids). All yeast growth was done at 30 °C and at 250 rpm for liquid culture.

4.5.2 E. coli Depletion Activity Assay

4.5.2.1 Preparation of LysEFm5 Control Variants

Template DNA for all control lysins was obtained from previous work (Supplemental Table S4.2) [40]. 10 ng of template DNA was amplified in a single 25 µL polymerase chain reaction (PCR) with primers designed to append the respective periplasm secretion signal peptide on the 5’ end of the genes of interest (see Supplemental Table S4.1 for primer sequences). 1.25 and 0.75 µL of the pET-OmpA-Fwd-1 and -2 primers, respectively, were used with 1.25 µL of the pET-LysV7-amp-R primer to append the OmpA signal peptide upstream of lysin

constructs (annealing temperature of 72 °C). 1.25, 0.75, and 0.5 µL of the pET-TorA-Fwd-1, -2, and -3 primers, respectively, were used with 1.25 µL of the pET-LysV7-amp-R primer to append the TorA signal peptide upstream of lysin constructs (annealing temperature of 65 °C). All primer stocks here and throughout the methods were kept at 10 µM in ultrapure water. All PCRs here and throughout the methods were 25 µL reactions run for 30 cycles which included 0.25 µL of Q5 DNA polymerase (New England Biolabs) and 0.5 µL of 10 mM deoxyribonucleotide triphosphate mix (dNTPs), unless otherwise noted. DNA encoding the OmpA and TorA signal peptides were identified from the iGEM Registry of Standard Biological Parts, parts BBa_K3114002 and BBa_K3114005, respectively. Final DNA inserts were isolated by gel electrophoresis.

4.5.2.2 Preparation of E. coli LysEFm5 Variant OmpA/TorA Expression Plasmids

An *E. coli* cell stock expressing the pETh production plasmid was obtained from previous work[40] (sequence provided in supplemental materials). *E. coli* was grown overnight in liquid LB supplemented with kanamycin, then DNA was extracted to obtain high concentration plasmid DNA. The plasmid was digested with NheI-HF and BamHI-HF restriction enzymes (New England Biolabs) according to manufacturer's protocols and isolated by gel electrophoresis. Individual DNA inserts for each control were assembled into the digested pETh plasmid via HiFi DNA Assembly (New England Biolabs), transformed into T7 Express *lysY/I^q* Competent *E. coli* (C3013I, New England Biolabs), plated on solid LB supplemented with kanamycin, and grown overnight. Individual colonies were grown up in liquid LB supplemented with kanamycin, sequence verified, and used to generate 30% glycerol stocks stored at -80 °C.

4.5.2.3 Individual Growth Depletion Assay Experiments

Cultures of *E. coli* expressing LysEFm5 variants with OmpA or TorA signal peptides in the pETh expression plasmid were inoculated from glycerol stocks into 5 mL of liquid LB supplemented with kanamycin and grown overnight. Cultures were diluted 1000-fold in fresh LB supplemented with kanamycin with varying levels of IPTG and 200 µL of culture dilutions were plated in triplicate in 96-well plates. 96-well plates were monitored in a microplate reader (Biotek Synergy H1), and the optical density at 600 nm (OD_{600}) was measured at 5-minute intervals for up to 18 hours.

4.5.2.4 Population Growth Depletion Assay with Known Controls

Cultures of *E. coli* expressing LysEFm5 variants with OmpA or TorA signal peptides in the pETh expression plasmid were inoculated from glycerol stocks into 5 mL of liquid LB supplemented with kanamycin and grown overnight. Cells from each of the controls were added in equal densities in 5 mL of LB supplemented with kanamycin to a final OD_{600} of 0.1. Cultures were grown to an OD600 of 0.5-0.7, diluted 10-fold in 5 mL of fresh LB supplemented with kanamycin and 0, 0.1, or 0.5 mM IPTG, and grown for 8 hours. 100 µL samples were taken from cultures at 0-, 1-, 2-, 4-, and 8-hour time points. DNA was extracted and stored at -20 °C until preparation for Illumina sequencing.

4.5.2.5 Preparation of E. coli Library Expression Plasmids with N-terminal TorA Signal Peptides

Following construction of each library as detailed below, a single 25 µL PCR was conducted to append the TorA signal peptide upstream of the library genes encoding CD variants as described previously. 10-50 ng of library DNA in 1µL of

ultrapure water was used as template, which supplied >1000-fold coverage of each library diversity. Final DNA inserts were isolated by gel electrophoresis.

pETh plasmid was extracted from overnight culture of *E. coli* stock and digested with NheI-HF and BamHI-HF restriction enzymes according to manufacturer's protocols. Isolated library genes were assembled into digested pETh vector via HiFi DNA Assembly. A 20 µL HiFi reaction consisted of 150 ng of digested vector, 50-75 ng of library insert, and 10 µL of 2X HiFi reaction mix. The reaction was incubated at 50 °C for one hour, then purified via MinElute PCR Purification Kit (Qiagen) and eluted in 10 µL of ultrapure water. 5 µL of purified reaction product was transformed into 10-beta Electrocompetent *E. coli* (New England Biolabs) according to manufacturer's protocols. Following the outgrowth step in 1 mL of outgrowth medium at 37 °C, the outgrowth culture was 2-fold serially diluted in 5 mL liquid LB supplemented with kanamycin and grown overnight to allow collection of a subset of transformants, as previous literature shows ~10- to 100-fold sequencing coverage of libraries is needed for depletion assay consistency[31]. A small fraction of the outgrowth culture was also 10-fold serial diluted and plated on LB-agar plates supplemented with kanamycin to estimated total number of transformants. Transformations consistently yielded 1-5 million transformants. After colony counts, a subset of ~$2.5*10^5$ transformants was collected from overnight liquid cultures and library DNA was extracted.

4.5.2.6 Library Depletion Screening

pETh plasmids containing each individual library were transformed into T7 Express *lysY/I^q* Competent *E. coli* (New England Biolabs), according to manufacturer's protocols. 300 ng of plasmid DNA was transformed into 50 µL of competent cells

for each library per transformation. The Designed and RM libraries were transformed 4X for each replicate, the Triplet library was transformed 3X for each replicate, and the GM library was transformed 1X for each replicate, based on each library's theoretical diversity. Following transformation, cultures were outgrown in 1 mL of super optimal broth (SOC) media at 37 °C and 250 rpm for 1 hour. Outgrowths for all transformations for a given library were then mixed, and a fraction of the final mixture was 10-fold serially diluted and plated to estimate number of transformants. The remaining outgrowth mixture was centrifuged for 2 minutes at $2000 \times$g, the media was removed, the cell pellet was resuspended in 5 mL of LB supplemented with kanamycin, and the culture was grown for 1 hour. 500 µL of culture was then added to 4.5 mL of fresh LB supplemented with kanamycin and a final concentration of 0.5 mM IPTG and grown for 8 hours at 37 °C and 250 rpm. After growth, the culture was centrifuged and DNA was extracted and stored at -20 °C until preparation for Illumina sequencing. The GM library was screened in duplicate, all other libraries were screened in triplicate, and all replicates were conducting on different days.

4.5.3 On-yeast Protease Stability Assay

4.5.3.1 Preparation of Yeast Surface Display pCT Plasmid

We originally intended to engineer an assay that used yeast surface display of lysins to functionally screen their catalytic activity. Prior literature has demonstrated that many lysins which target Gram-positive bacteria require a free N-termini for enzymatic function[195], so we constructed a pCT expression vector for display on the N-termini of Aga2. While our efforts to engineer a yeast display-based assay for quantifying enzymatic activity did not yield consistent results, we

used our N-terminal display pCT expression vector for our on-yeast protease stability assay. Details of plasmid construction are detailed below.

A pCT expression vector for display of proteins off the C-termini of Aga2 was obtained from previous work[30]. The pCT plasmid was digested using EcoRI-HF and XhoI restriction enzymes (New England Biolabs) and the larger digestion product was isolated by gel electrophoresis. This digestion step removed all Aga2 display machinery from the pCT plasmid and only the backbone was retained.

The individual components for reconstructing a pCT expression vector for display on the N-termini of Aga2 were amplified using primers described in Supplemental Table 1 and the original pCT vector as template: the Aga2SP-Amp-F and Aga2SP-Gene-Amp-R primers were used to amplify the Aga2 signal peptide; the Gene-Amp-F and Gene-Cmyc-Amp-R primers were used to amplify the gene for a Gp2 scaffold variant; and the Aga2-Amp-F and Aga2-pCT-Amp-R primers amplified the Aga2 protein. The Cmyc-G4S-Aga2-Insert is an oligonucleotide that encoded the cMyc protein tag and a $(Gly_4\text{-}Ser)_3$ linker. All primers were designed to have sufficient overlap to allow for HiFi DNA Assembly. PCR products were isolated by gel electrophoresis and then all individual components were HiFi assembled to yield the final expression construct in pCT: Aga2 signal peptide-Gp2 scaffold-cMyc-$(G_4S)_3$-Aga2. The Gp2 scaffold acts as a placeholder in this construct and can be removed via digestion with NheI and BamHI restriction enzymes to allow for insertion of other proteins. This HiFi product was transformed into 10-beta Competent *E. coli* (New England Biolabs), plated on solid LB supplemented with ampicillin, and grown overnight. Individual colonies

were grown up in liquid LB supplemented with ampicillin, sequence verified, and used to generate a 30% glycerol stocks stored at -80 °C. Proper expression of parental LysV7 construct in this plasmid was confirmed via flow cytometry. This construct is the pCT vector used in all following protocols. The full sequence of this pCT plasmid is provided in the supplemental materials.

4.5.3.2 Library Assembly into pCT Plasmid

Following construction of each library as detailed below, a single 25 µL PCR was conducted to append an HA tag upstream of the library genes encoding CD variants as well as overlap with the pCT vector at specified restriction enzyme sites using pCT-LysV7-Insert-F and -R primers. 10-50 ng of library DNA in 1µL of ultrapure water was used as template, which supplied >1000-fold coverage of each library diversity. Final DNA inserts were isolated by gel electrophoresis. Library genes were then amplified via PCR with Phusion Polymerase (New England Biolabs) according to manufacturer's protocol using pCT-LysV7-Amp-F and -R primers in a 100 µL 35-cycle reaction with 5 µL of isolated products acting as template. 5 µL of Phusion PCR products were run on gel electrophoresis to confirm product formation and the remaining product was concentrated via ethanol precipitation: 2 µL Pellet Paint Co-Precipitant (Millipore), 10 µL 3 M sodium acetate at pH 5.2, and 400 µL ethanol was added to the post-PCR product and the mixture was then incubated at 4 °C for 10 min. The insoluble DNA was pelleted via centrifugation at 15,000 × g for 20 min at 4 °C. The DNA was then washed with 500 µL 70% ethanol in dH_2O, centrifuged, washed with 500 µL ethanol, and centrifuged. The liquid was discarded and the pellet was dried overnight open to

room temperature air. The precipitate was then resuspended in 15 µL ultrapure water.

pCT plasmid for N-terminal expression on Aga2 was extracted from overnight culture of *E. coli* stock, digested with NheI-HF and BamHI-HF restriction enzymes according to manufacturer's protocols, and concentrated via ethanol precipitation. The concentrated lysin gene libraries were inserted into the pCT yeast display vector (described above) via homologous recombination into *S. cerevisiae* yeast (EBY100) as described previously[196]. Dilutions of transformed libraries were plated on solid SD-CAA plates to quantify number of transformants: EVCouplings-informed, PROSS-informed, Triplets, error-prone PCR (20 cycles) and error-prone PCR (25 cycles) all yielded 1.4-3.7*10^7 transformants while the GM library yielded 4*10^6 transformants. For on-yeast protease stability sorting, the EVCouplings- and PROSS-informed libraries were merged at equal densities to yield the Designed library and the error-prone PCR libraries (20 and 25 cycles) were merged at equal densities to yield the RM library.

4.5.3.3 On-yeast Protease Stability Screening

Yeast cells transformed with pCT plasmids containing the lysin libraries were inoculated and grown to an OD_{600} of ~1.0 in SD-CAA. Cells were centrifuged at 6000 × g for 2 minutes, resuspended in SG-CAA and induced overnight at 20 °C and 250 rpm. Following induction, for a single sample, approximately 10^7 yeast cells were centrifuged at 2000 × g for 2 minutes, washed twice in PBSACM (phosphate-buffered saline with 0.1% (w/v) bovine serum albumin, 1 mM $CaCl_2$ and 0.5 mM Mg_2SO_4), and resuspended in 200 µL PBSACM and incubated at 37 °C for 10 minutes. Cell densities were estimated by OD_{600} measurements of a

dilution of the yeast population. To account for library diversities, the number of sorted yeast cells differed across libraries. To allow for this, for a given protease replicate, three samples were prepared for the Designed and RM libraries, two samples were prepared for the Triplet library, and one sample was prepared for the GM library. Samples were merged after labeling prior to FACS.

Protease concentrations were tested and selected to provide a range of protein cleavage for each library. Libraries were treated with trypsin at 50 nM, chymotrypsin at 1 µM, and proteinase K at 0.5 U/L. Proteases were diluted to 2X these concentrations in PBSACM and incubated at 37 °C for 10 minutes. Then 200 µL of protease dilutions were added to each yeast sample, mixed briefly via pipetting, and incubated at 37 °C for 10 minutes. After incubation, 600 µL of ice-cold PBSACM was added to stop protease activity and samples were centrifuged. All following steps occurred on ice or at 4 °C in centrifuges.

Cells were washed twice in PBSACM, labeled in 200 µL PBSACM with 5 µg/mL anti-c-myc (clone 9E10, catalog number 626802; BioLegend) and 5 µg/mL anti-HA (chicken anti-HA, catalog number ab9111; Abcam) antibodies for 1 hour. Cells were then washed twice in 500 µL PBSACM and labeled with 4 µg/mL goat anti-mouse Alexa Fluor 647 (catalog number A-21235; Thermo Fisher Scientific) and 4 µg/mL goat anti-chicken Alexa Fluor 488 (catalog number A-11039; Thermo Fisher Scientific) for 30 minutes. Yeast cells were washed twice in 500 µL PBSACM then resuspended in 750 µL PBSACM for FACS sorting using a FACSAria II instrument (Becton, Dickinson Bioscience) at the University of Minnesota Flow Cytometry Resource facilities.

cMyc positive cells were sorted into four sorting gates based upon N-terminal HA to C-terminal cMyc ratio: top 5% of N-terminal:C-terminal ratio (score: 0.975, gate P5), next 10% (0.9, gate P6), all remaining N-terminal-positive cells (0.425, gate P7), and N-terminal-negative cells (0, gate P8) (Supplemental Figure 4). Number of cells collected in each gate per replicate per library are shown in Supplemental Table 3. Following sorting, cells were centrifuged, buffer was removed, and DNA was extracted via Zymoclean Gel DNA Recovery Kit (Zymo Research) following the manufacturer's protocol. Following DNA elution into 30 µL of ultrapure water, 15 µL of the DNA elution was mixed with 2 µL ExoI (New England Biolabs), 1 µL of Lambda Exonuclease (New England Biolabs) and 2 µL of 10X Lambda Exonuclease Buffer, incubated at 30 °C for 90 min to remove genomic DNA, and 80 °C for 20 min to inactivate the enzymes. The DNA was then stored at -20 °C until preparation for Illumina sequencing.

4.5.4 Illumina Sequencing and Read Analysis

4.5.4.1 DNA preparation for Illumina Sequencing

Following DNA extraction from depletion assays, Illumina adapters were appended to samples in two sequential PCRs using primers specified in Supplemental Table 11: LysEFm5-Illumina-p13-Fwd-N1-3 and OmpA/TorA Illumina Rev N1-3 primers were used for amplifying genes from proof-of-concept population experiments with LysEFm5 variants; Oligopool Illumina Fwd and Rev N1-3 were used for amplifying GM library samples; LysEFm5-Illumina-p13-Fwd-N1-3 and LysEFm5-Illumina-p157-Rev-N1-3 primers were used for amplifying Designed and RM library samples; and LysEFm5-Illumina-p5-Fwd-N1-3 and LysEFm5-Illumina-p145-Rev-N1-3 or LysEFm5-Illumina-p40-Fwd-N1-3 and LysEFm5-Illumina-p185-Rev-N1-3

primers were used for amplifying Triplet samples. Triplet samples had potential DNA diversity between positions 5 and 180, so mutations could be >500 nucleotides away from each other, preventing sequencing on one reading on the Illumina NovaSeq platform. To circumvent this, we amplified Triplet DNA samples in two regions (from positions 5-145 and 40-185) to enable sequencing on the NovaSeq with our other libraries. Following initial amplification, correct DNA products were isolated by gel electrophoresis and used as template in the second PCR which used Ni5N501-N508 and Ni7N701-N712 primers to append Illumina adapters. Correct products were isolated via gel electrophoresis and mixed for sequencing as described below.

All Illumina sequencing was conducted by the University of Minnesota Genomics Center. Proof-of-concept population experiments of LysEFm5 controls were sequenced on on an Illumina iSeq 100, yielding X reads. All samples were equally mixed to provide equal coverage. For all library sequencing, samples were mixed to provide greater coverage of depletion experiment samples, as literature has shown >10-fold read coverage is necessary for accurate depletion experiments[31]. GM samples were sequenced using version 3 chemistry for 2X300 paired-end sequencing on an Illumina MiSeq, yielding 10.9M filtered reads. DNA mixing was done to provide ~$2.5*10^5$ reads for each sorted yeast gate sample and ~$6.67*10^5$ reads for each depletion experiment sample. Designed, RM, and Triplet library samples were sequenced using 2X250 paired-end sequencing across two SP flow cells, yielding 411M filtered reads. DNA mixing was done to provide 10^6 reads for each sorted yeast gate sample and ~10^7 reads for each

depletion assay sample. Given this read allocation, the two higher performing FACS gates (P5 and P6) were sequenced at a greater depth (more reads per cells sorted) than the bottom two gates based on the assumption that better performing sequences would be more informative to sequence modeling than lower performing sequences.

Sequences were initially processed using computational resources of the Minnesota Supercomputing Institute. We used USearch v11[114] to merge, align, filter, and dereplicate all sequences. Merged reads were clipped to the region between amplifying primers and filtering was done to remove all sequences with >1 expected error. Sequences were then analyzed by assay and sample type to quantify sequence scores.

4.5.4.2 Converting Illumina Reads to Activity Scores

Sequences observed in depletion assays were only retained if they were observed in the initial pool. Sequences only observed in the induced pool were removed. If a sequence was observed only in the initial pool, it was assigned a read count of 1 in the induced pool to allow use of log2 analysis. The sum of total reads in the induced pool for calculating f_i below included these “pseudo-counts”. Sequence scores were then quantified as the log2 enrichment from the initial pool to the induced pool as shown below:

$$score_{activity} = \log_2 \frac{f_{induced}}{f_{initial}} \quad (4.1)$$

where f_i is the fraction of a sequence in population i. These activity scores were calculated for sequences within each experimental replicate. Sequences observed across all replicates were then identified and the mean and standard deviation of

activity scores was calculated. To increase rigor, we applied an additional filtering step, requiring sequences to be depleted in all replicates (activity score < 0), enriched in all replicates (activity score > 0), or have a standard deviation < 0.5 (to evaluate variants with activity scores near 0).

4.5.4.3 Converting Illumina Reads to Stability Scores

Our FACS collection gates were drawn to bin cells based on proteolytic stability. We defined a stability score which correlates to the relative position of a sequence across these gates, as described previously[30]. For each population, the read frequency of every sequence was converted to a theoretical number of cells collected via FACS, then the assay score for a sequence was calculated as the cell-averaged scored across gates using gate scores (0.975, 0.9, 0.425, 0) defined above. We report individual protease assay scores (trypsin, chymotrypsin, and proteinase K) and a stability score which is the average of trypsin, chymotrypsin, and proteinase K stability scores. Sequences had to be observed in at least one gate in a replicate to be included in analysis.

4.5.4.4 Merging sequences across assays

Sequences were compared across all assays to identify those which had both activity and stability scores. During processing for the GM library, DNA sequences were translated to amino acid sequences immediately after scoring within a specific assay. Then sequences were compared to identify common sequences present in multiple assays. This allowed potential amino acid sequences containing silent mutations to be identified as present in multiple assays, most notably GM4. However, for processing of the Designed, RM, and Triplet libraries, to avoid this, DNA sequences were not translated until after comparison across all

assays. This allowed replicate amino acid sequences (containing silent mutations) to be observed multiple times. This distinction is important, as we expect and observed that different DNA sequences could yield very different results, particularly in the *E. coli* depletion assay, as codon frequency impacts the producibility of a protein sequence.

4.5.5 Production and Testing of Individual Lysins

4.5.5.1 Construction of Lysin Expression Plasmids

Genes expressing lysin CDs or CD-CWBD fusions were constructed via PCR or ordered as G-blocks (Integrated DNA Technologies), amplified to append plasmid overlap, and inserted into pETh via HiFi DNA Assembly (New England Biolabs).

Variants GM1-5 were ordered as G-blocks (Supplemental Table 9), amplified with primers GMLibVars-Fwd and GMLibVars-CD-Rev or GMLibVars-CD-CW-Rev, and PCR products isolated via gel electrophoresis. The LysEFm5 CWBD gene was amplified with primers GMLibVars-CWBD-Fwd and -Rev and isolated via gel electrophoresis. GMLibVars-CD-Rev amplified the gene with overlap for insertion into pETh as a CD and GMLibVars-CD-CW-Rev amplified the gene with overlap with the CWBD gene for insertion in pETh as a CD-CWBD fusion.

All variants identified in Designed, RM, and Triplets library experimental were constructed via PCR insertion of mutations as CD-CWBD fusions using the LysV7 CD-CWBD gene as template. Correct DNA products were isolated via gel electrophoresis. Model-informed variants DP1-2 and RP1-2 were ordered as G-blocks encoding all mutations in the LysV7 sequence for HiFi DNA Assembly with the CWBD into pETh. G-blocks were amplified with primers LysV7-pET-Fwd and

Model-Pred-Rev and isolated via gel electrophoresis. The unmodified remainder of the CD and the CWBD were amplified with Model-Pred-Fwd and LysV7-CWBD-pET-Rev primers using the LysV7 CD-CWBD gene as template and isolated via gel electrophoresis. Genes were assembled into pETh via HiFi DNA Assembly, transformed into T7 Express Competent E. coli (New England Biolabs), and sequence-confirmed prior to individual production and testing as detailed below.

4.5.5.2 Production and Purification of Lysins

Lysin variants were produced as described previously. For each clone, a cell culture tube containing 3 mL of LB supplemented with kanamycin was inoculated with cells from a glycerol stock and incubated at 37 °C and 250 rpm overnight. After overnight growth, 100 mL of fresh LB was inoculated with 100 µL of confluent culture. When the OD_{600} was within the range of ~0.6-0.8, IPTG was added at a final concentration of 0.5 mM and the culture was left to incubate at 20 °C and 250 rpm for overnight. 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 Na_2HPO_4, 2 mM PBS, 5% glycerol, 3.1 g/liter 3-[(3-cholamidopropyl)-dimethylammonio]-1-propanesulfonate [CHAPS], 1.7 g/liter imidazole, with a Pierce Protease Inhibitor Mini Tablet, EDTA free [1 tablet per 10 ml buffer]) was added. Each culture was supplemented with $MgSO_4$ to a final concentration of 20 mM, 2U 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 and room temperature, respectively. Cell material was then centrifuged at 12000 × g for 10 minutes, and the supernatant was filtered, 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 and washed three times with wash buffer. Three elutions were then performed with 400 µL of elution buffer (with 50 mM sodium phosphate, 300 mM NaCl, 150 mM imidazole, 5% glycerol) to constitute the protein sample in a volume of ~1.2 mL. A small fraction of this elution was used for SDS-PAGE analysis of purified protein products. The remaining protein elutions were desalted via Zeba Spin Desalting Columns (ThermoFisher Scientific, cat. # 89889), eluted into PBS with 5% glycerol, divided into 100 µL aliquots, and snap-frozen until needed. All subsequent analysis was performed on aliquots thawed on ice immediately before use. Protein concentrations were quantified via NanoDrop One C (ThermoFisher Scientific).

4.5.5.4 Quantification of Lysin Enzymatic Activity via Turbidity Reduction Assay

E. faecium cell material was produced as described previously[40]. One hundred mL of BHI broth was inoculated with *E. faecium* strain 8E9 and grown overnight at 250 rpm and 37°C. The culture was autoclaved and centrifuged, the liquid was discarded, and the *E. faecium* pellet was washed three times with phosphate-buffered saline (PBS). The pellet was then resuspended in PBS at a concentration of ~0.3 g/mL and the stock was stored at 4°C until future use.

In a 96-well plate, specified amounts of purified lysin CDs or CD-CWBD fusions were combined with *E. faecium* cell material at an OD*600* ~ 1.0 in a total volume of 200 µL PBS. OD_{600} was measured every 2 minutes for up to 4 hours in a plate reader (Biotek Synergy H1) and catalytic activity was quantified as the highest rate of change in OD_{600} over any 6-minute interval.

4.5.5.5 Quantification of Lysin Melting Temperature via Thermal Denaturation Assay

Forty-five µL of lysin variants were aliquoted into optically clear PCR tubes. The stock solution of Sypro orange (Thermo Fisher) was diluted to 200X in PBS and 5 µL 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- to 490-nm excitation and 560- to 590-nm emission (FRET measurement setting on CFX PCR system). The melting temperature was defined as the temperature which yielded the maximum change of fluorescence with temperature via smoothing with local second-degree polynomials having widths of 2.5°C using the Savitzsky-Golay filter of the sklearn package in Python.

4.5.6 CD Library Construction

4.5.6.1 Genome-mined Library Construction

A blastp search was conducted on the BLAST server[197,198] using the LysV7 CD amino acid sequence as the query sequence. 100 lysin CD homologs were identified with sequence similarity to LysV7 ranging from 55-99%. All homologs were aligned with LysV7 and a relatively highly conserved region from positions 87-92 (sequence IGYEVC in LysV7) was identified. Oligonucleotides encoding positions 1-92 and 87-185 of each homolog (when aligned with the LysV7 sequence) and LysV7 were designed to allow for recombination via PCR and synthesized via Twist Biosciences. Each set of oligonucleotides encoding CD fragments was PCR amplified individually using Oligopool 1 or Oligopool 2 Fwd

and Rev primers (Supplemental Table S4.1), then the two products were PCR amplified together with Oligopool 1 Fwd and Oligopool 2 Rev to give a diverse array of full-length CDs. Oligopool 1 Rev 1 and 2 and Oligopool 2 Fwd 1 and 2 primers, respectively, were mixed equally prior to initial oligopool amplification to allow for diversity at the overlap site. Amplified DNA fragments were isolated via gel electrophoresis and used as template for construction into pETh or pCT as described above. To include homologs that were shorter than the LysV7 sequence, we allowed a broad tolerance of final DNA gene lengths. Unfortunately, this ultimately resulted in high presence of truncated CD sequences (of less than 100 amino acids in length) that significantly limited the depth to which we sampled the GM library in the following work. For all sequence analysis and assay results for the genome-mined library, we incorporated a minimum DNA length filter of 400 base pairs to identify “true” CD sequences.

4.5.6.2 Designed library construction

Towards applying new protein engineering strategies with higher functional hit rates, two LysEFm5-V7 libraries were generated using the EVCouplings and PROSS online servers (EV and PR libraries, respectively)[181,192]. For each library, a genomic database search was conducted on LysEFm5-V7 on the servers. The EVCouplings server ran a search on JackHMMER that yielded 2.8 x 10^5 homologs with an E-value $\leq 10^{-5}$, while the PROSS server, which uses a much smaller multiple sequence alignment for training, ran a BLAST search that yielded 346 homologs. Output from the EVCouplings server included a model used to predict fitness of all single and double mutants of LysEFm5-V7, while output from the PROSS server included a list of combinatorial mutations predicted to improve

LysEFm5-V7 stability. While double mutants of LysEFm5-V7 were predicted using the EVCouplings model, all top performing double mutants were combinations of top performing single mutants, so EVCoupling-informed substitutions were chosen based on predicted single-substitution mutant fitness. From these EVCouplings- and PROSS-predicted mutations, the top ~25 beneficial amino acid substitutions from each model/server were identified and combinatorially incorporated into the respective library designs to achieve a combined library diversity of ~1.2 x 10^7 (EV~1.18x10^7, PR~6x10^5, Supplemental Tables S4.4-5). Consecutive 60- and 90-mer oligonucleotides encoding all desired substitutions were synthesized for each of the libraries to construct via iterative PCRs (Supplemental Tables S4.6-7). DNA encoding LysV7 was used as template for these iterative PCRs and final PCR products were isolated via gel electrophoresis. These yielded genes for the EV and PR libraries, which together comprise the Designed library. EV and PR library genes were individually amplified for insertion into pETh for the depletion assay and amplified genes were mixed equally prior to HiFi DNA assembly to yield the Designed library used in *E. coli* experiments. EV and PR library genes were individually amplified and inserted into pCT via homologous recombination, as described above, then mixed at equal cell densities to yield the Designed library used in yeast experiments.

4.5.6.3 RM Library Construction

The RM library was constructed via error-prone PCR, directed between positions 17 and 150 of pETh-LysV7 template DNA with LysV7-epPCR-Fwd and -Rev primers, as previously described for a moderate level of mutation[36,199]. Two 50 uL PCRs were conducted with final concentrations of 1 μM of nucleoside analogs

8-oxo-dGTP and dPTP, each, using 0.25 µL of Taq DNA Polymerase (New England Biolabs). PCRs were run for 20 and 25 cycles to allow for differing degrees of mutagenesis. Gene fragments from positions 1-17 and 150-185 were amplified with pET-LysV7-Amp-F and LysV7-epPCR-Revtobeg primers and pET-LysV7-Amp-R and LysV7-epPCR-Fwdtoend primers, respectively. Final library gene constructs were then amplified together with pET-LysV7-Amp-F and -R to generate library gene products, yielding error-prone library (20 cycles, ep-20 library) and error-prone library (25 cycles, ep-25 library). Ep-20 and ep-25 library genes were individually amplified for insertion into pETh for the depletion assay and amplified genes were mixed equally prior to HiFi DNA assembly to yield the RM library used in *E. coli* experiments. Ep-20 and ep-25 library genes were individually amplified and inserted into pCT via homologous recombination, as described above, then mixed at equal cell densities to yield the RM library for yeast experiments.

4.5.6.4 Triplet Library Construction

The Triplet library was designed to allow for local combinatorial diversity at all *i*, *i+1*, *i+2* positions from position 5 to 180 in LysV7. To do this efficiently, the Triplet library was constructed via PCR with a series of 29 “tiles” of oligonucleotides, each of which mutated 6 sets of *i*, *i+1*, *i+2* positions. Each tile *i* consisted of 6 Tile *i*-Fwd primers, which were equally mixed for a PCR, and 1 Tile *i*-Rev primer (Supplemental Table S4.8). Tile *i*-Fwd primers were used with pET-LysV7-amp-R to insert desired diversity while amplifying the LysV7 gene from tile *i* to the end of the CD and Tile *i*-Rev primers were used with pET-LysV7-amp-F to amplify the beginning of the LysV7 gene until the start of tile *i*. These products were isolated via gel electrophoresis and then PCR amplified with pET-LysV7-amp-F and -R to

yield the final tile *i* product. Following construction of each individual tile, genes were merged into the final Triplet library. For sequencing this library, the diversified DNA length ((180-5) X 3 = 525 base pairs) exceeded Illumina NovaSeq read length (500 base pairs), we chose to sequence two reading frames of the Triplet library, ranging from positions 5-140 and positions 45-180. Sequence scores were calculated based on frequencies within a reading frame, then all unique sequences within each reading frame were merged and scores for sequences seen in both reading frames were averaged. Sequences which appeared as the parental LysV7 sequence were removed from the population, as they could have had diversity outside of the reading frame.

4.5.7 Ridge Regression Modeling

4.5.7.1 Ridge Regression Modeling of Libraries and Sub-libraries

Regression modeling was done to identify CD sequence-function relationships and inform design of improved CDs. CD amino acid sequences were compared to the parental LysV7 protein sequence and transformed to one-hot-encoded matrices specifying (1) all individual substitutions from LysV7 or (2) all pairs of individual substitutions from LysV7. SW models were trained with the individual substitution matrix while PW models were trained with both the individual and pairs of substitutions matrix. Modeling was done using the LinearRegression, ElasticNet, Lasso, and Ridge models in the sklearn package in Python (https://scikit-learn.org/stable). However, initial modeling attempts used linear, lasso, and elastic net regression models led to overfitting, so all models discussed here are ridge regression models. A range of alpha values were initially tested for ridge regression models trained on activity and stability data sets, but we consistently

found that alpha = 1 provided highest model accuracy as measured by Pearson's correlation and coefficient of determination (R^2), so all models discussed here used alpha = 1.

For all modeling, experimental scores for each library were normalized from [0, 1], with higher scores for higher performing variants. For activity scores, this yielded model scores closer to 1 correlating to negative log2 enrichment values and model scores closer to 0 correlating to log2 enrichment values closer to or greater than 0. Model performance was assessed via 10-fold cross validation: the set of sequences within each library was randomly divided into 10 folds of equal sizes using the KFold function in the sklearn package in Python. For each fold, model performance was assessed via a model ridge regressed on other nine folds. The mean and standard deviation of Pearson's correlation, R^2 and MSE values across all ten folds are reported. For models trained on the RM library depletion data, 20000 random sequences were selected from the library for training, rather than the full 38514 sequences, due to increasing model parameter space and to allow comparison to the Designed library training size.

4.5.7.2 Sub-sampling of Designed and RM Libraries

To assess model performance when trained on fewer sequences, the Designed and RM library stability data sets were sub-sampled and ridge regression was performed. The sample_without_replacement function in the sklearn package in Python was used to randomly sample 100, 500, 1000, or 5000 observances within each library data set. Ridge regression with 10-fold cross validation was performed as described above and average performance metrics were reported.

4.5.7.3 Predicting Novel Variants with PW Models

We generated CD variants containing up to seven combinatorial mutations as predicted by our models to test their ability to generate previously unseen, high-performing CD variants. For both the Designed and RM libraries, two PW models were trained on the set of sequences observed within both stability and activity data sets. Models were separately trained on CD experimental activity and stability data. We analyzed the model coefficients from the PW stability model to identify the single substitutions predicted to be most beneficial. We generated all pairs of predicted beneficial single mutations and assessed each new variant, containing two mutations, for stability as predicted by the model. A score threshold was assigned to identify the top ~100-200 2-mutation variants, which were then combined with the predicted beneficial single mutations to generate variants containing three simultaneous mutations and stability scores calculated. This process was iterated to generate sets of variants containing 3-7 simultaneous mutations (Supplemental Figures S4.24-27). Variant activity scores were predicted by the PW model trained on activity data. The individual variants containing seven mutations with the highest predicted activity score or the highest stability score from each library were synthesized as G-blocks and individual produced and tested, as described above.

4.5.8 Statistical analysis

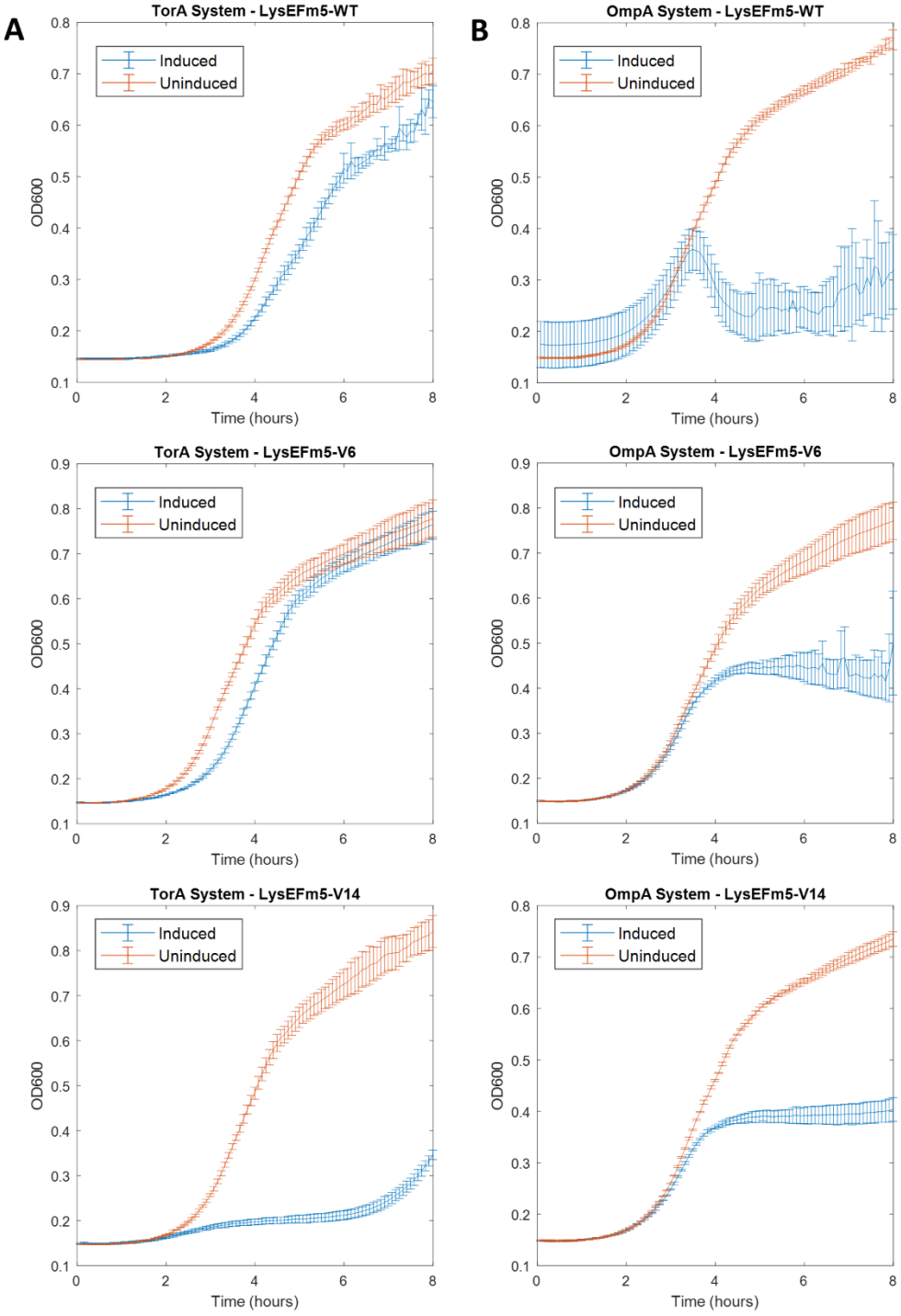
All statistical comparisons were done using functions in the sklearn package in Python. All comparisons between model performance were assessed via a two-sample t-test from model statistics using the ttest_ind_from_stats function.

Comparisons made between sub-library performances were assessed via a Mann-Whitney U test on populations using the mannwhitneyu function.

4.6 Acknowledgements

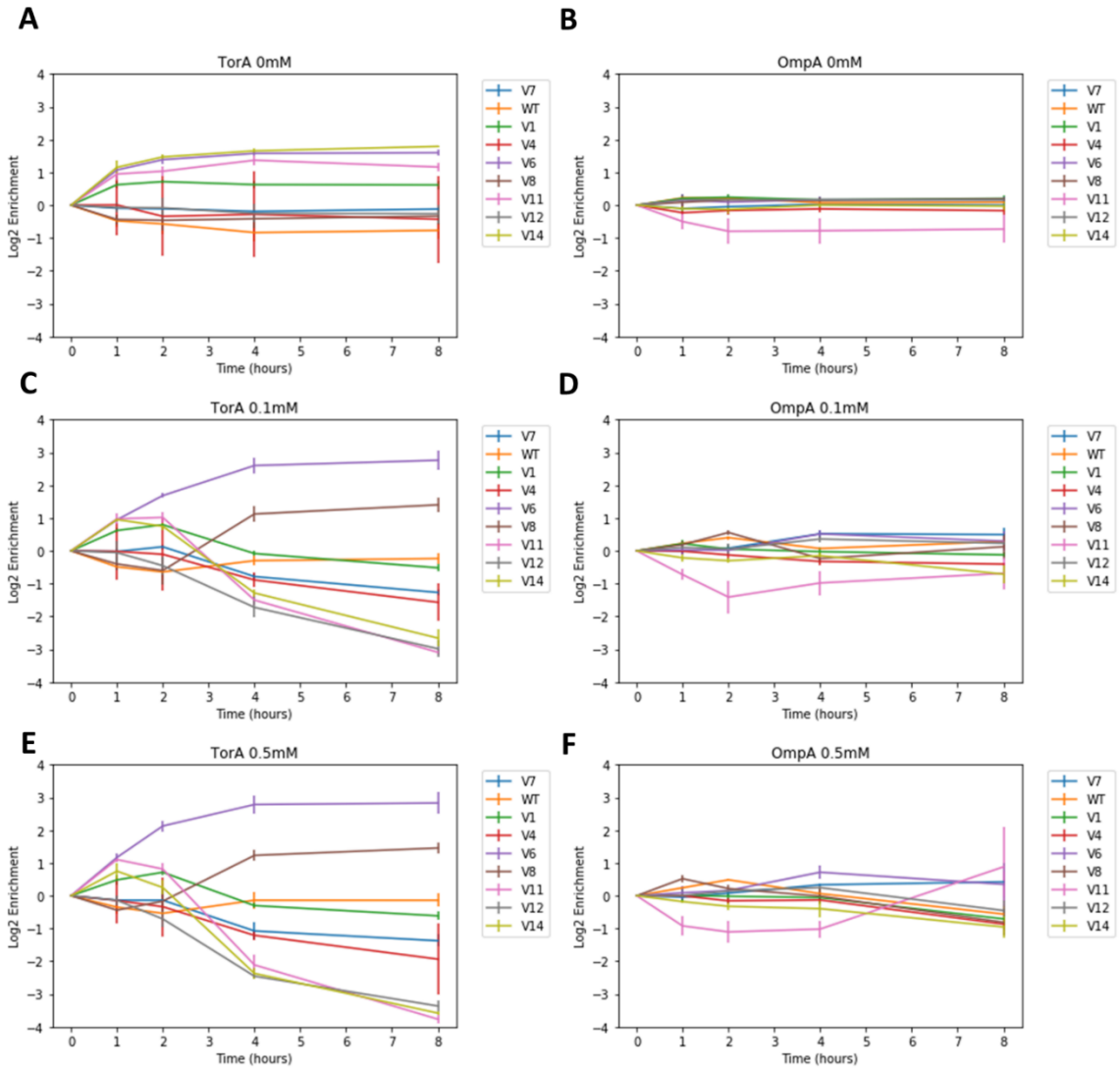
This work was funded by the National Institutes of Health (R01 GM121777). We appreciate assistance from the University of Minnesota Flow Cytometry Core, University of Minnesota Genomics Center, and the Minnesota Supercomputing Institute (MSI) at the University of Minnesota. We thank Patricia Ferrieri for donation of the *E. faecium* strain used in this study.

4.7 Supplement



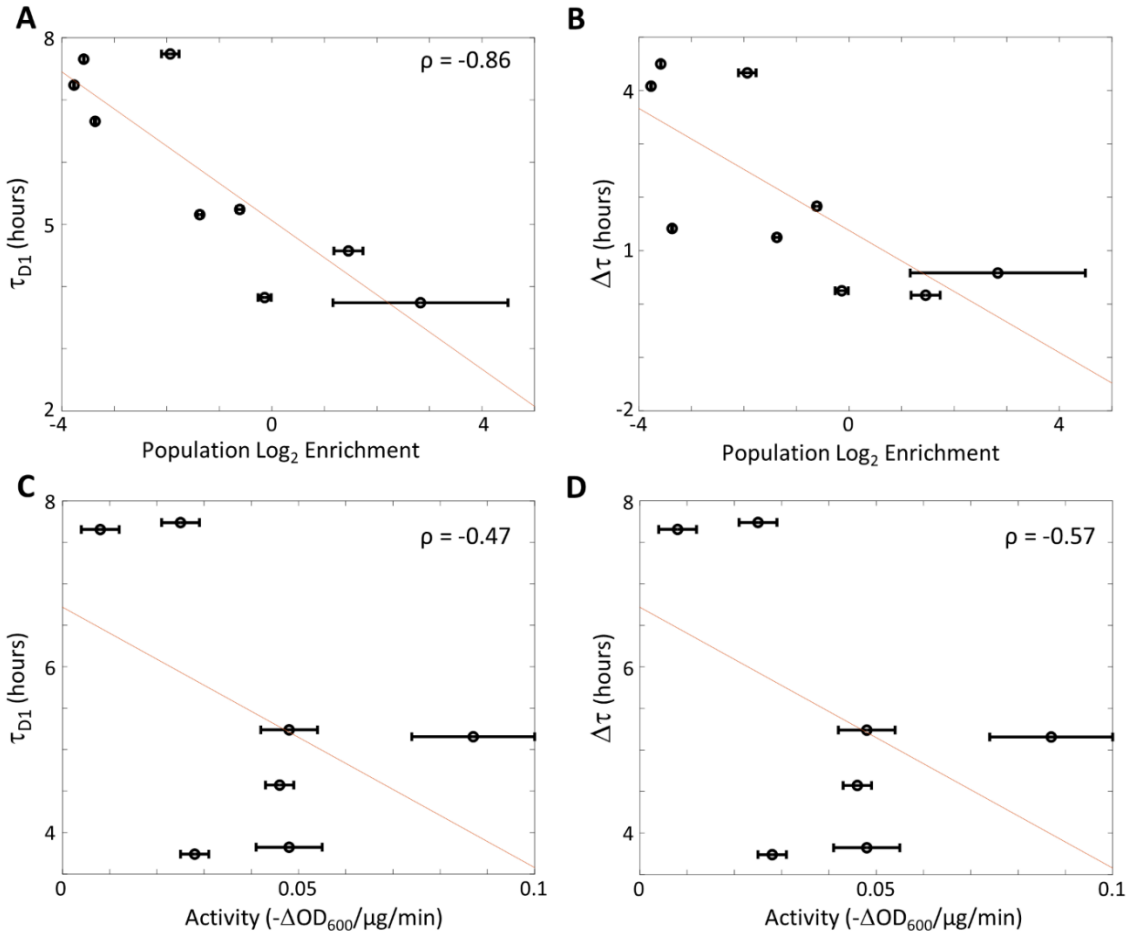
Supplemental Figure S4.1 Growth curves of additional controls with N-terminal TorA (A) and OmpA (B) signal peptides.

Controls are previously tested variants of LysEFm5. Mutations are shown in Supplemental Table 2.



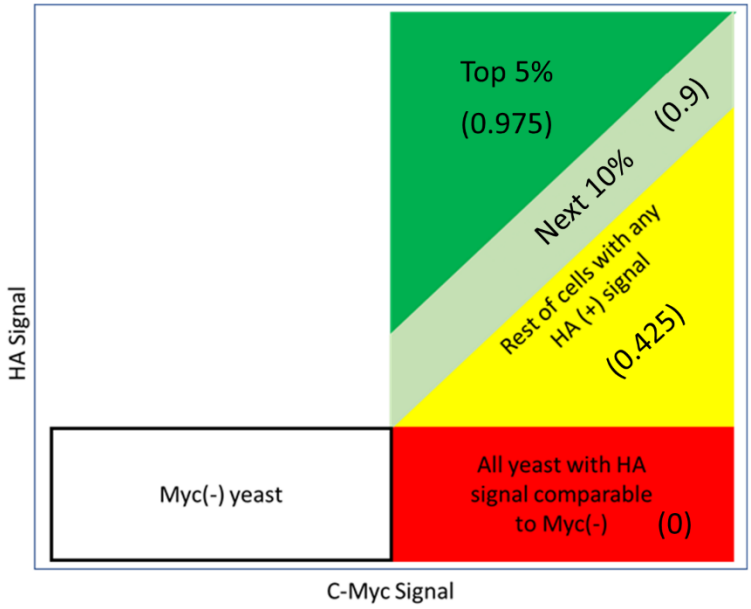
Supplemental Figure S4.2 Enrichment results for LysEFm5 controls expressed with N-terminal TorA (A, C, E) or OmpA (B, D, F) signal peptides across IPTG concentrations.

Data and error bars shown are the mean and standard deviation across quadruplicate samples. 0.5 mM IPTG concentration plots (E, F) are the same as those shown in main text.



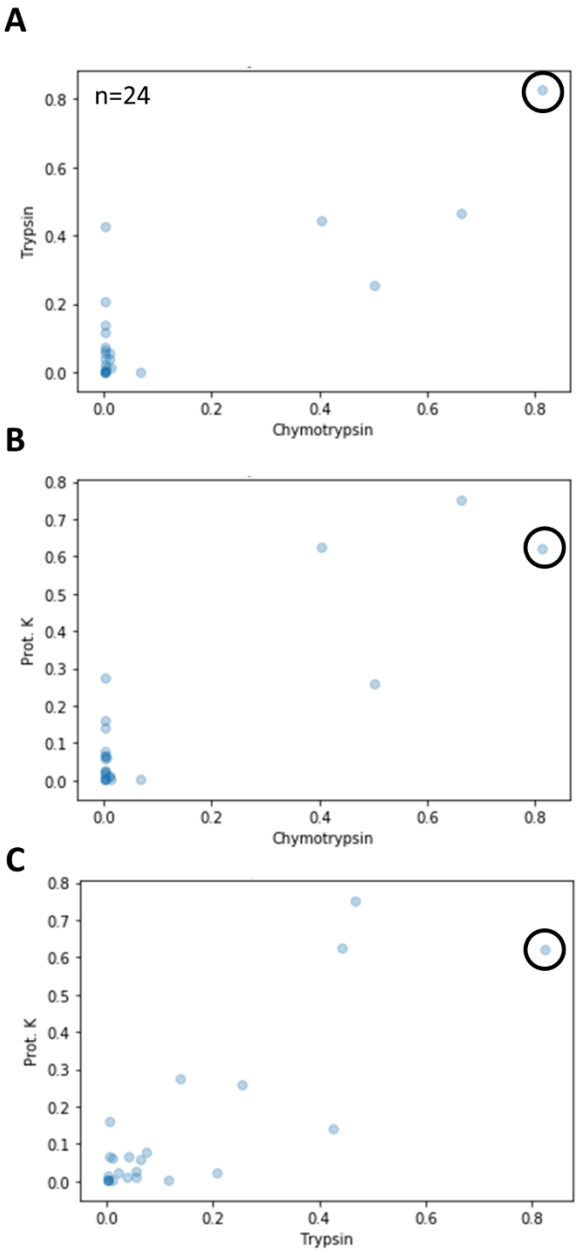
Supplemental Figure S4.3 Comparison of individual growth kinetics with population enrichment and previously characterized molecular activity[40] *for expression with the TorA signal peptide.*

(AB) Comparison of τ_{D1} (A) and $\Delta\tau$ (B) values with population enrichment. Population enrichment values are from the 0.5 mM IPTG, 8-hour time point samples. (CD) Comparison of τ_{D1} (C) and $\Delta\tau$ (D) values with molecular activity.



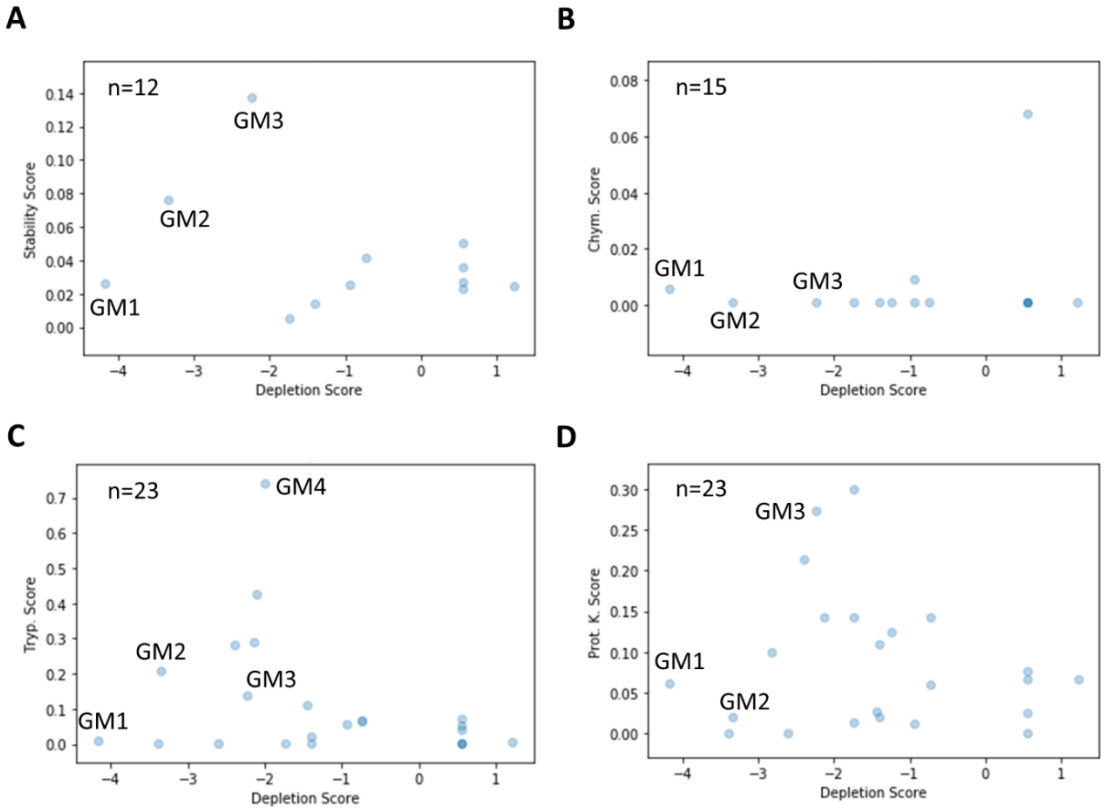
Supplemental Figure S4.4 Diagram of yeast sorting strategy for assessing protein stability.

Within each population, we collected yeast cells with (1) top 5% of N-terminal:C-terminal ratio (score: 0.975), next 10% (0.9), all remaining N-terminal-positive cells (0.425), and N-terminal-negative cells (0).



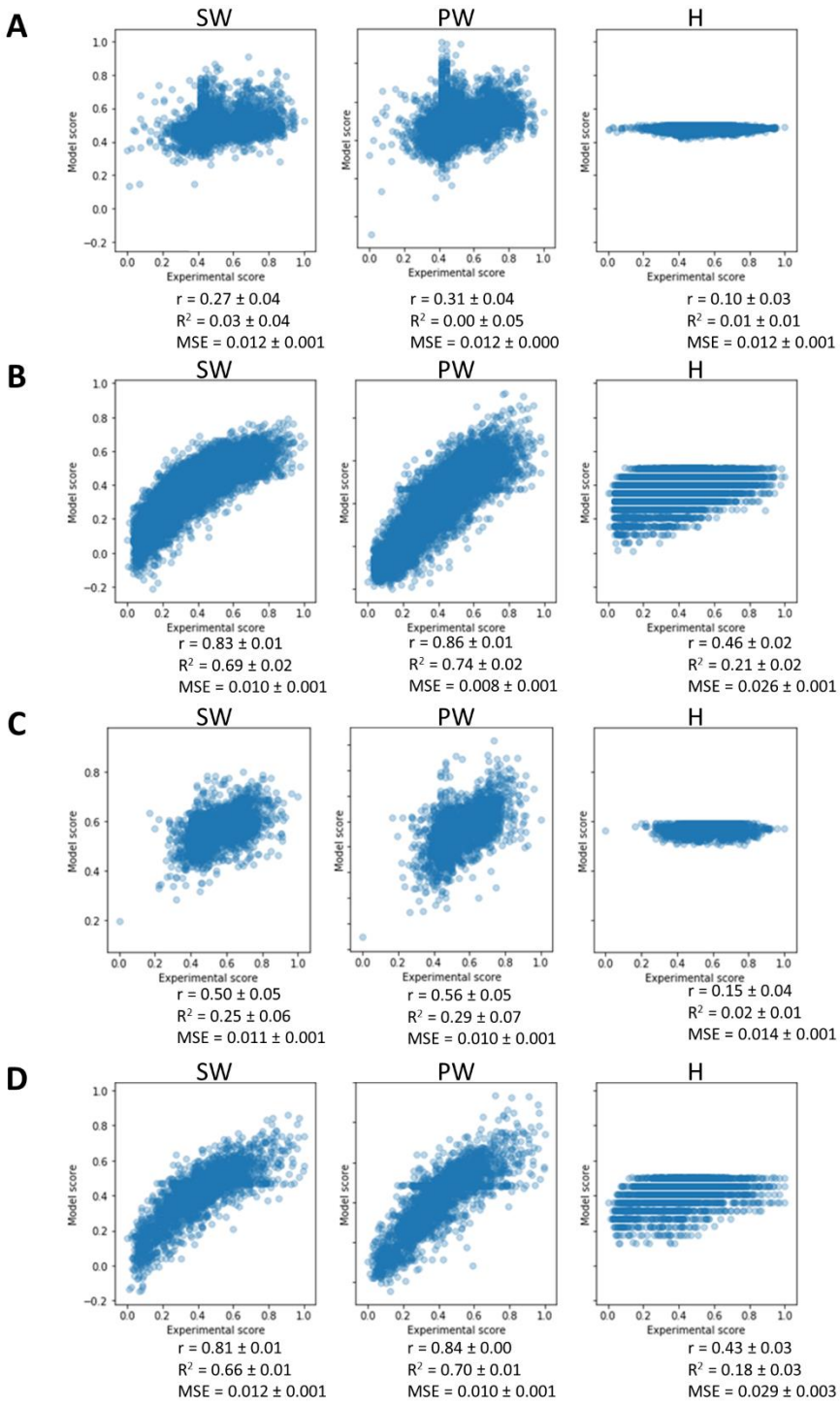
Supplemental Figure S4.5 Comparison of proteolytic stabilities of all variants in the GM library observed across all replicates of trypsin, chymotrypsin, and proteinase K experiments (n=24).

Figures show average stability to trypsin compared to chymotrypsin (A), proteinase K compared to chymotrypsin (B), and proteinase K compared t trypsin (C). Values represent the mean across all replicates. The parental LysV7 sequence is circled.



Supplemental Figure S4.6 Comparison of stability and activity scores of the GM library.

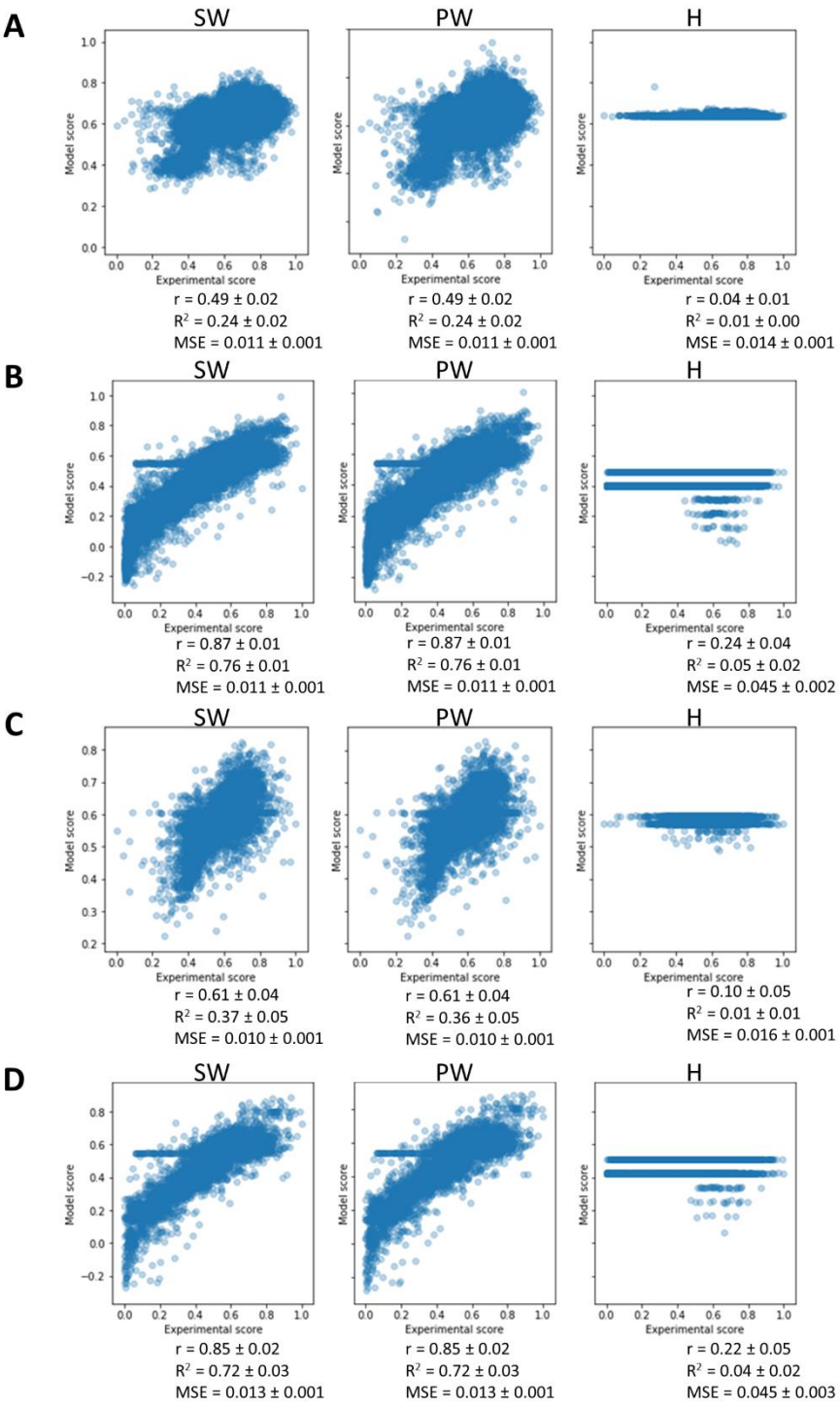
(A) Average stability score (A) or individual protease stability scores (B-D) compared to activity score. Average stability score is the mean of each individual protease stability score.



Supplemental Figure S4.7 Comparison of model and experimental scores for activity and stability in the Designed library.

(AB) Sitewise- (SW), pairwise- (PW), and Hamming distance-informed (H) models predicted activity (A) or stability (B) for all sequences observed in those specific assays. (CD) Sitewise- (SW), pairwise- (PW), and Hamming distance-informed (H) models predicted activity (C) or stability (D) for all sequences observed in both

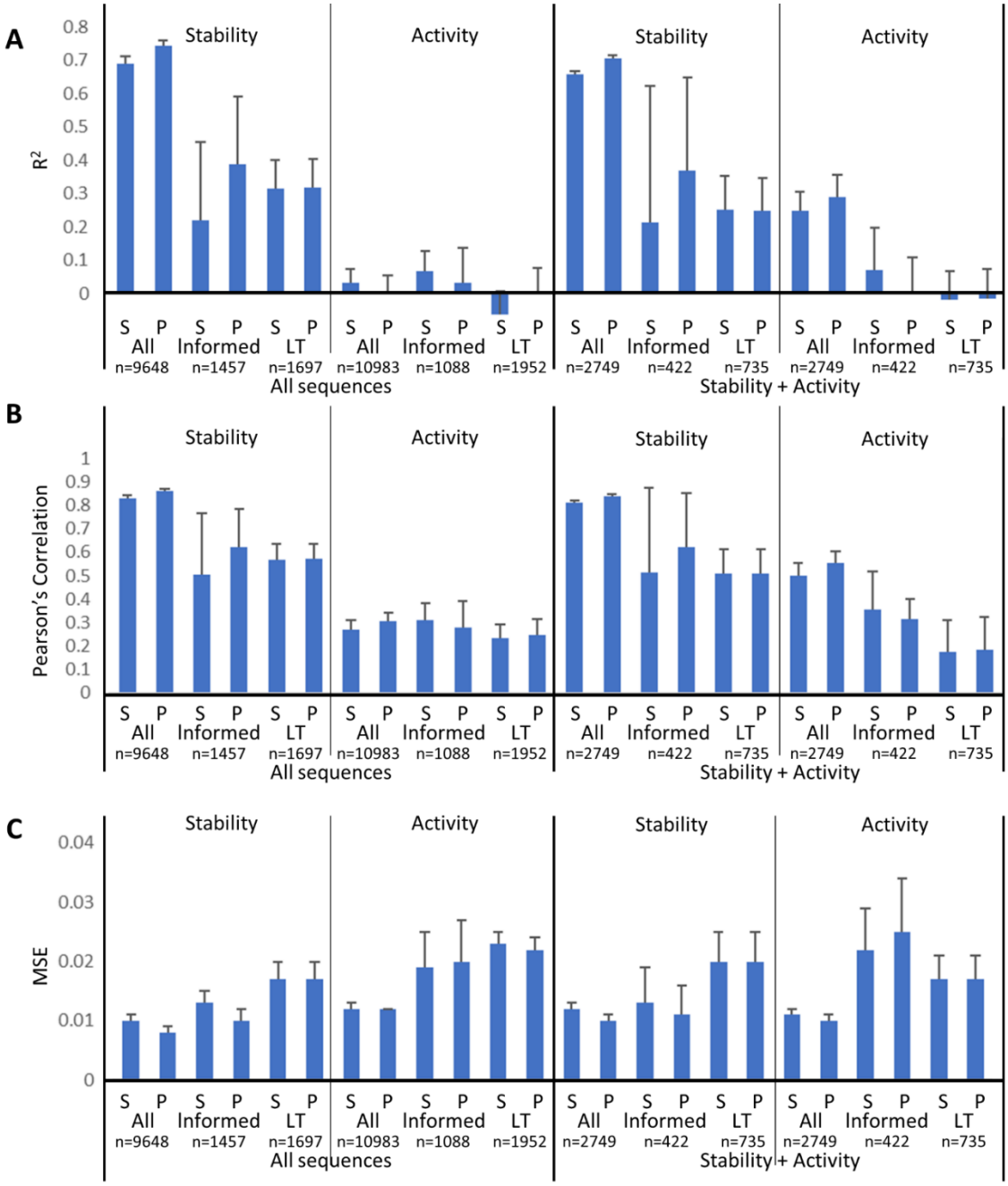
activity and stability assays. Model performance metrics are included below each plot. All experimental scores were normalized from [0, 1] for modeling.



Supplemental Figure S4.8 Comparison of model and experimental scores for activity and stability in the RM library.

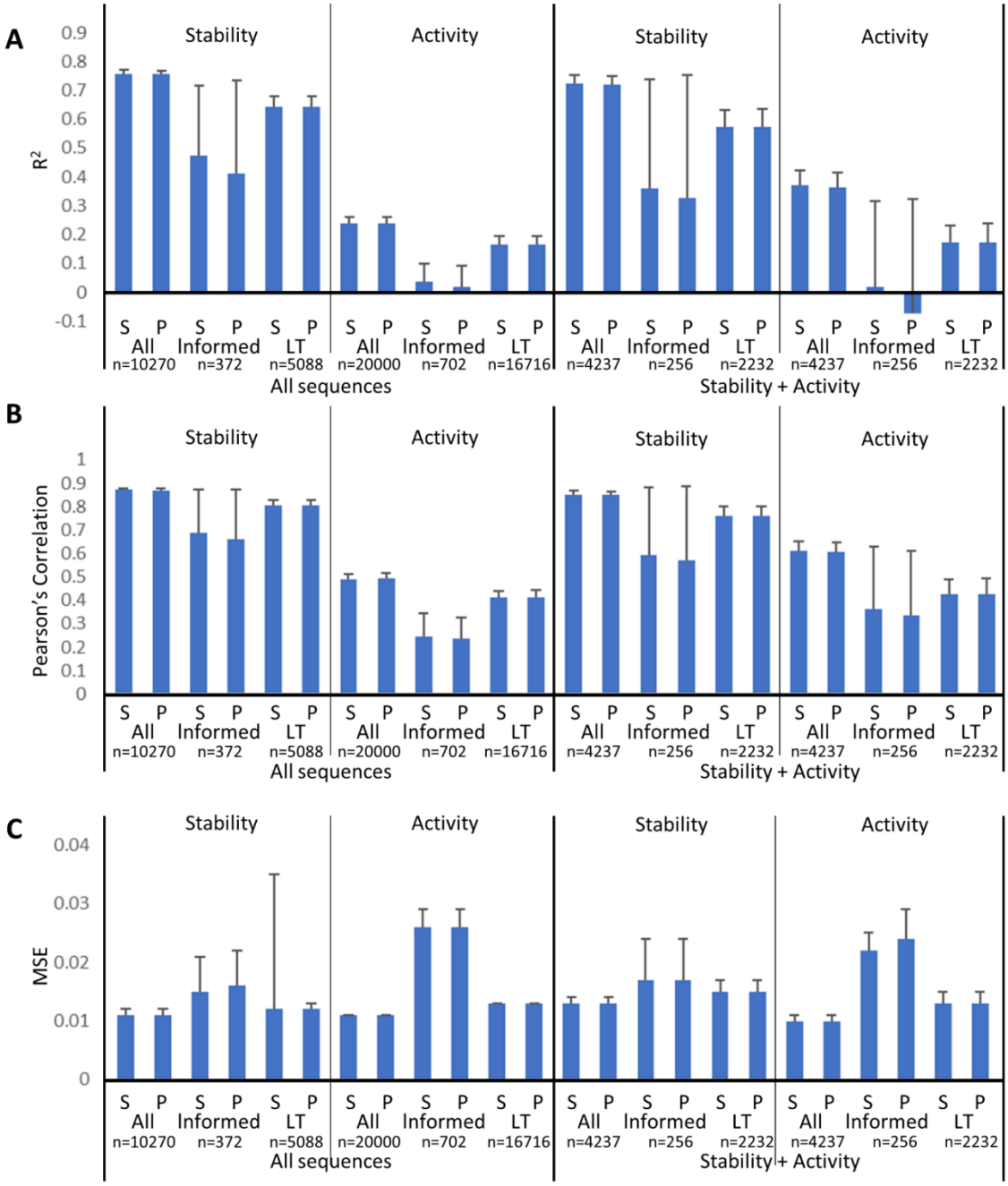
(AB) Sitewise- (SW), pairwise- (PW), and Hamming distance-informed (H) models predicted activity (A) or stability (B) for all sequences observed in those specific assays. (CD) Sitewise- (SW), pairwise- (PW), and Hamming distance-informed (H) models predicted activity (C) or stability (D) for all sequences observed in both

activity and stability assays. Model performance metrics are included below each plot. All experimental scores were normalized from [0, 1] for modeling.



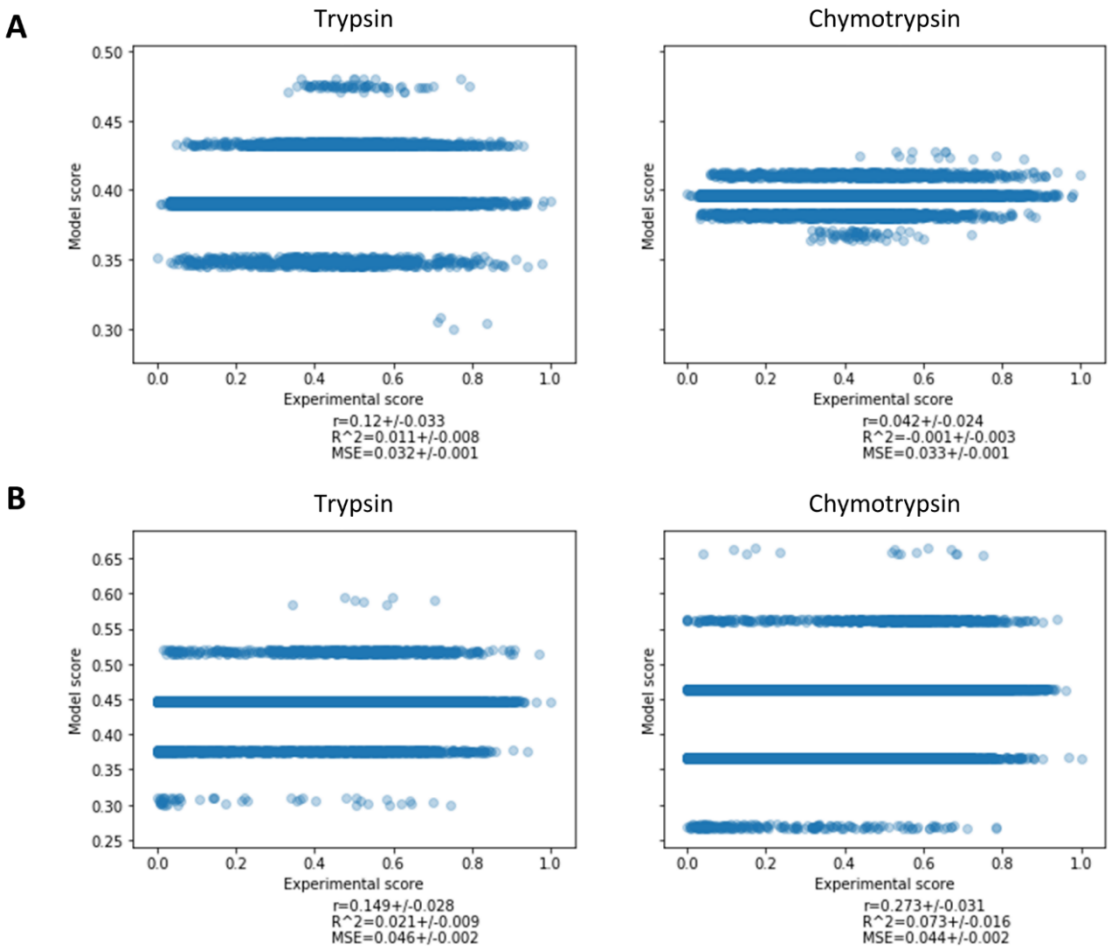
Supplemental Figure S4.9 All model metrics for models trained on the Designed library.

(ABC) R^2 (A), Pearson's correlation (B), and MSE (C) values for SW and PW models trained on all and subsets of Designed library for stability and activity. Models were trained on all sequences observed within activity or stability assays (left) or on all sequences observed in both activity and stability assays (right). Total number of sequences within each group is included below group titles. S and P denote SW and PW models, respectively.

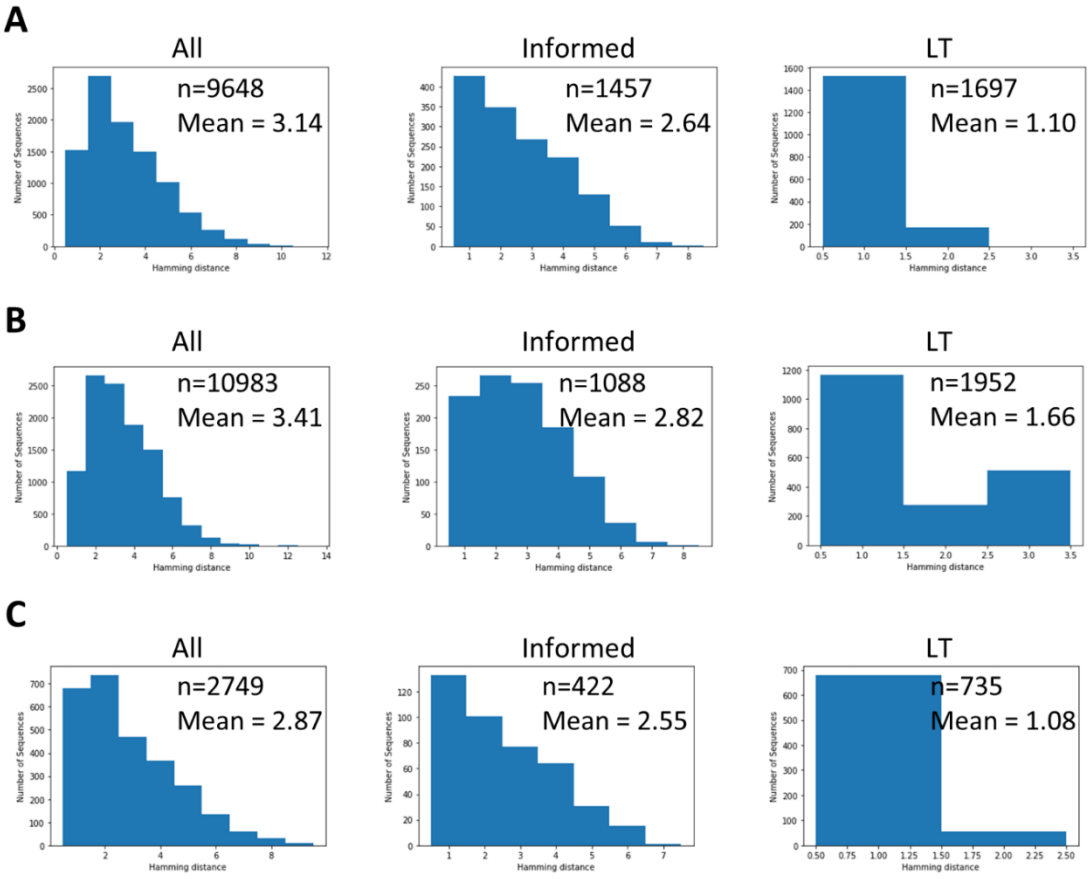


Supplemental Figure S4.10 All model metrics for models trained on the RM library.

(ABC) R^2 (A), Pearson's correlation (B), and MSE (C) values for SW and PW models trained on all and subsets of RM library for stability and activity. Models were trained on all sequences observed within activity or stability assays (left) or on all sequences observed in both activity and stability assays (right). Total number of sequences within each group is included below group titles. S and P denote SW and PW models, respectively.

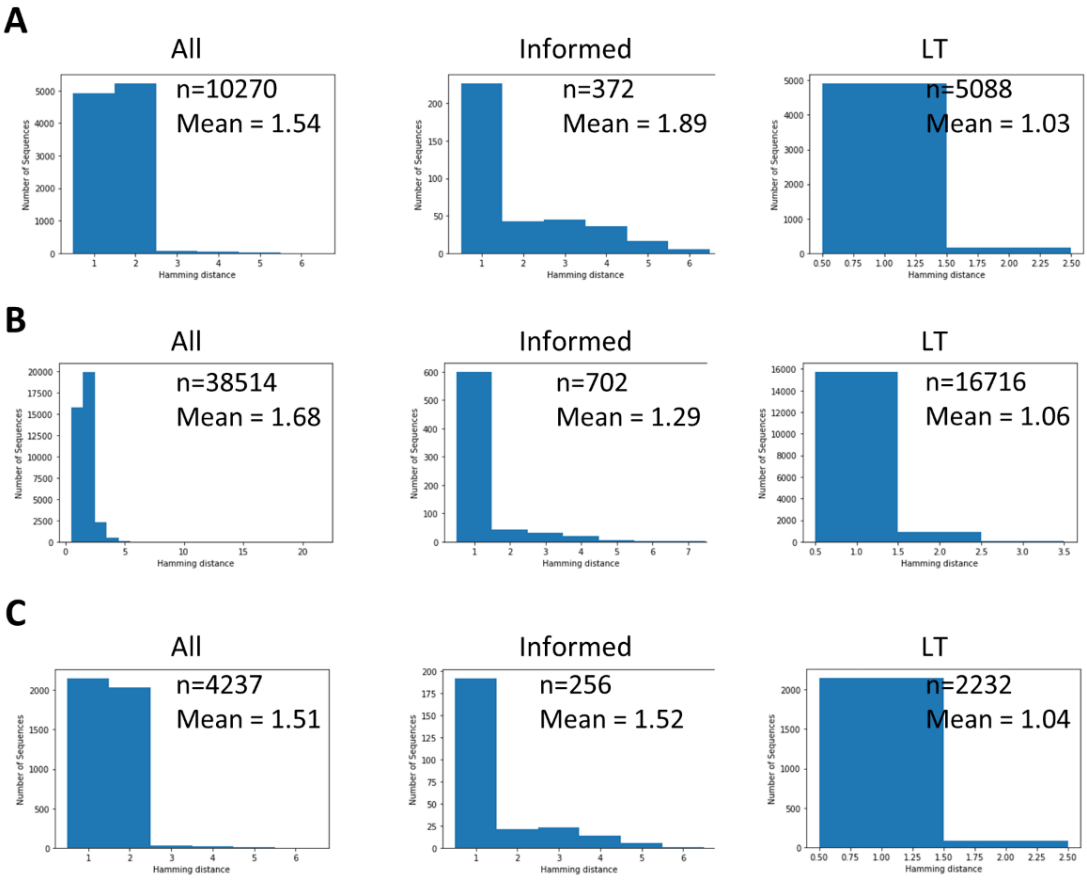


Supplemental Figure S4.11 Ridge regression model performance for predicting stability to trypsin or chymotrypsin when trained on number of cleavage sites. (AB) Model and experimental scores for ridge regression models trained on the number of trypsin or chymotrypsin cleavage sites for variants observed within the Designed (A) and RM (B) libraries. Model metrics are shown below each plot. Trypsin cleavage sites were identified as lysine and arginine amino acids and chymotrypsin cleavage sites were identified as tryptophan, phenylalanine, and tyrosine amino acids. All experimental scores were normalized from [0, 1] for modeling.



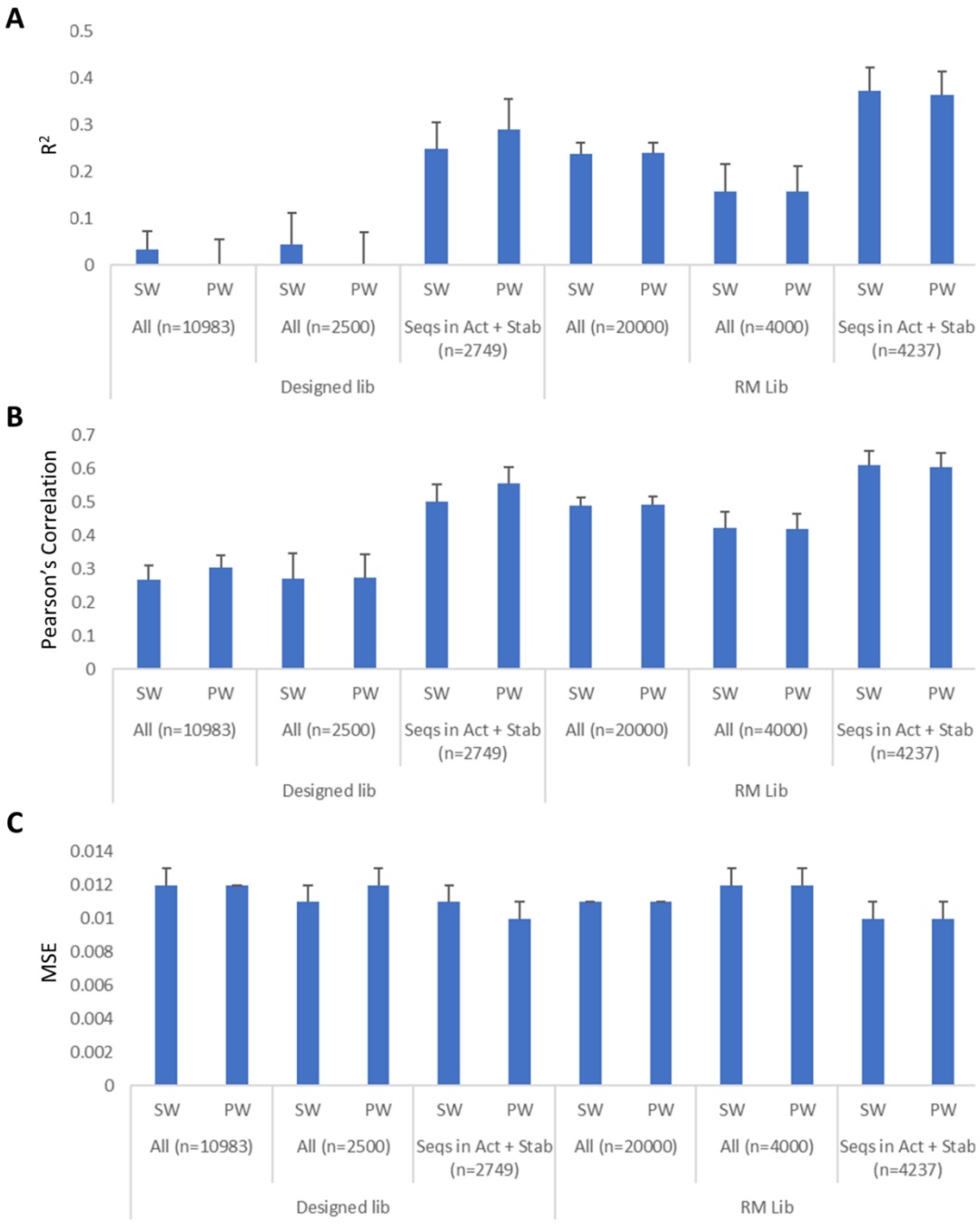
Supplemental Figure S4.12 Histograms of Hamming distances observed within each sub-library in the Designed library.

(ABC) Histogram of Hamming distance across sub-libraries observed in the stability assay (A), depletion assay (B), or both (C). “All” denotes all variants, “Informed” denotes variants with EV- and PROSS-informed mutations, and “LT” denotes variants with mutations at some i, i+1, i+2 positions.



Supplemental Figure S4.13 Histograms of Hamming distances observed within each sub-library in the RM library.

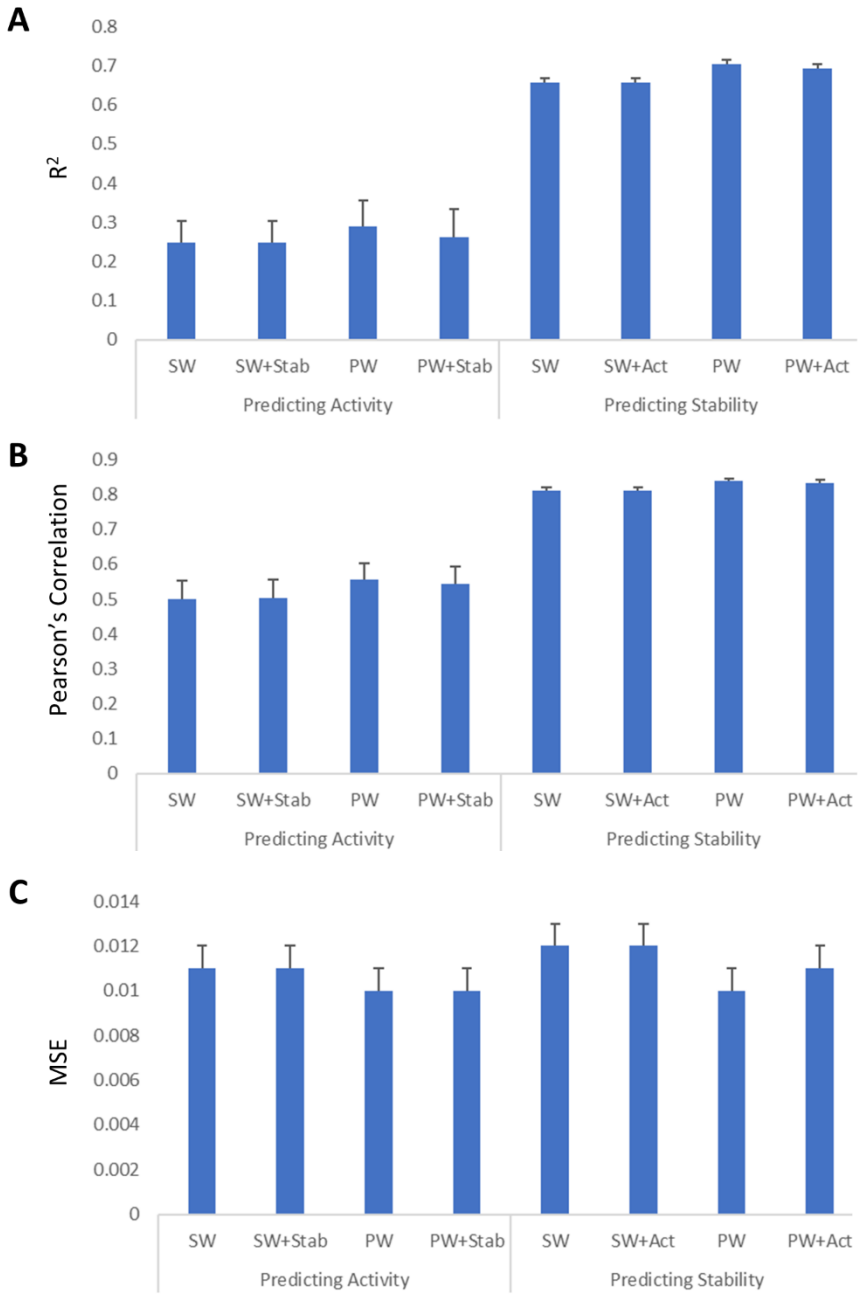
(ABC) Histogram of Hamming distance across sub-libraries observed in the stability assay (A), depletion assay (B), or both (C). “All” denotes all variants, “Informed” denotes variants with EV- and PROSS-informed mutations, and “LT” denotes variants with mutations at some i, i+1, i+2 positions.



Supplemental Figure S4.14 All activity model metrics for models trained on sequences present in activity and depletion data compared to sub-sampling of the Designed and RM libraries.

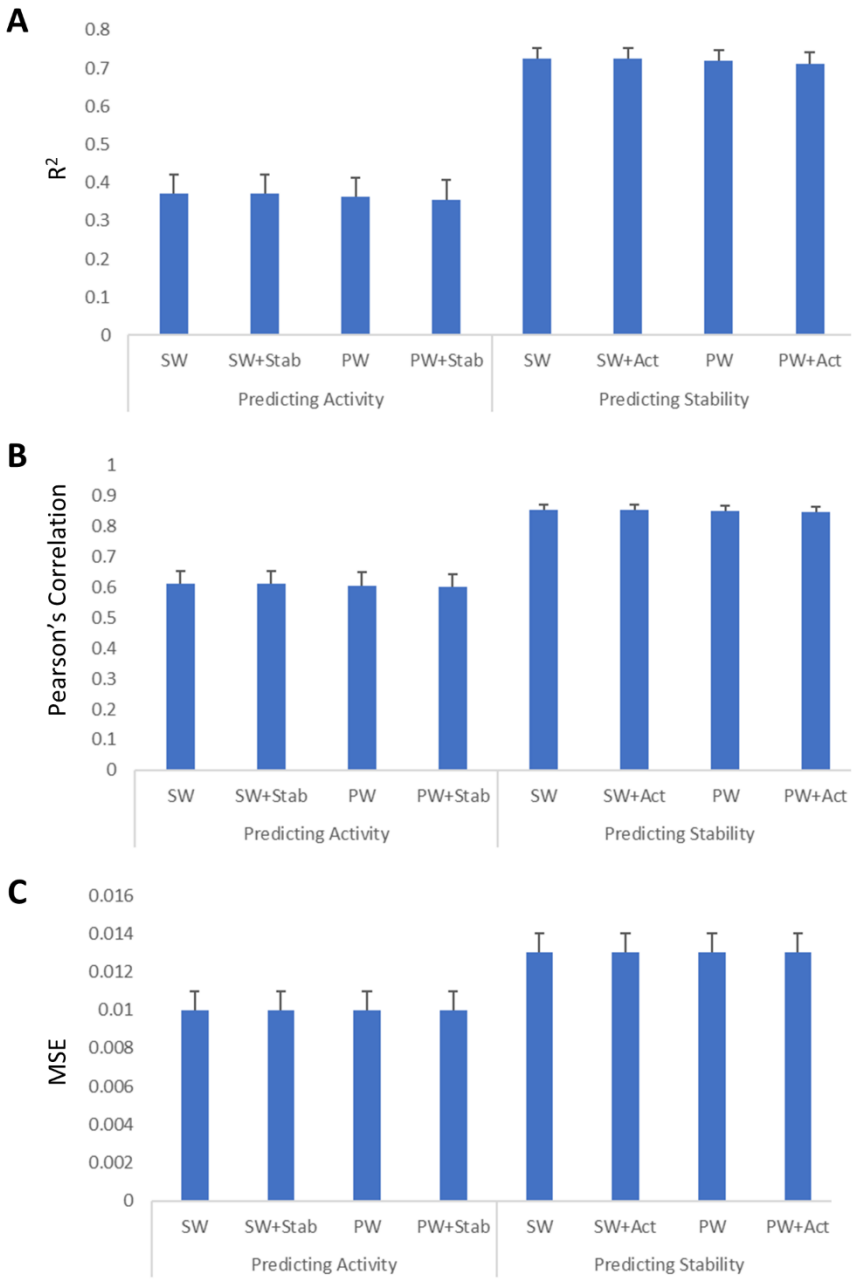
(ABC) R^2 (A), Pearson's correlation (B), and MSE (C) values for SW and PW models trained on all of the specified library, all sequences observed in both activity and stability data sets, or a comparable number of random sequences from the specified library. Values are shown for the Designed library (left) and the RM library (right). 2500 and 4000 sequences were randomly sampled from the Designed and RM libraries, respectively (compared to 2749 and 4237 sequences

observed in both activity and stability data sets). Total number of sequences within each group is included below group titles.



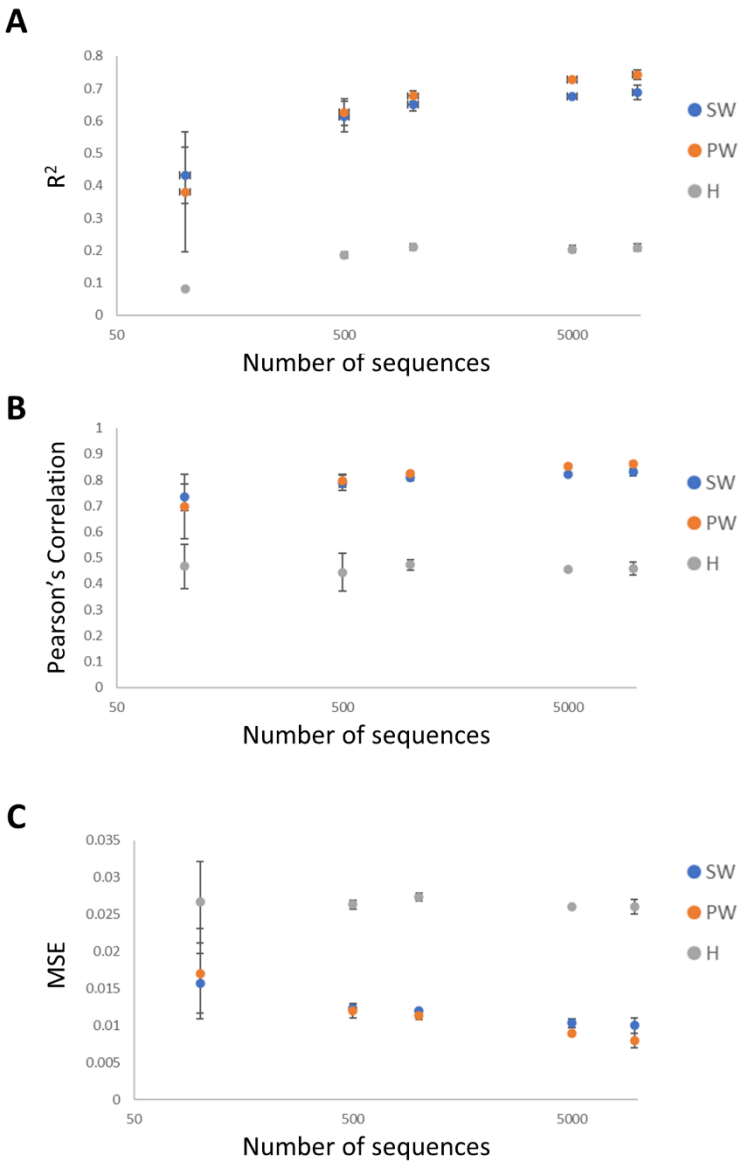
Supplemental Figure S4.15 Model performance metrics when trained with stability data to predict activity or activity data to predict stability for variants in the Designed library.

(ABC) R^2 (A), Pearson's correlation (B), and MSE (C) values for SW and PW models predicted with stability or activity data to predict activity or stability, respectively. "SW/PW+Stab" denotes models trained with stability data and "SW/PW+Act" denotes models trained with activity data. SW and PW model metrics (with no activity or stability data) are the same as those shown in the main text and Supplemental Figure 9.



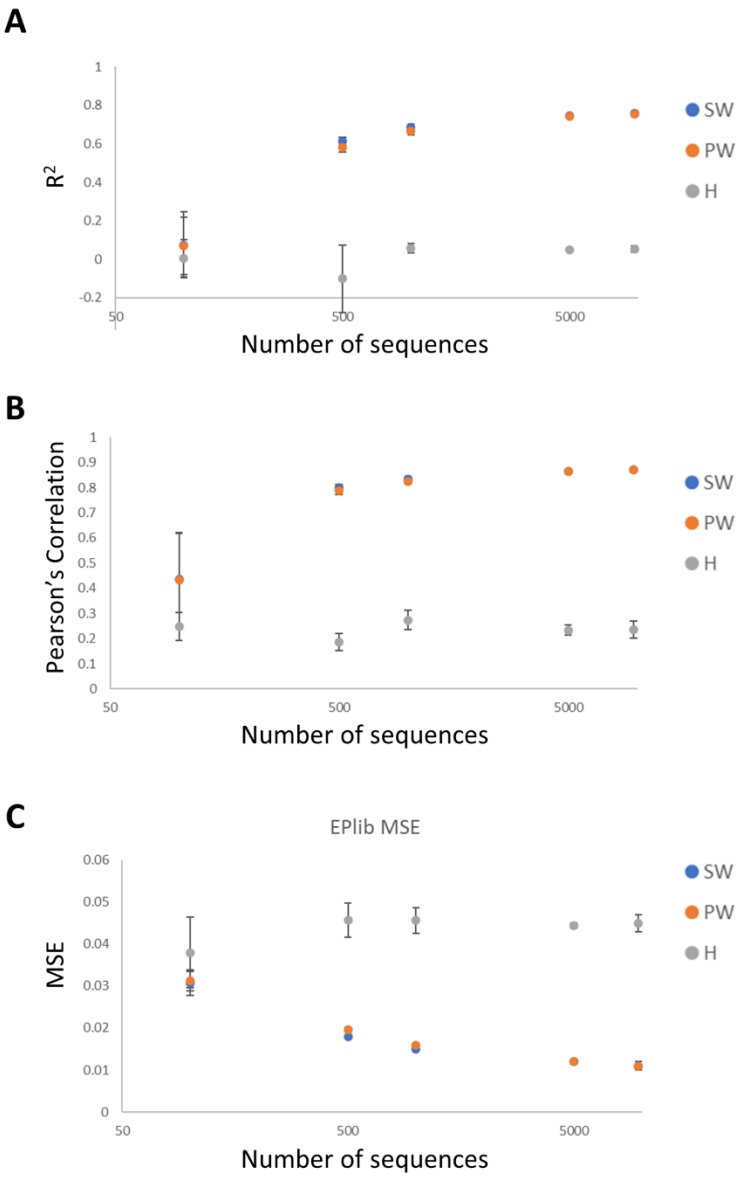
Supplemental Figure S4.16 Model performance metrics when trained with stability data to predict activity or activity data to predict stability for variants in the RM library.

(ABC) R^2 (A), Pearson's correlation (B), and MSE (C) values for SW and PW models predicted with stability or activity data to predict activity or stability, respectively. "SW/PW+Stab" denotes models trained with stability data and "SW/PW+Act" denotes models trained with activity data. SW and PW model metrics (with no activity or stability data) are the same as those shown in the main text and Supplemental Figure 10.



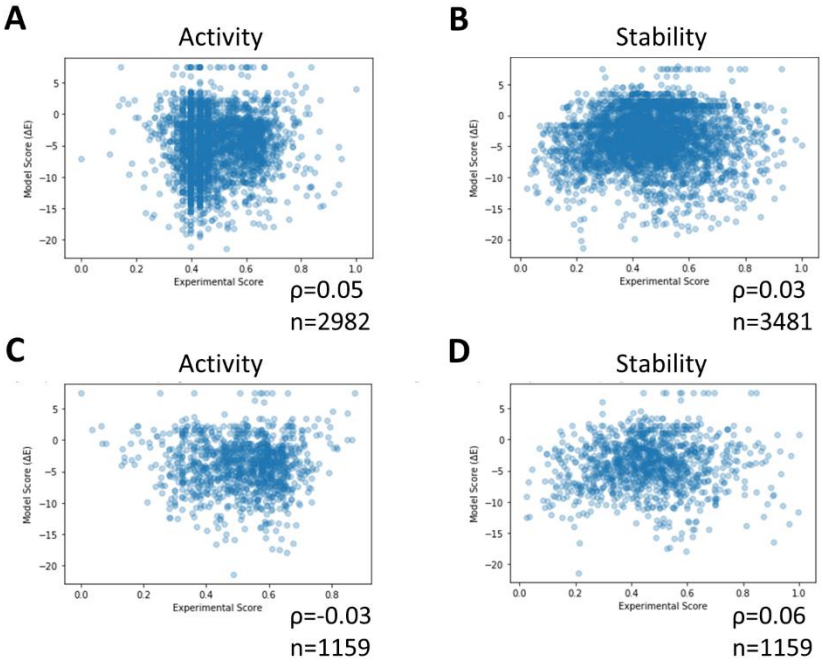
Supplemental Figure S4.17 Model performance metrics when subsampling the Designed library for predicting stability.

(ABC) Comparison of R^2 (A), Pearson's correlation (B), and MSE (C) values for SW-, PW-, and Hamming-distance informed models predicting stability with 100, 500,1000, 5000, or all of the sequences in the data set.

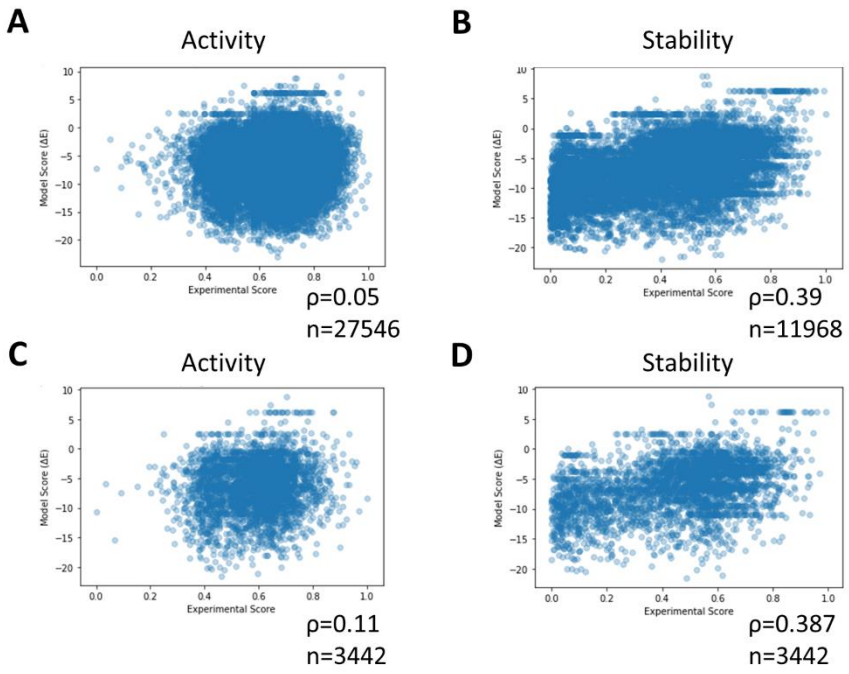


Supplemental Figure S4.18 Model performance metrics when subsampling the RM library for predicting stability.

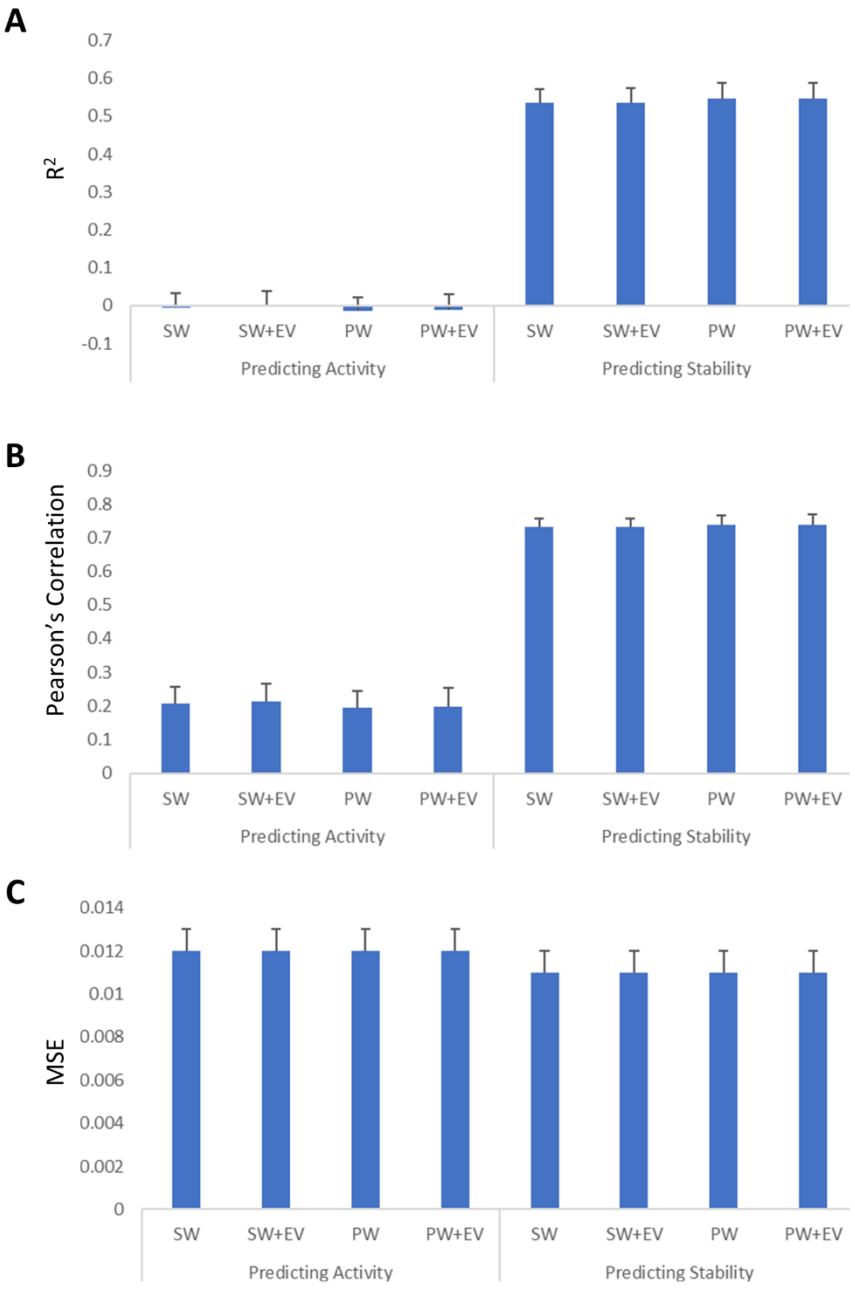
(ABC) Comparison of R^2 (A), Pearson's correlation (B), and MSE (C) values for SW-, PW-, and Hamming-distance informed models predicting stability with 100, 500,1000, 5000, or all of the sequences in the data set.



Supplemental Figure S4.19 Comparison of EVCouplings-predicted statistical fitness for variants containing one or two mutations in the Designed library.
(AB) EVCouplings-predicted statistical fitnesses and experimental activity (A) or stabililty (B) scores for all variants observed in those assays within the Designed library. (CD) EVCouplings-predicted statistical fitnesses and experimental activity (C) or stabililty (D) scores for all variants observed in both of those assays within the Designed library. Correlation and number of variants are given below each plot. All experimental scores were normalized from [0, 1] for modeling.

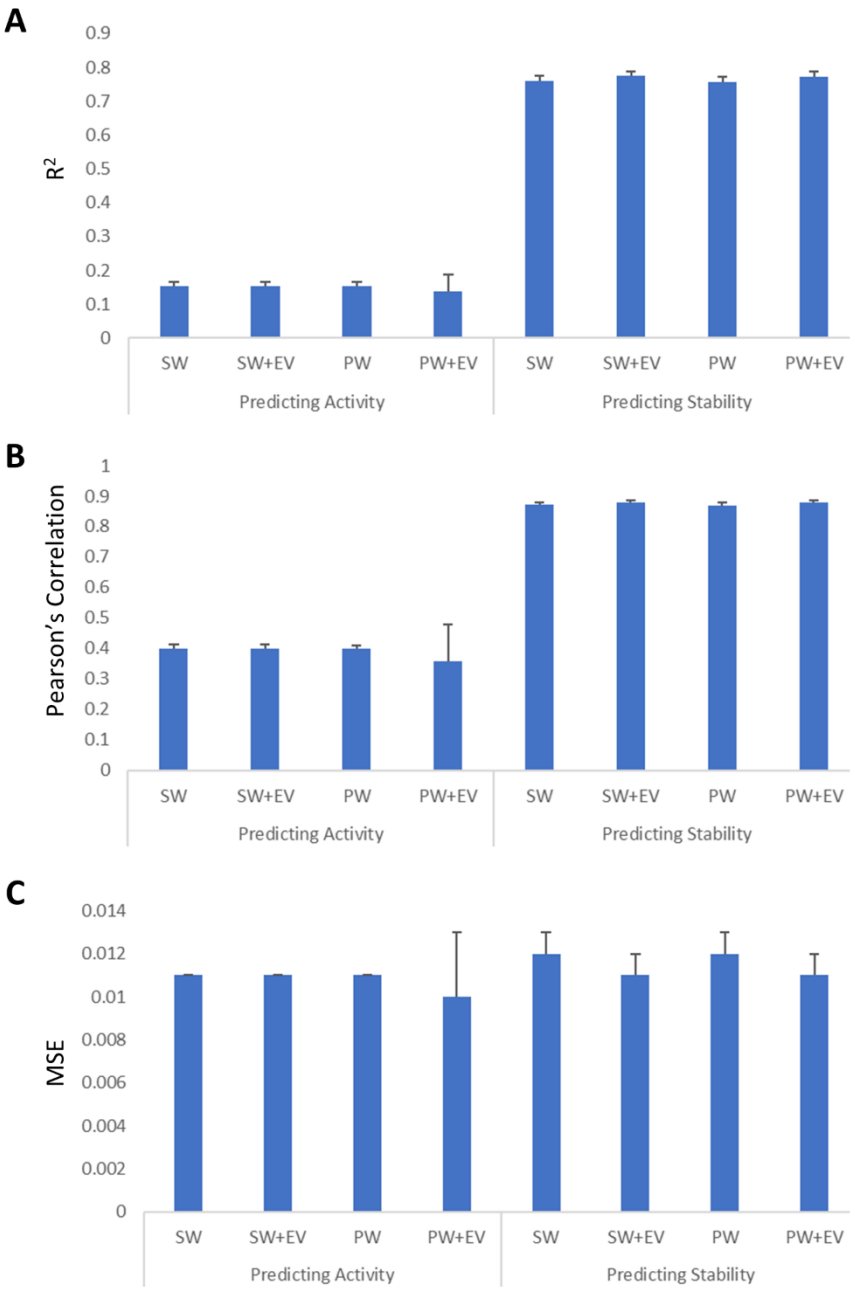


Supplemental Figure S4.20 Comparison of EVCouplings-predicted statistical fitness for variants containing one or two mutations in the RM library.
(AB) EVCouplings-predicted statistical fitnesses and experimental activity (A) or stabililty (B) scores for all variants observed in those assays within the RM library. (CD) EVCouplings-predicted statistical fitnesses and experimental activity (C) or stabililty (D) scores for all variants observed in both of those assays within the RM library. Correlation and number of variants are given below each plot. All experimental scores were normalized from [0, 1] for modeling.



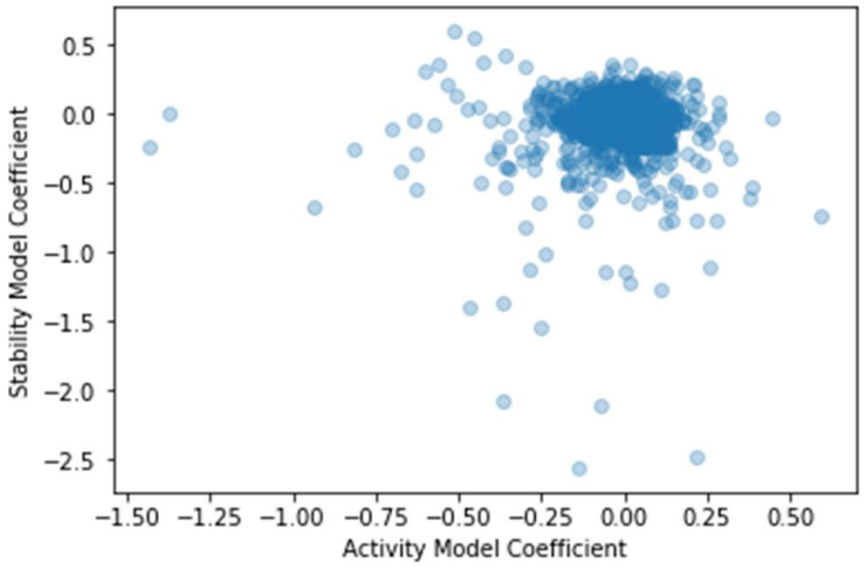
Supplemental Figure S4.21 Model performance metrics when trained with EVCouplings information to predict activity or stability for variants in the Designed library.

(ABC) R^2 (A), Pearson's correlation (B), and MSE (C) values for SW and PW models predicted with stability or activity data to predict activity or stability, respectively. "SW/PW+EV" denotes models trained with EVCouplings data. SW and PW model metrics (with no activity or stability data) are the same as those shown in the main text and Supplemental Figure 9.



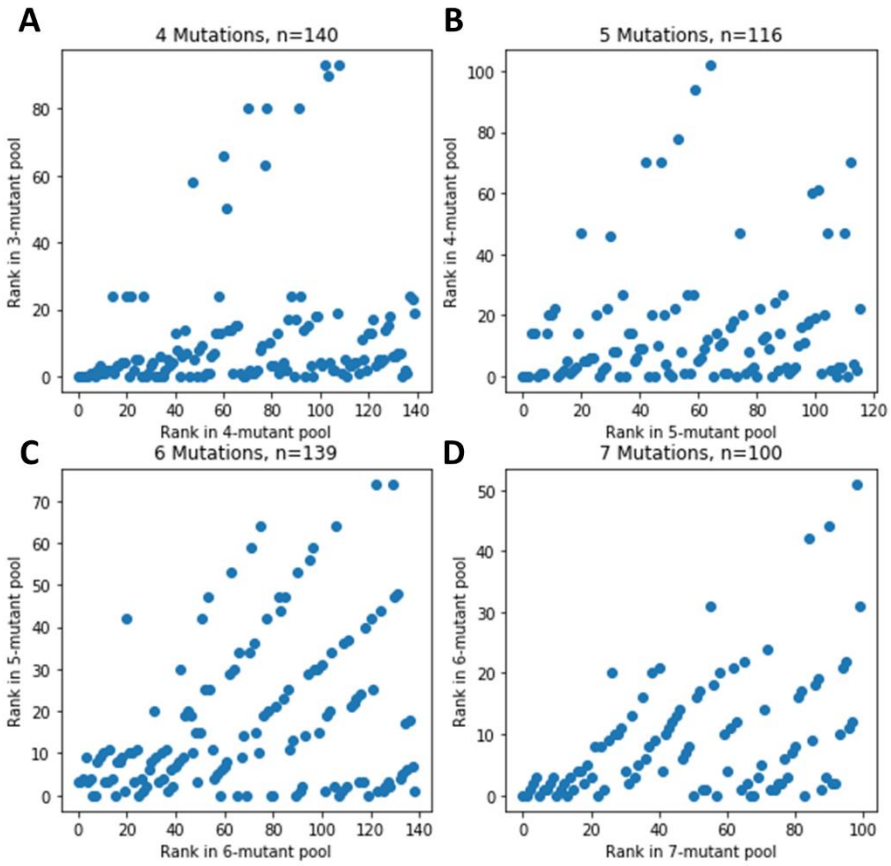
Supplemental Figure S4.22 Model performance metrics when trained with EVCouplings information to predict activity or stability for variants in the RM library.

(ABC) R^2 (A), Pearson's correlation (B), and MSE (C) values for SW and PW models predicted with stability or activity data to predict activity or stability, respectively. "SW/PW+EV" denotes models trained with EVCouplings data. SW and PW model metrics (with no activity or stability data) are the same as those shown in the main text and Supplemental Figure 10.



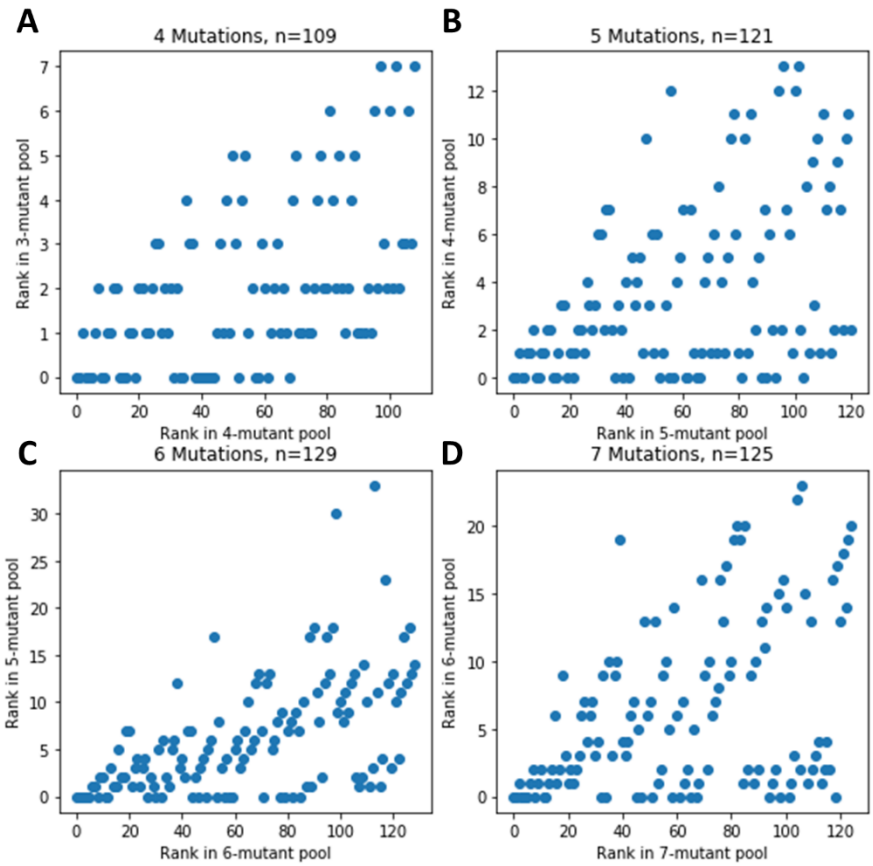
Supplemental Figure S4.23 Comparison of stability and activity model coefficients.

Model coefficients are those from heat maps shown in Figure 6 A (stability) and B (activity).



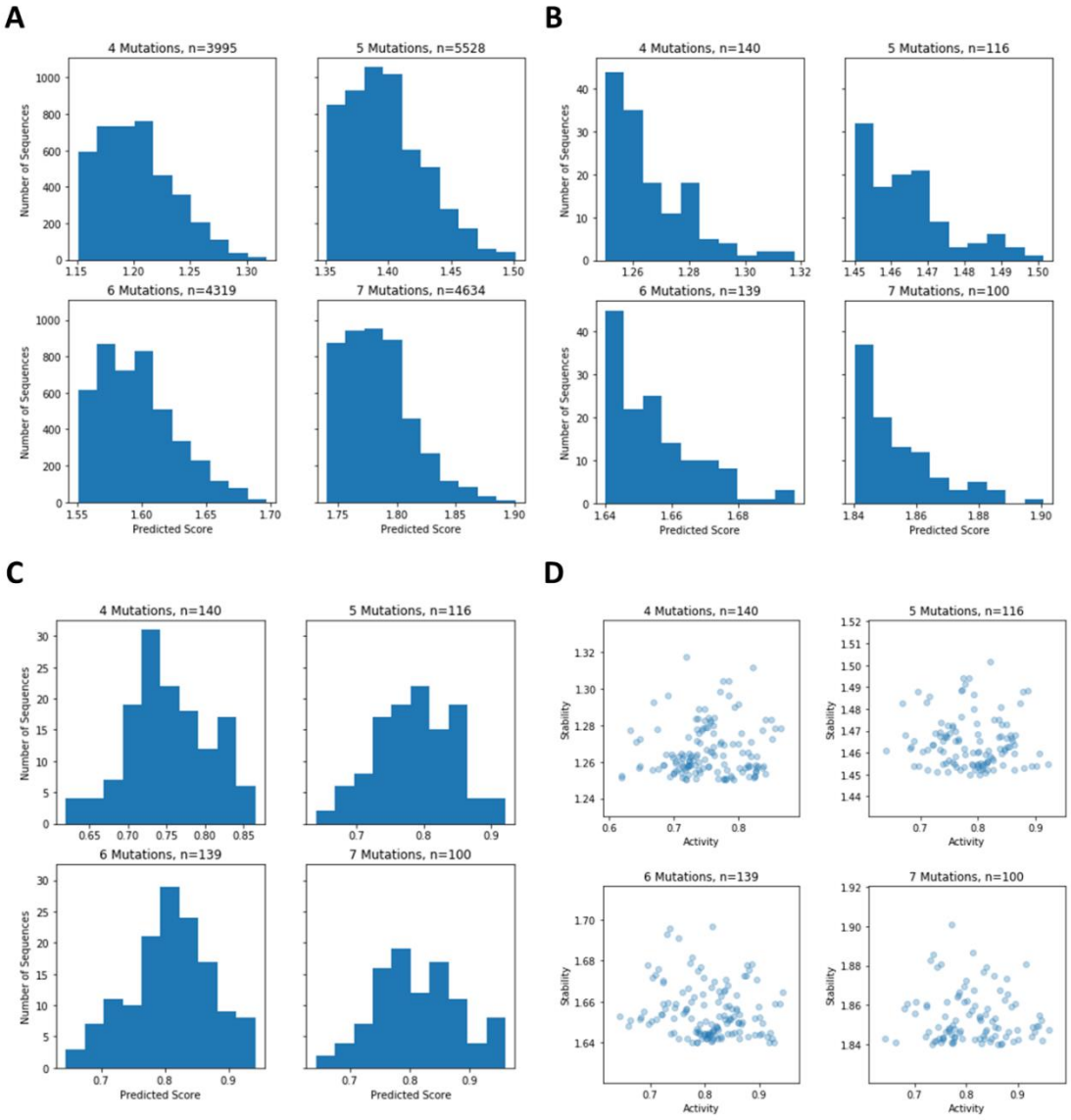
Supplemental Figure S4.24 Comparison of rank of predicted variants in the i^{th} pool with highest performing parental variant in the $i\text{-}1^{th}$ pool for variants predicted by the PW model from the Designed library.

(ABCD) Plots show the rank of best variants with four mutations compared to the rank of best performing parent with 3 mutations (A), rank of variants with five mutations compared to rank of parent with 4 mutations (B), rank of variants with 6 mutations compared to rank of parent with 5 mutations (C), and rank of variants with 7 mutations compared to rank of parent with 6 mutations (D). PW model was informed by variants observed in both activity and stability data within the Designed library. Total number of predicted variants in i^{th} pool are shown above each plot.



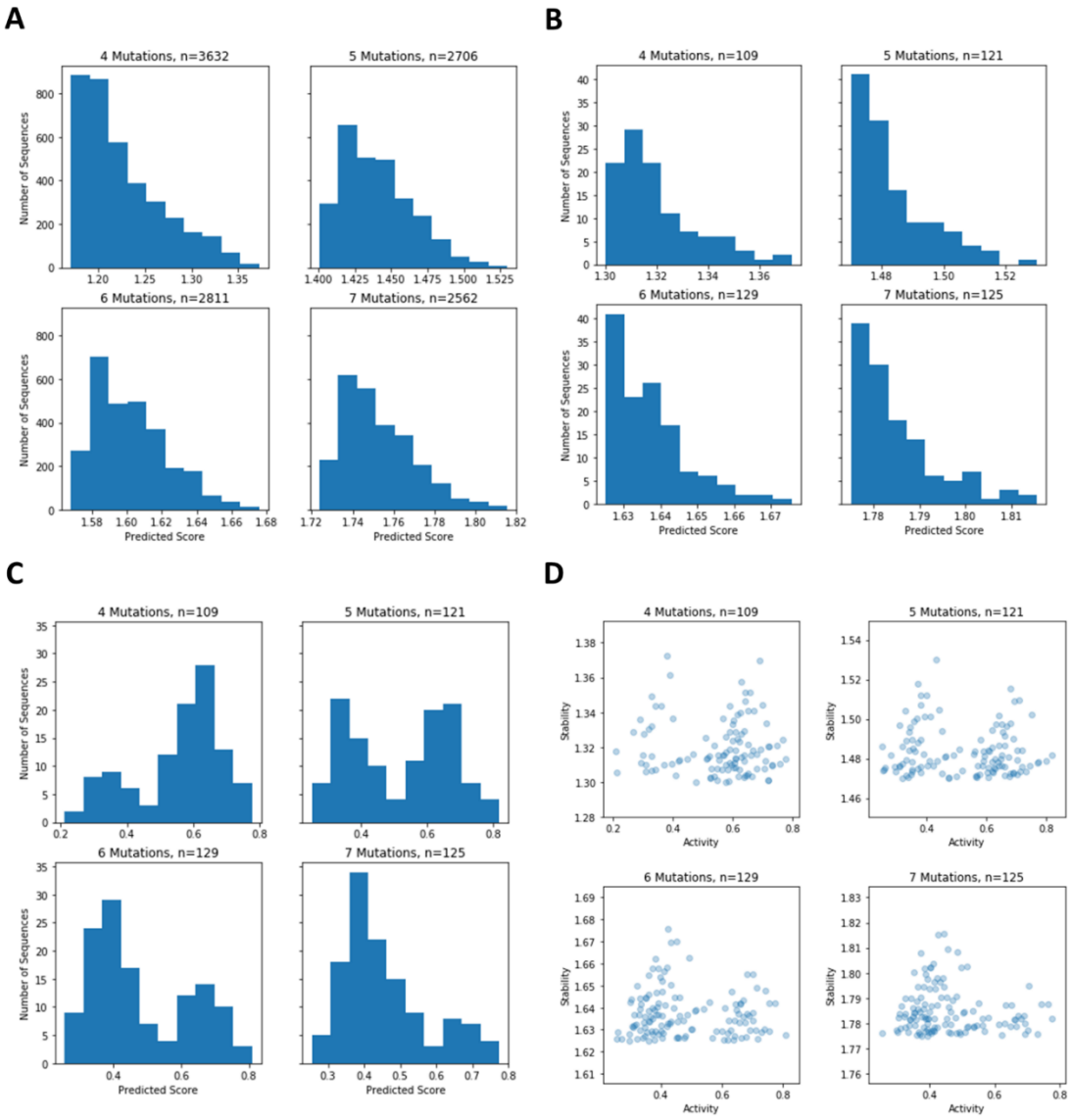
Supplemental Figure S4.25 Comparison of rank of predicted variants in the i^{th} pool with highest performing parental variant in the $i\text{-}1^{th}$ pool for variants predicted by the PW model from the RM library.

(ABCD) Plots show the rank of best variants with four mutations compared to the rank of best performing parent with 3 mutations (A), rank of variants with five mutations compared to rank of parent with 4 mutations (B), rank of variants with 6 mutations compared to rank of parent with 5 mutations (C), and rank of variants with 7 mutations compared to rank of parent with 6 mutations (D). PW model was informed by variants observed in both activity and stability data within the RM library. Total number of predicted variants in i^{th} pool are shown above each plot.



Supplemental Figure S4.26 Model activity and stability scores of variants predicted by Designed library PW model.

(A) Histogram of stability scores of all predicted variants tested containing 4-7 mutations during identification of variants DP1-2. (B) Histogram of stability scores of predicted variants containing 4-7 mutations used to construct next round of variants containing i+1 mutations. Histogram of variants containing 7 mutations is shown for comparison. (C) Histogram of predicted activity scores of all variants shown in B. (D) Comparison of predicted activity and stability scores of variants shown in B and C. Number of sequences is shown above each plot.



Supplemental Figure S4.27 Model activity and stability scores of variants predicted by RM library PW model.

(A) Histogram of stability scores of all predicted variants tested containing 4-7 mutations during identification of variants RP1-2. (B) Histogram of stability scores of predicted variants containing 4-7 mutations used to construct next round of variants containing i+1 mutations. Histogram of variants containing 7 mutations is shown for comparison. (C) Histogram of predicted activity scores of all variants shown in B. (D) Comparison of predicted activity and stability scores of variants shown in B and C. Number of sequences is shown above each plot.

Supplemental Table S4.1 DNA primers for construction of pCT expression plasmid and insertion of lysin catalytic domains into pCT and pET plasmids for high throughput screening.

Primer Name	DNA
pCT-LysV7-insert-F	TTATTGCTTCAGTTTTAGCAGCTAGCTACCCATACGACGTTCCAGACTACGCTCTGCAGGCTAGTATGGTTGAGATTATTAATAAAACTG
pCT-LysV7-insert-R	TCAGAAATAAGCTTTTGTTCGGATCCCAGGCTCAGATAGATGGAGTCAATGCGGACG
pCT-LysV7-amp-F	TTATTGCTTCAGTTTTAGCAG
pCT-LysV7-amp-R	TCAGAAATAAGCTTTTGTTC
Aga2SP-Amp-F	GTCAAGGAGAAAAAACTATAGAATTC
Aga2SP-Gene-Amp-R	GTCGCCCAAAATTTGCTAGCTGCTAAAACTGAAGCAATAACAGAAAATATTG
Gene-Amp-F	GCTAGCAAATTTTGGGCGACTGTAGAATCCTCTGA
Gene-Cmyc-Amp-R	CAAGTCCTCTTCAGAAATAAGCTTTTGTTC
Cmyc-G4S-Aga2-Insert	TTATTTCTGAAGAGGACTTGGGTGGTGGTGGTTCTGGTGGTGGTGGTTCTGGTGGTGGTGGTTCTCTGCAGCAGGAACTGACAACTATAT
Aga2-Amp-F	CTGCAGCAGGAACTGACAACTATATG
Aga2-pCT-Amp-R	CTACACTGTTGTTATCAGATCTCGAGCTATTAAAAAACATACTGTGTGTTTATGGGGCTG
pET-ompA-Fwd 1	AAGGAGATATACATATGGCTAGCATGAAAAAAACCGCGATCGCGATCGCGGTTGCGCTGGCGGGTTTCGCGACCGTTGCGCAGGCGATGG
pET-ompA-Fwd 2	GACCGTTGCGCAGGCGATGGTTGAGATTATTAATAAAACTGTTACTCGTGG
pET-TorA-Fwd 1	AAGGAGATATACATATGGCTAGCATGAACAACAACGACCTGTTCCAGGCGTCTCGTCGTCGTTTCCTGGCGCAGCTGGGTGGTCTGACCG
pET-TorA-Fwd 2	CGCAGCTGGGTGGTCTGACCGTTGCGGGTATGCTGGGTCCGTCTCTGCTGACCCCGCGTC
pET-TorA-Fwd 3	GCTGACCCCGCGTCGTGCGACCGCGATGGTTGAGATTATTAATAAAACTGTTACTCGTGG
pET-LysV7-amp-F	AAGGAGATATACATATGGCTAGC
pET-LysV7-amp-R	TCAGTGATGATGGTGATGGTGGGATCC
Oligopool 1 Fwd	GCTCTGCAGGCTAGT
Oligopool 1 Rev 1	GCRGACTTCATATCCAAK

Oligopool 1 Rev 2	KTTGACTTCATATCCAAK
Oligopool 2 Fwd 1	MTTGGATATGAAGTCYGC
Oligopool 2 Fwd 2	MTTGGATATGAAGTCAAM
Oligopool 2 Rev	ATCTGGTTTCAAACTGGG
LysV7-epPCR-Fwd	GTTGCTGGTCGTCGC
LysV7-epPCR-Rev	TTGAGCCTGGGTGGAATTAAA
LysV7-epPCR-Fwdtoend	TTTAATTCCACCCAGGCTCAA
LysV7-epPCR-Revtobeg	GCGACGACCAGCAAC

Supplemental Table S4.2 Previously evaluated lysin controls used in this study[40].

LysEFm5 Variant	Mutation	Reported Activity ($-\Delta OD_{600}$/min/μg)
LysV1	T40P, N47V	0.048 ± 0.006
LysV4	N32G, E38T	0.025 ± 0.004
LysV6	N47A, V91P	0.028 ± 0.003
LysV7	T34S, A35V	0.087 ± 0.013
LysV8	T40A	0.046 ± 0.003
LysV11	M45P, I87G	ND
LysV12	A35G	ND
LysV14	A74T, N83Y	0.008 ± 0.004

Supplemental Table S4.3 Number of cells collected in FACS for each library, protease replicate, and gate.

Genome-mined Library	Chymotrypsin			Trypsin			Proteinase K		
Gate		Replicate 2	Replicate 3	Replicate 1	Replicate 2	Replicate 3	Replicate 1	Replicate 2	Replicate 3
P5		50000	50000	50000	50000	50000	50000	50000	50000
P6		101034	115296	100000	132611	134013	84511	80661	131632
P7		29191	36698	310893	280246	245234	28122	32678	256874
P8		1150125	1001989	425000	832706	894777	1167045	1275537	856032
Designed Library	Chymotrypsin			Trypsin			Proteinase K		
Gate	Replicate 1	Replicate 2	Replicate 3	Replicate 1	Replicate 2	Replicate 3	Replicate 1	Replicate 2	Replicate 3
P5	799194	565963	810828	617437	1000000	838792	1000000	1000000	1000000
P6	1189218	951554	1362510	956180	1286674	1880952	1410751	1962004	1483579
P7	3006871	3425927	4000000	2765088	3820773	4000000	7981596	4000000	4000000
P8	8500000	13532535	4000000	8500000	4000000	4000000	8500000	4000000	4000000
EP Library	Chymotrypsin			Trypsin			Proteinase K		
Gate	Replicate 1	Replicate 2	Replicate 3	Replicate 1	Replicate 2	Replicate 3	Replicate 1	Replicate 2	Replicate 3
P5	582781	908245	459071	953988	1000000	851856	1000000	1000000	1000000
P6	1804527	2101034	802315	1115275	1590836	1794758	1499506	1532304	2219193
P7	4903921	12851428	4000000	5668124	4000000	4000000	8500000	4000000	4000000
P8	4730942	16942965	4000000	7622494	4000000	4000000	8500000	4000000	3665408
Tiled Library	Chymotrypsin			Trypsin			Proteinase K		
Gate	Replicate 1	Replicate 2	Replicate 3	Replicate 1	Replicate 2	Replicate 3	Replicate 1	Replicate 2	Replicate 3
P5	100000	500000	500000	500000	500000	500000	222762	500000	500000
P6	539196	882500	769809	1000000	923178	1257780	691660	1165397	1049150
P7	1539944	5314872	4000000	4250000	4000000	4000000	4250000	4000000	4000000
P8	2408003	7452351	4000000	4250000	4000000	4000000	4250000	4000000	1988575

Supplemental Table S4.4 EVCouplings-informed library design[142,181].

Mutated positions	WT Codon	WT Amino Acid	Library Codons	Library Amino Acids	Diversity
20	GCT	A	GCT, RAM	A, K, N, E, D	5
26	TTT	F	NTT	F, I, V, L	4
30	TGG	W	TGG, GCA	W, A	2
34	TCT	S	ASC	S, T	2
35	GTG	V	GYG	V, A	2
38	GAA	E	GAA, TAT	E, Y	2
40	ACT	T	AMT	T, N	2
48	AAC	N	GCC, AAM	N, A, K	3
51	GCC	A	GMG	A, E	2
61	AAG	K	RAA	K, E	2
67	ACG	T	ACC, GTG	T, V	2
71	TTC	F	TWT	F, Y	2
73	GGT	G	GGT, ATT	G, I	2
80	GGC	G	GSG, CST	G, A, P, R	4
95	GAT	D	KWT	D, Y, F, V	4
100	AAG	K	AAG, GCG	K, A	2
109	ACA	T	RCA	T, A	2
113	ATT	I	ATT, GYG	I, A, V	3

Total Diversity	1.18E+07

Supplemental Table S4.5 PROSS-informed library design[192].

Mutated positions	WT Codon	WT Amino Acid	Library Codons	Library Amino Acids	Diversity
20	GCT	A	GCT, AAN	A, N, K	3
21	GTC	V	GTC, CCG	V, P	2
39	GCA	A	GCA,AAA,CGT	A, N, R	3
40	ACT	T	AWT	T, N	2
44	GCG	A	GCG, AAC	A, N	2
51	GCC	A	GMG	A, E	2
52	GCT	A	GCT, AAC	A, N	2
67	ACG	T	RCG	T, A	2
71	TTC	F	TTC, GAT	F, D	2
78	GCC	A	GSC	A, G	2
94	AAC	N	ARC	N, S	2
95	GAT	D	GAT, ATG	D, M	2
97	ACT	T	RCT	T, A	2
99	CTG	L	CTG, GAT	L, D	2
104	CAG	Q	SAG	Q, E	2
105	GCC	A	GCC, AAC	A, N	2
124	GTG	V	GTG, CCG	V, P	2
152	AAT	N	RAT	N, D	2

Total Diversity	5.90E+05

Supplemental Table S4.6 Primers for construction of EVCouplings-informed library.

Primer Name	DNA
EV F1	ATGGTTGAGATTATTAATAAAACTGTTACTCGTGGTGTTGCTGGTCGTCGCGGA
EV F2-1	GGTCGTCGCGGAGGGGCTGTCAAGGGTGTCGTT
EV F2-2	GGTCGTCGCGGAGGGRAMGTCAAGGGTGTCGTT
EV F3-1	GTCAAGGGTGTCGTTNTTCATAATACTTGGGGTAATTCAASCGYGAAACAAGAAGCAA MTCGCCTTGCGGCGATGAATAAT
EV F3-2	GTCAAGGGTGTCGTTNTTCATAATACTGCAGGTAATTCAASCGYGAAACAAGAAGCAA MTCGCCTTGCGGCGATGAATAAT
EV F3-3	GTCAAGGGTGTCGTTNTTCATAATACTTGGGGTAATTCAASCGYGAAACAATATGCAAM TCGCCTTGCGGCGATGAATAAT
EV F3-4	GTCAAGGGTGTCGTTNTTCATAATACTGCAGGTAATTCAASCGYGAAACAAGAAGCAA MTCGCCTTGCGGCGATGAATAAT
EV F4-1	CTTGCGGCGATGAATAATAAMCAGCTGGMGGCTGGCTTCGCGCACTATTATATTGAC
EV F4-2	CTTGCGGCGATGAATAATGCCCAGCTGGMGGCTGGCTTCGCGCACTATTATATTGAC
EV F5	TCGCGCACTATTATATTGACRAAAATACCATTTGGCGC
EV F6-1	AATACCATTTGGCGCACCGAAGACACGTWTAATGGTGCATGGCACACTGCCAAT
EV F6-2	AATACCATTTGGCGCGTGGAAGACACGTWTAATGGTGCATGGCACACTGCCAAT
EV F6-3	AATACCATTTGGCGCACCGAAGACACGTWTAATATTGCATGGCACACTGCCAAT
EV F6-4	AATACCATTTGGCGCGTGGAAGACACGTWTAATATTGCATGGCACACTGCCAAT
EV F7-1	TGGCACACTGCCAATGSGGATGGTAATATGAACTATATCGGATATGAAGTCTGTGGCAA C
EV F7-2	TGGCACACTGCCAATCSTGATGGTAATATGAACTATATCGGATATGAAGTCTGTGGCAA C
EV F8-1	TATGAAGTCTGTGGCAACKWTCAGACTCCCCTGAAGGACTTTTTGCAGGCCGAGGAGA AC
EV F8-2	TATGAAGTCTGTGGCAACKWTCAGACTCCCCTGGCGGACTTTTTGCAGGCCGAGGAGA AC
EV F9-1	CAGGCCGAGGAGAACRCATTTTGGCAAATTGCGCAAGATCTGAAGTATTACGGTTTGCC T
EV F9-2	CAGGCCGAGGAGAACRCATTTTGGCAAGYGGCGCAAGATCTGAAGTATTACGGTTTGCC T
EV F10	AAGTATTACGGTTTGCCTGTGAATCGTAATACAGTTCGCCTGCACCATGAATTCAGCGCT

EV F11	CATGAATTCAGCGCTACTCAGTGCCCAAAACGTTCCCTTATCATTCATACTGGGTTTAAT
EV F12	CTTATCATTCATACTGGGTTTAATTCCACCCAGGCTCAACCTGCCAACGTGACAAACGCC
EV F13	AACGTGACAAACGCCATGAAGGACTACGTGATCAAGAACGTCTTGAAGTACTACAACAA C
EV F14	GTCTTGAAGTACTACAACAACCCCAGTTTGAAACCAGATGGATCCGAACAAAAGCTTATT
EV R1	TCAGAAATAAGCTTTTGTTCGGATCC

Supplemental Table S4.7 Primers for construction of PROSS-informed library.

Primer Name	DNA
PROSS F1	ATGGTTGAGATTATTAATAAAACTGTTACTCGTGGTGTTGCTGGTCGTCGC
PROSS F2-1	GCTGGTCGTCGCGGAGGGGCTGTCAAGGGTGTCGTTTTTCATAATACTTGG
PROSS F2-2	GCTGGTCGTCGCGGAGGGAANGTCAAGGGTGTCGTTTTTCATAATACTTGG
PROSS F2-3	GCTGGTCGTCGCGGAGGGGCTCCGAAGGGTGTCGTTTTTCATAATACTTGG
PROSS F2-4	GCTGGTCGTCGCGGAGGGAANCCGAAGGGTGTCGTTTTTCATAATACTTGG
PROSS F3	GTCGTTTTTCATAATACTTGGGGTAATTCATCTGTGAAACAA
PROSS F4-1	GGTAATTCATCTGTGAAACAAGAAGCAAWTCGCCTTGCGGCGATGAATAATAACCAGC TG
PROSS F4-2	GGTAATTCATCTGTGAAACAAGAAAAAAWTCGCCTTGCGGCGATGAATAATAACCAGC TG
PROSS F4-3	GGTAATTCATCTGTGAAACAAGAACGTAWTCGCCTTGCGGCGATGAATAATAACCAGC TG
PROSS F4-4	GGTAATTCATCTGTGAAACAAGAAGCAAWTCGCCTTGCGAACATGAATAATAACCAGCT G
PROSS F4-5	GGTAATTCATCTGTGAAACAAGAAAAAAWTCGCCTTGCGAACATGAATAATAACCAGCT G
PROSS F4-6	GGTAATTCATCTGTGAAACAAGAACGTAWTCGCCTTGCGAACATGAATAATAACCAGCT G
PROSS F5-1	ATGAATAATAACCAGCTGGMGGCTGGCTTCGCGCACTATTATATTGACAAGAATACCAT T
PROSS F5-2	ATGAATAATAACCAGCTGGMGAACGGCTTCGCGCACTATTATATTGACAAGAATACCAT T
PROSS F6-1	TATTATATTGACAAGAATACCATTTGGCGCRCGGAAGACACGTTCAATGGTGCATGGCA C
PROSS F6-2	TATTATATTGACAAGAATACCATTTGGCGCRCGGAAGACACGGATAATGGTGCATGGCA C

PROSS F7	GGTGCATGGCACACTGSCAATGGCGATGGTAATATGAACTATATCGGATATGAAGTCTGT
PROSS F8-1	ATCGGATATGAAGTCTGTGGCARCGATCAGRCTCCCCTGAAGGACTTTTTGSAGGCCGAGGAGAACACATTTTGGCAAATT
PROSS F8-2	ATCGGATATGAAGTCTGTGGCARCATGCAGRCTCCCCTGAAGGACTTTTTGSAGGCCGAGGAGAACACATTTTGGCAAATT
PROSS F8-3	ATCGGATATGAAGTCTGTGGCARCGATCAGRCTCCCGATAAGGACTTTTTGSAGGCCGAGGAGAACACATTTTGGCAAATT
PROSS F8-4	ATCGGATATGAAGTCTGTGGCARCATGCAGRCTCCCGATAAGGACTTTTTGSAGGCCGAGGAGAACACATTTTGGCAAATT
PROSS F8-5	ATCGGATATGAAGTCTGTGGCARCGATCAGRCTCCCCTGAAGGACTTTTTGSAGAACGAGGAGAACACATTTTGGCAAATT
PROSS F8-6	ATCGGATATGAAGTCTGTGGCARCATGCAGRCTCCCCTGAAGGACTTTTTGSAGAACGAGGAGAACACATTTTGGCAAATT
PROSS F8-7	ATCGGATATGAAGTCTGTGGCARCGATCAGRCTCCCGATAAGGACTTTTTGSAGAACGAGGAGAACACATTTTGGCAAATT
PROSS F8-8	ATCGGATATGAAGTCTGTGGCARCATGCAGRCTCCCGATAAGGACTTTTTGSAGAACGAGGAGAACACATTTTGGCAAATT
PROSS F9	AACACATTTTGGCAAATTGCGCAAGATCTGAAGTATTACGGTTTG
PROSS F10-1	CTGAAGTATTACGGTTTGCCTGTGAATCGTAATACAGTTCGCCTGCACCATGAATTCAGC
PROSS F10-2	CTGAAGTATTACGGTTTGCCTCCGAATCGTAATACAGTTCGCCTGCACCATGAATTCAGC
PROSS F11	GCACCATGAATTCAGCGCTACTCAGTGCCCAAAACGTTCCCTTATCATTCATACTGGG
PROSS F12-1	TCCCTTATCATTCATACTGGGTTTRATTCCACCCAGGCGCAACCTGCCAACGTGACA
PROSS F12-2	TCCCTTATCATTCATACTGGGTTTRATTCCACCCAGCGCCAACCTGCCAACGTGACA
PROSS F13-1	CCTGCCAACGTGACAAACGCCATGAAGGACTACGTGATCAAGAAC
PROSS F13-2	CCTGCCAACGTGACAAACAAAATGAAGGACTACGTGATCAAGAAC
PROSS F14-1	GACTACGTGATCAAGAACGTCTTGAAGTACTACAACAACCCCAGTTTGAAACCAGATGGA
PROSS F14-2	GACTACGTGATCAAGAACGTCAAGAAGTACTACAACAACCCCAGTTTGAAACCAGATGGA
PROSS R1	TCAGAAATAAGCTTTTGTTCGGATCCATCTGGTTTCAAACT

Supplemental Table S4.8 Primers for construction of Triplet library via linear triple "tiles"

Primer	DNA Sequence
Tile 1 F1	AGTATGGTTGAGATTATTNNKNNKNNKGTTACTCGTGGTGTTGCTGGTCGTCGCGGAGGG
Tile 1 F2	AGTATGGTTGAGATTATTAATNNKNNKNNKACTCGTGGTGTTGCTGGTCGTCGCGGAGGG
Tile 1 F3	AGTATGGTTGAGATTATTAATAAANNKNNKNNKCGTGGTGTTGCTGGTCGTCGCGGAGGG
Tile 1 F4	AGTATGGTTGAGATTATTAATAAAACTNNKNNKNNKGGTGTTGCTGGTCGTCGCGGAGGG
Tile 1 F5	AGTATGGTTGAGATTATTAATAAAACTGTTNNKNNKNNKGTTGCTGGTCGTCGCGGAGGG
Tile 1 F6	AGTATGGTTGAGATTATTAATAAAACTGTTACTNNKNNKNNKGCTGGTCGTCGCGGAGGG
Tile 1 R1	AATAATCTCAACCATACTAGCCTGCAG
Tile 2 F1	AATAAAACTGTTACTCGTNNKNNKNNKGGTCGTCGCGGAGGGGCTGTCAAGGGTGTCGTT
Tile 2 F2	AATAAAACTGTTACTCGTGGTNNKNNKNNKCGTCGCGGAGGGGCTGTCAAGGGTGTCGTT
Tile 2 F3	AATAAAACTGTTACTCGTGGTGTTNNKNNKNNKCGCGGAGGGGCTGTCAAGGGTGTCGTT
Tile 2 F4	AATAAAACTGTTACTCGTGGTGTTGCTNNKNNKNNKGGAGGGGCTGTCAAGGGTGTCGTT
Tile 2 F5	AATAAAACTGTTACTCGTGGTGTTGCTGGTNNKNNKNNKGGGGCTGTCAAGGGTGTCGTT
Tile 2 F6	AATAAAACTGTTACTCGTGGTGTTGCTGGTCGTNNKNNKNNKGCTGTCAAGGGTGTCGTT
Tile 2 R1	ACGAGTAACAGTTTTATTAATAATCTCAACC
Tile 3 F1	GGTGTTGCTGGTCGTCGCNNKNNKNNKGTCAAGGGTGTCGTTTTTCATAATACTTGGGGT
Tile 3 F2	GGTGTTGCTGGTCGTCGCGGANNKNNKNNKAAGGGTGTCGTTTTTCATAATACTTGGGGT
Tile 3 F3	GGTGTTGCTGGTCGTCGCGGAGGGNNKNNKNNKGGTGTCGTTTTTCATAATACTTGGGGT
Tile 3 F4	GGTGTTGCTGGTCGTCGCGGAGGGGCTNNKNNKNNKGTCGTTTTTCATAATACTTGGGGT
Tile 3 F5	GGTGTTGCTGGTCGTCGCGGAGGGGCTGTCNNKNNKNNKGTTTTTCATAATACTTGGGGT
Tile 3 F6	GGTGTTGCTGGTCGTCGCGGAGGGGCTGTCAAGNNKNNKNNKTTTCATAATACTTGGGGT
Tile 3 R1	GCGACGACCAGCAAC
Tile 4 F1	GGAGGGGCTGTCAAGGGTNNKNNKNNKCATAATACTTGGGGTAATTCATCTGTGAAACAA
Tile 4 F2	GGAGGGGCTGTCAAGGGTGTCNNKNNKNNKAATACTTGGGGTAATTCATCTGTGAAACAA

Tile 4 F3	GGAGGGGCTGTCAAGGGTGTCGTTNNKNNKNNKACTTGGGGTAATTCATCTGTGAAACAA
Tile 4 F4	GGAGGGGCTGTCAAGGGTGTCGTTTTTNNKNNKNNKTGGGGTAATTCATCTGTGAAACAA
Tile 4 F5	GGAGGGGCTGTCAAGGGTGTCGTTTTTCATNNKNNKNNKGGTAATTCATCTGTGAAACAA
Tile 4 F6	GGAGGGGCTGTCAAGGGTGTCGTTTTTCATAATNNKNNKNNKAATTCATCTGTGAAACAA
Tile 4 R1	ACCCTTGACAGCCCC
Tile 5 F1	GTCGTTTTTCATAATACTNNKNNKNNKTCATCTGTGAAACAAGAAGCAACTCGCCTTGCG
Tile 5 F2	GTCGTTTTTCATAATACTTGGNNKNNKNNKTCTGTGAAACAAGAAGCAACTCGCCTTGCG
Tile 5 F3	GTCGTTTTTCATAATACTTGGGGTNNKNNKNNKGTGAAACAAGAAGCAACTCGCCTTGCG
Tile 5 F4	GTCGTTTTTCATAATACTTGGGGTAATNNKNNKNNKAAACAAGAAGCAACTCGCCTTGCG
Tile 5 F5	GTCGTTTTTCATAATACTTGGGGTAATTCANNKNNKNNKCAAGAAGCAACTCGCCTTGCG
Tile 5 F6	GTCGTTTTTCATAATACTTGGGGTAATTCATCTNNKNNKNNKGAAGCAACTCGCCTTGCG
Tile 5 R1	AGTATTATGAAAAACGACACCCTTGAC
Tile 6 F1	TGGGGTAATTCATCTGTGNNKNNKNNKGCAACTCGCCTTGCGGCGATGAATAATAACCAG
Tile 6 F2	TGGGGTAATTCATCTGTGAAANNKNNKNNKACTCGCCTTGCGGCGATGAATAATAACCAG
Tile 6 F3	TGGGGTAATTCATCTGTGAAACAANNKNNKNNKCGCCTTGCGGCGATGAATAATAACCAG
Tile 6 F4	TGGGGTAATTCATCTGTGAAACAAGAANNKNNKNNKCTTGCGGCGATGAATAATAACCAG
Tile 6 F5	TGGGGTAATTCATCTGTGAAACAAGAAGCANNKNNKNNKGCGGCGATGAATAATAACCAG
Tile 6 F6	TGGGGTAATTCATCTGTGAAACAAGAAGCAACTNNKNNKNNKGCGATGAATAATAACCAG
Tile 6 R1	CACAGATGAATTACCCCA
Tile 7 F1	AAACAAGAAGCAACTCGCNNKNNKNNKATGAATAATAACCAGCTGGCCGCTGGCTTCGCG
Tile 7 F2	AAACAAGAAGCAACTCGCCTTNNKNNKNNKAATAATAACCAGCTGGCCGCTGGCTTCGCG
Tile 7 F3	AAACAAGAAGCAACTCGCCTTGCGNNKNNKNNKAATAACCAGCTGGCCGCTGGCTTCGCG
Tile 7 F4	AAACAAGAAGCAACTCGCCTTGCGGCGNNKNNKNNKAACCAGCTGGCCGCTGGCTTCGCG
Tile 7 F5	AAACAAGAAGCAACTCGCCTTGCGGCGATGNNKNNKNNKCAGCTGGCCGCTGGCTTCGCG
Tile 7 F6	AAACAAGAAGCAACTCGCCTTGCGGCGATGAATNNKNNKNNKCTGGCCGCTGGCTTCGCG
Tile 7 R1	GCGAGTTGCTTCTTGTTTCACAGATGAATTACC

Tile 8 F1	CTTGCGGCGATGAATAATNNKNNKNNKGCCGCTGGCTTCGCGCACTATTATATTGACAAG
Tile 8 F2	CTTGCGGCGATGAATAATAACNNKNNKNNKGCTGGCTTCGCGCACTATTATATTGACAAG
Tile 8 F3	CTTGCGGCGATGAATAATAACCAGNNKNNKNNKGGCTTCGCGCACTATTATATTGACAAG
Tile 8 F4	CTTGCGGCGATGAATAATAACCAGCTGNNKNNKNNKTTCGCGCACTATTATATTGACAAG
Tile 8 F5	CTTGCGGCGATGAATAATAACCAGCTGGCCNNKNNKNNKGCGCACTATTATATTGACAAG
Tile 8 F6	CTTGCGGCGATGAATAATAACCAGCTGGCCGCTNNKNNKNNKCACTATTATATTGACAAG
Tile 8 R1	ATTATTCATCGCCGC
Tile 9 F1	AACCAGCTGGCCGCTGGCNNKNNKNNKTATTATATTGACAAGAATACCATTTGGCGCACG
Tile 9 F2	AACCAGCTGGCCGCTGGCTTCNNKNNKNNKTATATTGACAAGAATACCATTTGGCGCACG
Tile 9 F3	AACCAGCTGGCCGCTGGCTTCGCGNNKNNKNNKATTGACAAGAATACCATTTGGCGCACG
Tile 9 F4	AACCAGCTGGCCGCTGGCTTCGCGCACNNKNNKNNKGACAAGAATACCATTTGGCGCACG
Tile 9 F5	AACCAGCTGGCCGCTGGCTTCGCGCACTATNNKNNKNNKAAGAATACCATTTGGCGCACG
Tile 9 F6	AACCAGCTGGCCGCTGGCTTCGCGCACTATTATNNKNNKNNKAATACCATTTGGCGCACG
Tile 9 R1	GCCAGCGGCCAGCTG
Tile 10 F1	TTCGCGCACTATTATATTNNKNNKNNKACCATTTGGCGCACGGAAGACACGTTCAATGGT
Tile 10 F2	TTCGCGCACTATTATATTGACNNKNNKNNKATTTGGCGCACGGAAGACACGTTCAATGGT
Tile 10 F3	TTCGCGCACTATTATATTGACAAGNNKNNKNNKTGGCGCACGGAAGACACGTTCAATGGT
Tile 10 F4	TTCGCGCACTATTATATTGACAAGAATNNKNNKNNKCGCACGGAAGACACGTTCAATGGT
Tile 10 F5	TTCGCGCACTATTATATTGACAAGAATACCNNKNNKNNKACGGAAGACACGTTCAATGGT
Tile 10 F6	TTCGCGCACTATTATATTGACAAGAATACCATTNNKNNKNNKGAAGACACGTTCAATGGT
Tile 10 R1	AATATAATAGTGCGCGAA
Tile 11 F1	GACAAGAATACCATTTGGNNKNNKNNKGACACGTTCAATGGTGCATGGCACACTGCCAAT
Tile 11 F2	GACAAGAATACCATTTGGCGCNNKNNKNNKACGTTCAATGGTGCATGGCACACTGCCAAT
Tile 11 F3	GACAAGAATACCATTTGGCGCACGNNKNNKNNKTTCAATGGTGCATGGCACACTGCCAAT

Tile 11 F4	GACAAGAATACCATTTGGCGCACGGAANNKNNKNNKAATGGTGCATGGCACACTGCCAAT
Tile 11 F5	GACAAGAATACCATTTGGCGCACGGAAGACNNKNNKNNKGGTGCATGGCACACTGCCAAT
Tile 11 F6	GACAAGAATACCATTTGGCGCACGGAAGACACGNNKNNKNNKGCATGGCACACTGCCAAT
Tile 11 R1	CCAAATGGTATTCTTGTCAATATAATAGTGC
Tile 12 F1	CGCACGGAAGACACGTTCNNKNNKNNKTGGCACACTGCCAATGGCGATGGTAATATGAAC
Tile 12 F2	CGCACGGAAGACACGTTCAATNNKNNKNNKCACACTGCCAATGGCGATGGTAATATGAAC
Tile 12 F3	CGCACGGAAGACACGTTCAATGGTNNKNNKNNKACTGCCAATGGCGATGGTAATATGAAC
Tile 12 F4	CGCACGGAAGACACGTTCAATGGTGCANNKNNKNNKGCCAATGGCGATGGTAATATGAAC
Tile 12 F5	CGCACGGAAGACACGTTCAATGGTGCATGGNNKNNKNNKAATGGCGATGGTAATATGAAC
Tile 12 F6	CGCACGGAAGACACGTTCAATGGTGCATGGCACNNKNNKNNKGGCGATGGTAATATGAAC
Tile 12 R1	GAACGTGTCTTCCGTG
Tile 13 F1	AATGGTGCATGGCACACTNNKNNKNNKGATGGTAATATGAACTATATCGGATATGAAGTC
Tile 13 F2	AATGGTGCATGGCACACTGCCNNKNNKNNKGGTAATATGAACTATATCGGATATGAAGTC
Tile 13 F3	AATGGTGCATGGCACACTGCCAATNNKNNKNNKAATATGAACTATATCGGATATGAAGTC
Tile 13 F4	AATGGTGCATGGCACACTGCCAATGGCNNKNNKNNKATGAACTATATCGGATATGAAGTC
Tile 13 F5	AATGGTGCATGGCACACTGCCAATGGCGATNNKNNKNNKAACTATATCGGATATGAAGTC
Tile 13 F6	AATGGTGCATGGCACACTGCCAATGGCGATGGTNNKNNKNNKTATATCGGATATGAAGTC
Tile 13 R1	AGTGTGCCATGCACC
Tile 14 F1	GCCAATGGCGATGGTAATNNKNNKNNKATCGGATATGAAGTCTGTGGCAACGATCAGACT
Tile 14 F2	GCCAATGGCGATGGTAATATGNNKNNKNNKGGATATGAAGTCTGTGGCAACGATCAGACT
Tile 14 F3	GCCAATGGCGATGGTAATATGAACNNKNNKNNKTATGAAGTCTGTGGCAACGATCAGACT
Tile 14 F4	GCCAATGGCGATGGTAATATGAACTATNNKNNKNNKGAAGTCTGTGGCAACGATCAGACT
Tile 14 F5	GCCAATGGCGATGGTAATATGAACTATATCNNKNNKNNKGTCTGTGGCAACGATCAGACT

Tile 14 F6	GCCAATGGCGATGGTAATATGAACTATATCGGANNKNNKNNKTGTGGCAACGATCAGACT
Tile 14 R1	ATTACCATCGCCATTGGC
Tile 15 F1	ATGAACTATATCGGATATNNKNNKNNKGGCAACGATCAGACTCCCCTGAAGGACTTTTTG
Tile 15 F2	ATGAACTATATCGGATATGAANNKNNKNNKAACGATCAGACTCCCCTGAAGGACTTTTTG
Tile 15 F3	ATGAACTATATCGGATATGAAGTCNNKNNKNNKGATCAGACTCCCCTGAAGGACTTTTTG
Tile 15 F4	ATGAACTATATCGGATATGAAGTCTGTNNKNNKNNKCAGACTCCCCTGAAGGACTTTTTG
Tile 15 F5	ATGAACTATATCGGATATGAAGTCTGTGGCNNKNNKNNKACTCCCCTGAAGGACTTTTTG
Tile 15 F6	ATGAACTATATCGGATATGAAGTCTGTGGCAACNNKNNKNNKCCCCTGAAGGACTTTTTG
Tile 15 R1	ATATCCGATATAGTTCATATTACCATC
Tile 16 F1	GAAGTCTGTGGCAACGATNNKNNKNNKCTGAAGGACTTTTTGCAGGCCGAGGAGAACACA
Tile 16 F2	GAAGTCTGTGGCAACGATCAGNNKNNKNNKAAGGACTTTTTGCAGGCCGAGGAGAACACA
Tile 16 F3	GAAGTCTGTGGCAACGATCAGACTNNKNNKNNKGACTTTTTGCAGGCCGAGGAGAACACA
Tile 16 F4	GAAGTCTGTGGCAACGATCAGACTCCCNNKNNKNNKTTTTTGCAGGCCGAGGAGAACACA
Tile 16 F5	GAAGTCTGTGGCAACGATCAGACTCCCCTGNNKNNKNNKTTGCAGGCCGAGGAGAACACA
Tile 16 F6	GAAGTCTGTGGCAACGATCAGACTCCCCTGAAGNNKNNKNNKCAGGCCGAGGAGAACACA
Tile 16 R1	ATCGTTGCCACAGACTTC
Tile 17 F1	CAGACTCCCCTGAAGGACNNKNNKNNKGCCGAGGAGAACACATTTTGGCAAATTGCGCAA
Tile 17 F2	CAGACTCCCCTGAAGGACTTTNNKNNKNNKGAGGAGAACACATTTTGGCAAATTGCGCAA
Tile 17 F3	CAGACTCCCCTGAAGGACTTTTTGNNKNNKNNKGAGAACACATTTTGGCAAATTGCGCAA
Tile 17 F4	CAGACTCCCCTGAAGGACTTTTTGCAGNNKNNKNNKAACACATTTTGGCAAATTGCGCAA
Tile 17 F5	CAGACTCCCCTGAAGGACTTTTTGCAGGCCNNKNNKNNKACATTTTGGCAAATTGCGCAA
Tile 17 F6	CAGACTCCCCTGAAGGACTTTTTGCAGGCCGAGNNKNNKNNKTTTTGGCAAATTGCGCAA
Tile 17 R1	GTCCTTCAGGGGAGTCTG

Tile 18 F1	TTTTTGCAGGCCGAGGAGNNKNNKNNKTGGCAAATTGCGCAAGATCTGAAGTATTACG GT
Tile 18 F2	TTTTTGCAGGCCGAGGAGAACNNKNNKNNKCAAATTGCGCAAGATCTGAAGTATTACG GT
Tile 18 F3	TTTTTGCAGGCCGAGGAGAACACANNKNNKNNKATTGCGCAAGATCTGAAGTATTACG GT
Tile 18 F4	TTTTTGCAGGCCGAGGAGAACACATTTNNKNNKNNKGCGCAAGATCTGAAGTATTACG GT
Tile 18 F5	TTTTTGCAGGCCGAGGAGAACACATTTTGGNNKNNKNNKCAAGATCTGAAGTATTACG GT
Tile 18 F6	TTTTTGCAGGCCGAGGAGAACACATTTTGGCAANNKNNKNNKGATCTGAAGTATTACG GT
Tile 18 R1	CTCCTCGGCCTGCAA
Tile 19 F1	AACACATTTTGGCAAATTNNKNNKNNKCTGAAGTATTACGGTTTGCCTGTGAATCGTAA T
Tile 19 F2	AACACATTTTGGCAAATTGCGNNKNNKNNKAAGTATTACGGTTTGCCTGTGAATCGTAA T
Tile 19 F3	AACACATTTTGGCAAATTGCGCAANNKNNKNNKTATTACGGTTTGCCTGTGAATCGTAA T
Tile 19 F4	AACACATTTTGGCAAATTGCGCAAGATNNKNNKNNKTACGGTTTGCCTGTGAATCGTAA T
Tile 19 F5	AACACATTTTGGCAAATTGCGCAAGATCTGNNKNNKNNKGGTTTGCCTGTGAATCGTAA T
Tile 19 F6	AACACATTTTGGCAAATTGCGCAAGATCTGAAGNNKNNKNNKTTGCCTGTGAATCGTAA T
Tile 19 R1	AATTTGCCAAAATGTGTT
Tile 20 F1	GCGCAAGATCTGAAGTATNNKNNKNNKCCTGTGAATCGTAATACAGTTCGCCTGCACCA T
Tile 20 F2	GCGCAAGATCTGAAGTATTACNNKNNKNNKGTGAATCGTAATACAGTTCGCCTGCACCA T
Tile 20 F3	GCGCAAGATCTGAAGTATTACGGTNNKNNKNNKAATCGTAATACAGTTCGCCTGCACCA T
Tile 20 F4	GCGCAAGATCTGAAGTATTACGGTTTGNNKNNKNNKCGTAATACAGTTCGCCTGCACCA T
Tile 20 F5	GCGCAAGATCTGAAGTATTACGGTTTGCCTNNKNNKNNKAATACAGTTCGCCTGCACCA T
Tile 20 F6	GCGCAAGATCTGAAGTATTACGGTTTGCCTGTGNNKNNKNNKACAGTTCGCCTGCACCA T
Tile 20 R1	ATACTTCAGATCTTGCGCAATTTG
Tile 21 F1	TACGGTTTGCCTGTGAATNNKNNKNNKGTTCGCCTGCACCATGAATTCAGCGCTACTCA G
Tile 21 F2	TACGGTTTGCCTGTGAATCGTNNKNNKNNKCGCCTGCACCATGAATTCAGCGCTACTCA G

Tile 21 F3	TACGGTTTGCCTGTGAATCGTAATNNKNNKNNKCTGCACCATGAATTCAGCGCTACTCA G
Tile 21 F4	TACGGTTTGCCTGTGAATCGTAATACANNKNNKNNKCACCATGAATTCAGCGCTACTCA G
Tile 21 F5	TACGGTTTGCCTGTGAATCGTAATACAGTTNNKNNKNNKCATGAATTCAGCGCTACTCA G
Tile 21 F6	TACGGTTTGCCTGTGAATCGTAATACAGTTCGCNNKNNKNNKGAATTCAGCGCTACTCA G
Tile 21 R1	ATTCACAGGCAAACCGTA
Tile 22 F1	CGTAATACAGTTCGCCTGNNKNNKNNKTTCAGCGCTACTCAGTGCCCAAAACGTTCCCTT
Tile 22 F2	CGTAATACAGTTCGCCTGCACNNKNNKNNKAGCGCTACTCAGTGCCCAAAACGTTCCCT T
Tile 22 F3	CGTAATACAGTTCGCCTGCACCATNNKNNKNNKGCTACTCAGTGCCCAAAACGTTCCCTT
Tile 22 F4	CGTAATACAGTTCGCCTGCACCATGAANNKNNKNNKACTCAGTGCCCAAAACGTTCCCT T
Tile 22 F5	CGTAATACAGTTCGCCTGCACCATGAATTCNNKNNKNNKCAGTGCCCAAAACGTTCCCTT
Tile 22 F6	CGTAATACAGTTCGCCTGCACCATGAATTCAGCNNKNNKNNKTGCCCAAAACGTTCCCTT
Tile 22 R1	CAGGCGAACTGTATTACG
Tile 23 F1	CACCATGAATTCAGCGCTNNKNNKNNKCCAAAACGTTCCCTTATCATTCATACTGGGTTT
Tile 23 F2	CACCATGAATTCAGCGCTACTNNKNNKNNKAAACGTTCCCTTATCATTCATACTGGGTTT
Tile 23 F3	CACCATGAATTCAGCGCTACTCAGNNKNNKNNKCGTTCCCTTATCATTCATACTGGGTTT
Tile 23 F4	CACCATGAATTCAGCGCTACTCAGTGCNNKNNKNNKTCCCTTATCATTCATACTGGGTTT
Tile 23 F5	CACCATGAATTCAGCGCTACTCAGTGCCCANNKNNKNNKCTTATCATTCATACTGGGTTT
Tile 23 F6	CACCATGAATTCAGCGCTACTCAGTGCCCAAAANNKNNKNNKATCATTCATACTGGGTT T
Tile 23 R1	AGCGCTGAATTCATG
Tile 24 F1	ACTCAGTGCCCAAAACGTNNKNNKNNKATTCATACTGGGTTTAATTCCACCCAGGCTCA A
Tile 24 F2	ACTCAGTGCCCAAAACGTTCCNNKNNKNNKCATACTGGGTTTAATTCCACCCAGGCTCA A
Tile 24 F3	ACTCAGTGCCCAAAACGTTCCCTTNNKNNKNNKACTGGGTTTAATTCCACCCAGGCTCAA
Tile 24 F4	ACTCAGTGCCCAAAACGTTCCCTTATCNNKNNKNNKGGGTTTAATTCCACCCAGGCTCAA

Tile 24 F5	ACTCAGTGCCCAAAACGTTCCCTTATCATTNNKNNKNNKTTTAATTCCACCCAGGCTCAA
Tile 24 F6	ACTCAGTGCCCAAAACGTTCCCTTATCATTCATNNKNNKNNKAATTCCACCCAGGCTCAA
Tile 24 R1	ACGTTTTGGGCACTGAGT
Tile 25 F1	TCCCTTATCATTCATACTNNKNNKNNKTCCACCCAGGCTCAACCTGCCAACGTGACAAAC
Tile 25 F2	TCCCTTATCATTCATACTGGGNNKNNKNNKACCCAGGCTCAACCTGCCAACGTGACAAA C
Tile 25 F3	TCCCTTATCATTCATACTGGGTTTNNKNNKNNKCAGGCTCAACCTGCCAACGTGACAAAC
Tile 25 F4	TCCCTTATCATTCATACTGGGTTTAATNNKNNKNNKGCTCAACCTGCCAACGTGACAAAC
Tile 25 F5	TCCCTTATCATTCATACTGGGTTTAATTCCNNKNNKNNKCAACCTGCCAACGTGACAAAC
Tile 25 F6	TCCCTTATCATTCATACTGGGTTTAATTCCACCNNKNNKNNKCCTGCCAACGTGACAAAC
Tile 25 R1	AGTATGAATGATAAGGGAACGTTTT
Tile 26 F1	GGGTTTAATTCCACCCAGNNKNNKNNKGCCAACGTGACAAACGCCATGAAGGACTACG TG
Tile 26 F2	GGGTTTAATTCCACCCAGGCTNNKNNKNNKAACGTGACAAACGCCATGAAGGACTACG TG
Tile 26 F3	GGGTTTAATTCCACCCAGGCTCAANNKNNKNNKGTGACAAACGCCATGAAGGACTACG TG
Tile 26 F4	GGGTTTAATTCCACCCAGGCTCAACCTNNKNNKNNKACAAACGCCATGAAGGACTACGT G
Tile 26 F5	GGGTTTAATTCCACCCAGGCTCAACCTGCCNNKNNKNNKAACGCCATGAAGGACTACGT G
Tile 26 F6	GGGTTTAATTCCACCCAGGCTCAACCTGCCAACNNKNNKNNKGCCATGAAGGACTACGT G
Tile 26 R1	CTGGGTGGAATTAAACCC
Tile 27 F1	GCTCAACCTGCCAACGTGNNKNNKNNKATGAAGGACTACGTGATCAAGAACGTCTTGA AG
Tile 27 F2	GCTCAACCTGCCAACGTGACANNKNNKNNKAAGGACTACGTGATCAAGAACGTCTTGA AG
Tile 27 F3	GCTCAACCTGCCAACGTGACAAACNNKNNKNNKGACTACGTGATCAAGAACGTCTTGAA G
Tile 27 F4	GCTCAACCTGCCAACGTGACAAACGCCNNKNNKNNKTACGTGATCAAGAACGTCTTGAA G
Tile 27 F5	GCTCAACCTGCCAACGTGACAAACGCCATGNNKNNKNNKGTGATCAAGAACGTCTTGA AG
Tile 27 F6	GCTCAACCTGCCAACGTGACAAACGCCATGAAGNNKNNKNNKATCAAGAACGTCTTGA AG

Tile 27 R1	CACGTTGGCAGGTTG
Tile 28 F1	ACAAACGCCATGAAGGACNNKNNKNNKAAGAACGTCTTGAAGTACTACAACAACCCCA GT
Tile 28 F2	ACAAACGCCATGAAGGACTACNNKNNKNNKAACGTCTTGAAGTACTACAACAACCCCA GT
Tile 28 F3	ACAAACGCCATGAAGGACTACGTGNNKNNKNNKGTCTTGAAGTACTACAACAACCCCA GT
Tile 28 F4	ACAAACGCCATGAAGGACTACGTGATCNNKNNKNNKTTGAAGTACTACAACAACCCCA GT
Tile 28 F5	ACAAACGCCATGAAGGACTACGTGATCAAGNNKNNKNNKAAGTACTACAACAACCCCA GT
Tile 28 F6	ACAAACGCCATGAAGGACTACGTGATCAAGAACNNKNNKNNKTACTACAACAACCCCA GT
Tile 28 R1	GTCCTTCATGGCGTTTGT
Tile 29 F1	TACGTGATCAAGAACGTCNNKNNKNNKTACAACAACCCCAGTTTGAAACCAGATGGATC C
Tile 29 F2	TACGTGATCAAGAACGTCTTGNNKNNKNNKAACAACCCCAGTTTGAAACCAGATGGATC C
Tile 29 F3	TACGTGATCAAGAACGTCTTGAAGNNKNNKNNKAACCCCAGTTTGAAACCAGATGGAT CC
Tile 29 F4	TACGTGATCAAGAACGTCTTGAAGTACNNKNNKNNKCCCAGTTTGAAACCAGATGGATC C
Tile 29 F5	TACGTGATCAAGAACGTCTTGAAGTACTACNNKNNKNNKAGTTTGAAACCAGATGGATC C
Tile 29 F6	TACGTGATCAAGAACGTCTTGAAGTACTACAACNNKNNKNNKTTGAAACCAGATGGATC C
Tile 29 R1	GACGTTCTTGATCACGTA

Supplemental Table S4.9 DNA sequences for full lysin catalytic or cell wall-binding domains.

Variant	DNA
LysEFm5-V7 Cat. domain[a]	ATGGTTGAGATTATTAATAAAACTGTTACTCGTGGTGTTGCTGGTCGTCGCGGAGGGGC TGTCAAGGGTGTCGTTTTTCATAATACTTGGGGTAATTCATCTGTGAAACAAGAAGCAAC TCGCCTTGCGGCGATGAATAATAACCAGCTGGCCGCTGGCTTCGCGCACTATTATATTGA CAAGAATACCATTTGGCGCACGGAAGACACGTTCAATGGTGCATGGCACACTGCCAATG GCGATGGTAATATGAACTATATCGGATATGAAGTCTGTGGCAACGATCAGACTCCCCTG AAGGACTTTTTGCAGGCCGAGGAGAACACATTTTGGCAAATTGCGCAAGATCTGAAGTA TTACGGTTTGCCTGTGAATCGTAATACAGTTCGCCTGCACCATGAATTCAGCGCTACTCA GTGCCCAAAACGTTCCCTTATCATTCATACTGGGTTTAATTCCACCCAGGCTCAACCTGCC AACGTGACAAACGCCATGAAGGACTACGTGATCAAGAACGTCTTGAAGTACTACAACAA CCCCAGTTTGAAACCAGAT
LysEFm5 CWBD[b]	GGTAAGGCGCCATCTACCAGTGGCCAAACGCCACCGTCTGGGGCGAACGTAACTCCCTC GACTCCGTCACAACATGATAAAGCGGTTGCTGCAACTAAACCCAAGCATCAGGGAAATG CATGGGGTAAACTGGACTATTTCAATGGGCACGGAAAGGATCAGATCCGCGTGGCCGG ATGGCTTGTTCCCGACAAGCCGCAAGGGAGCATTGGCAAATACGCCTATGTGATTTTTA TGCAACATGGGACGGGGAAAGAGTTAACGCGCGTCCAATCCGCTGGCATCAAGCGTCC TGATGTCAAAAAAGCATACGGTTATCAAGGGGGCCAAGAGCTTGGTTTCGATGTAACA GTAAAAAAAGCACAGTTCAAAGGCAAAAAAGTAGACGTTATTCTTCGTCGCGCCAACAA AGCCAATGGGGAGGGAGCGGTTAATGACGTCCGCATTGACTCCATCTATCTGAGCCTGG GATCCATGAC
GM1	AAGGAGATATACATATGGCTAGCATGGCAGGGAAACGCGCCGGCAGTGTTAAAGGGG TGGTTATCCACAACACATGGACCAACACTACTGCAGAACAAGAGATGAATCGTTTGGCA AATATGACGCCAAAACAATTGGAGGCAGGCTTTGCGCACTATTATGTAGACGAAAAGAC GATTATACGCACCGAGGATACTTATAATCGGGCGTGGCACGTGGCAAATAGTGACGGA AATAATAGTTATCTTGGATATGAAGTCTGCCAGTCTCTGGGGGCATCGGACAAGGACTT TTTGGCCAATGAGCAGGCGACATTTAAACAAGTCGCGGAGGATCTGAAGTTTTACGGTT TGAAGGCGAATCGTGATACAGTTCGCCTGCACCGCGAATTCGTGGCTACTGCGTGCCCA CATCGTTCCTGGGAATTACATGGCAAGTCAGTCAACAGTGTTAAGGATTACTTCATAGCC CAAATAAACAAATATCTTGGAGTGCCCAGTTTGAAACCAGAT
GM2	AAGGAGATATACATATGGCTAGCATGGCAGGGAAACGCGCCGGCAGTGTTAAAGGGG TGGTTATCCACAACACATGGACCAACACTACTGCAGAACAAGAGATGAATCGTTTGGCA AATATGACGCCAAAACAATTGGAGGCAGGCTTTGCGCACTATTATGTAGACGAAAAGAC GATTATACGCACCGAGGATACTTATAATCGGGCGTGGCACGTGGCAAATAGTGACGGA AATAATAGTTATCTTGGATATGAAGTCTGCCAGTCTCTTGGCGCTAGCGATAAGGACTTT TTGGCAAATGAGCAAGCAACATTTAAACAAGTAGCGGAAGATCTGAAGTTTTACGGTTT GAAAGCAAATCGTGACACAGTTCGCCTGCACCGGGAATTCGTGGCTACTGCGTGCCCAC ACCGTTCCTGGGAATTGCATGGGAAAAGTATCAATTCCGTGAAAGATTATTTTATTGCGC AGATCAATAAGTACATGGGTGTGCCCAGTTTGAAACCAGAT

GM3	AAGGAGATATACATATGGCTAGCATGAAGGTCATTAACAATGCGGTGTGCCGTGGAGT AGCCGGAAAACGTATTGGAAATGTTAAAGGCGTAGTCTTACATAACACATGGGATAATA AGTCAGCTAACTCTCATATAGAACGCTTGGGAAAGATGAACAATAAACAGTTGGAAGAA GGCTTTGCACACTACTATGTTGATGAAGAGACAATCGTACGTGTGGAGGATACCTTTAA CAAGGCATGGCACACTGCCAATGTTGAGGGCAATGCTTACTACATTGGATATGAAGTCT GCCAGTCTCTGGGGGCATCGGACAAGGACTTTTTGGCCAATGAGCAGGCGACATTTAAA CAAGTCGCGGAGGATCTGAAGTTTTACGGTTTGAAGGCGAATCGTGATACAGTTCGCCT GCACCGCGAATTCGTGGCTACTGCGTGCCCACATCGTTCCTGGGAATTACATGGCAAGT CAGTCAACAGTGTTAAGGATTACTTCATAGCCCAAATAAACAAATATCTTGGAGTGCCCA GTTTGAAACCAGAT
GM4	AAGGAGATATACATATGGCTAGCATGGTTAAGGTAATTAATAAATCAGCTTGTCGTGGT GTTGCTGGTAAACGCGCGGGGAATGTCAAGGGTGTCGTTATACATAATGATGCGGGTG CGGTAGGGGCCACAGCGGAACTGTATGTGAAGCGGCTTGAGGCTATGACCAACAAACA GCTTGAGAACGGTTTTGCGCATTACTATATTGATCGCAACACGGTCGCGCGTGTCGAGG ACACATATAACAAAGCCTGGCATACAGCCAACCAGGATGGTAACGCGAATTACATTGGA TATGAAGTCTGCCAGTCATTGGGCGCCTCTGACAAGGACTTTTTGGCGAATGAGCAAGC CACATTTAAACAAGTAGCGGAGGATCTGAAGTTTTACGGTTTGAAGGCGAATCGTGATA CAGTTCGCCTGCACCGCGAATTCGTGGCTACTGCGTGCCCACATCGTTCCTGGGAATTAC ATGGCAAGTCAGTCAACAGTGTTAAGGATTACTTCATAGCCCAAATAAACAAATATCTTG GAGTGCCCAGTTTGAAACCAGAT
GM5	AAGGAGATATACATATGGCTAGCATGGTGGAGATTATTGTTAATTACGTCACCCGTGGT GTAGCCGGCCGGCGTTCTGGAGCCATTCAGGGAGCAGTCATTCATAATTCGTGGAGTTC CGCGACCGCCAAACAGGAGGCTGATCGCCTTGCACGTATGACTCCGGCTCAGTTAGAAG CCGGTTTTGCGCATGAGTATATAGACTCTAATACCGTGTATGTTACCGAAAATCATCTTA ACCGCGCTTGGCATGTGGCCAACTCCGTAGGTAACAATGGGTTTATTGGATATGAAGTC CGCGGCAACCGGGAGACTCCCAAGGCGGTATTTTTGCAGGCCGAGCAAAACGCTTTTTG GCAAGCAGCGGAGGATCTGCGTTTCTACGGTTTGCCTGTGAATCGTGATACAGTTAAAT GTCACCATCAATTCAGCGCTACTGAGTGCCCAAAACGTTCCCTTATGGAGCATTGCGGGT ATGATTCCACCCTTGCTGTCCCTGCCGCTATAACAGTGCAGATGCAAGACTACTTTATCTC ACAAATCAAGAAGTACTACGACAACCCCAGTTTGAAACCAGAT
DP1	ATGGTTGAGATTATTAATAAAACTGTTACTCGTGGTGTTGCTGGTCGTCGCGGAGGGGC TGTCAAGGGTGTCGTTTTTCATAATACTTGGGGTCATTCATCTGTGAAACAAGAAGCAAC TCGCCTTGCGGCGATGAATAAAAACCAGCTGGCCGCTGGCTTCGCGCACTATGATATTG ACACCAAAACCAACTGGCGCACGGAAGACACGGTGAATGGTGCATGGCACACTGCCAA TGGCGATGGTAATATGAACTATATCGGATATGAAGTCTGTGGCAACGATCAGACTCCCC TGAAGGACTTTTTGCAGGCCGAGGAGAACACATTTTGGCAAATTGCGCAAGATCTGAAG TATTACGGTTTGCCTGTGAATCGTAATACAGTTCGCCTGCACCATGAATTCAGCGCTACT CAGTGCCCAAAACGTTCCCTTATCATTCATACTGGGTTTAATTCCACCCAGGCTCAACCTG CCAACGTGACAAACGCCATGAAGGACTACGTGATCAAGAACGTCTTGAAGTACTACAAC AACCCCAGTTTGAAACCAGAT

DP2	ATGGTTGAGATTATTAATAAAACTGTTACTCGTGGTGTTGCTGGTCGTCGCGGAGGGGC TGTCAAGGGTGTCGTTTTTCCGAATACTGGCGGTCATTCATCTGTGAAACAAGAAGCAA CTCGCCTTGCGGCGATGAATAAAAACCAGCTGGCCGCTGGCTTCGCGCACTATGATATT GACACCAAAACCATTTGGCGCACGGAAGACACGGTGAATGGTGCATGGCACACTGCCA ATGGCGATGGTAATATGAACTATATCGGATATGAAGTCTGTGGCAACGATCAGACTCCC CTGAAGGACTTTTTGCAGGCCGAGGAGAACACATTTTGGCAAATTGCGCAAGATCTGAA GTATTACGGTTTGCCTGTGAATCGTAATACAGTTCGCCTGCACCATGAATTCAGCGCTAC TCAGTGCCCAAAACGTTCCCTTATCATTCATACTGGGTTTAATTCCACCCAGGCTCAACCT GCCAACGTGACAAACGCCATGAAGGACTACGTGATCAAGAACGTCTTGAAGTACTACAA CAACCCCAGTTTGAAACCAGAT
RP1	ATGGTTGAGATTATTAATAAAACTGTTACTCGTGGTGTTGCTGGTCGTCTGGGAGGGGC TGTCAAGGGTGTCGTTTTTCATAATACTTGGGGTAATTCATCTGCGAAACAAGAAGCAA CTCGCCTTGCGGCGATGAATAATCGTCAGCTGGCCGCTGGCTTCGCGCACTATTATATTG ACAAGAATACCATTTGGCGCACGGAAGACACGTTCAATGGTGCATGGCACACTGCCAAT GGCGATGGTGATATGAACTATATCGGATATGAAGTCTATGGCAACGATCAGACTCCCCT GAAGGACTTTTTGCAGGCCGAGGAGAACACATTTTGGCAACCGGCGCAAGATCTGAAG TATTACGGTTTGCCTGTGAATCGTAATACAGTTCGCCTGCACCAGGAATTCAGCGCTACT CAGTGCCCAAAACGTTCCCTTATCATTCATACTGGGTTTAATTCCACCCAGGCTCAACCTG CCAACGTGACAAACGCCATGAAGGACTACGTGATCAAGAACGTCTTGAAGTACTACAAC AACCCCAGTTTGAAACCAGAT
RP2	ATGGTTGAGATTATTAATAAAACTGTTACTCGTGGTGTTGCTGGTCGTCTGGGAGGGGC TGTCAAGGGTGTCGTTTTTCATAATACTTGGGGTAATTCATCTGCGAAACAAGAAGCAA CTCGCCTTGCGGCGATGAATAATAACCAGCTGGCCGCTGGCTTCGCGCACTATTATATTG ACAAGAATACCATTTGGCGCACGGAAGACACGTTCAATGGTGCATGGCACACTGCCAAT GGCGATGGTGATATGAACTATATCGGATATGAAGTCTGTGGCAACGATCAGACTCCCCT GGAAGACTTTTTGCAGGCCGAGGAGAACACATTTTGGCAACCGGCGCAAGATCTGAAG TATTACGGTTTGAGCGTGAATCGTAATACAGTTCGCCTGCACCAGGAATTCAGCGCTACT CAGTGCCCAAAACGTTCCCTTATCATTCATACTGGGTTTAATTCCACCCAGGCTCAACCTG CCAACGTGACAAACGCCATGAAGGACTACGTGATCAAGAACGTCTTGAAGTACTACAAC AACCCCAGTTTGAAACCAGAT

[a] Sequence obtained from Baryakova *et al*.[40] [b] Sequence obtained from Gong *et al*.[132]

Supplemental Table S4.10 DNA primers for amplification and construction of lysin CD-CWBD genes into pET expression plasmid.

Primer	DNA Sequence
GMLibVars-Fwd	AAGGAGATATACATATGGCTAGC
GMLibVars-CD-Rev	ATGATGGTGATGGTGGGATCCATCTGGTTTCAAACTGGG
GMLibVars-CD-CW-Rev	CTGGTAGATGGCGCCTTACCATCTGGTTTCAAACTGGG
GMLibVars-CWBD-Fwd	CCCAGTTTGAAACCAGATGGTAAGGCGCCATCTA

GMLibVars-CWBD-Rev	GTGATGATGGTGATGGTG
LysV7 pET-Fwd	AGAAGGAGATATACATATGGCTAGCATGGTTGAGATTATTAATAAAACTGTTAC
LysV7-CWBD pET-Rev	TTATCAGTGATGATGGTGATGG
Model Pred Fwd	GCTCAACCTGCCAAC
Model Pred Rev	CGTTTGTCACGTTGGCA
Des Lib Control Fwd	AAGGGTGTCGTTTTTCATAATAACTGGGGTAATTCAAACGCGAAACAAGAAGCAGCGCGCCTTGCGGCGATGAATAAT
Des Lib Control Rev	ATTATGAAAAACGACACCCTT
EP Lib Control Fwd 1	ACTCGTGGTGTTGCTGGTCGTCTGGGAGGGGCTGTCAAGGGTGTC
EP Lib Control Fwd 2	ATTTGGCGCACGGAAGACACGCTGAATGGTGCATGGCACACTGCC
EP Lib Control Rev 1	ACGACCAGCAACACC
EP Lib Control Rev 2	CGTGTCTTCCGTGC
Tiled Lib Control Fwd	GGTAATTCATCTGTGAAACAAGTGGCAACTCGCCTTGCGGCGATG
Tiled Lib Control Rev	TTGTTTCACAGATGAATTACC
Des Lib C1 Fwd	GGTGTCGTTTTTCATAATACTGGCGGTAATTCATCTGTGAAACAA
Des Lib C1 Rev	AGTATTATGAAAAACGACACC
Des Lib C2 Fwd 1	GGTGTCGTTTTTCATAATACTTGGGGTAATTCAACCGCGAAACAAGAAGCAACTCGCCTT
Des Lib C2 Fwd 2	AATAACCAGCTGGCCGCTGGCGTGGCGCACTATTATATTGACAAG
Des Lib C2 Rev 2	CGGCCAGCTGGTTATT
EP Lib C1 Fwd 2	CTGAAGTATTACGGTTTGCCTCCGAATCGTAATACAGTTCGCCTG
EP Lib C1 Rev 1	GCCACAGACTTCATATCC
EP Lib C1 Rev 2	AGGCAAACCGTAATACTTCAG
EP Lib C2 Fwd	GGTGTCGTTTTTCATAATACTTGGGGTAATTCATCTGCGAAACAAGAAGCAACTCGCCTT
EP Lib C3 Fwd	GCCGCTGGCTTCGCGCACTATTATATTGACAAGGATACCATTTGGCGCACGGAAGACACGTTCAAT
EP Lib C3 Rev	ATAGTGCGCGAAGCC
EP Lib C4 Fwd	GGCAACGATCAGACTCCCCTGCGTGACTTTTTGCAGGCCGAGGAG
EP Lib C4 Rev	CAGGGGAGTCTGATCGT
EP Lib C5 Fwd	GGCAACGATCAGACTCCCCTGGAAGACTTTTTGCAGGCCGAGGAG

EP Lib C6 Fwd	GGTGTCGTTTTTCATAATACTTGGGGTAGCTCATCTGCGAAACAAGAAGCAACTCGCCTT
Tiled Lib C1 Fwd	GCATGGCACACTGCCAATGGCGATAGCACCGGCAACTATATCGGATATGAAGTC
Tiled Lib C1 Rev	ATCGCCATTGGCAGTG
Tiled Lib C2 Fwd	GCCGCTGGCTTCGCGCACTATCGTCCGAGCAAGAATACCATTTGGCGCACG
Tiled Lib C3 Fwd	GCATGGCACACTGCCAATGGCTGGCGTCTGATGAACTATATCGGATATGAA
Tiled Lib C4 Fwd	AATATGAACTATATCGGATATTGGTTTAACGGCAACGATCAGACTCCCCTG
Tiled Lib C4 Rev	GTTAAACCAATATCCGATATAGTTCATATT

Supplemental Table S4.11 DNA primers used for Illumina sequencing

Primer	DNA Sequence
Oligopool Illumina Fwd N1	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGNGCTCTGCAGGCTAGT
Oligopool Illumina Fwd N2	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGNNGCTCTGCAGGCTAGT
Oligopool Illumina Fwd N3	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGNNNGCTCTGCAGGCTAGT
Oligopool Illumina Rev N1	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGNATCTGGTTTCAAACTGGG
Oligopool Illumina Rev N2	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGNNATCTGGTTTCAAACTGGG
Oligopool Illumina Rev N3	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGNNNATCTGGTTTCAAACTGGG
OmpA/TorA Illumina Rev N1	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGNGGGAGTCTGATCGTTGCC
OmpA/TorA Illumina Rev N2	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGNNGGGAGTCTGATCGTTGCC
OmpA/TorA Illumina Rev N3	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGNNNGGGAGTCTGATCGTTGCC
LysEFm5-Illumina-p13-Fwd1-N1	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGNCGTGAGTTGCTGGTCGTCGC
LysEFm5-Illumina-p13-Fwd1-N2	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGNNCGTGAGTTGCTGGTCGTCGC
LysEFm5-Illumina-p13-Fwd1-N3	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGNNNCGTGAGTTGCTGGTCGTCGC
LysEFm5-Illumina-p13-Fwd2-N1	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGNACATCGTTGCTGGTCGTCGC
LysEFm5-Illumina-p13-Fwd2-N2	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGNNACATCGTTGCTGGTCGTCGC
LysEFm5-Illumina-p13-Fwd2-N3	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGNNNACATCGTTGCTGGTCGTCGC
LysEFm5-Illumina-p157-Rev-N1	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGNTTGAGCCTGGGTGGAATTAAA

LysEFm5-Illumina-p157-Rev-N2	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGNNTTGAGCCTGGGTGGAATTAAA
LysEFm5-Illumina-p157-Rev-N3	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGNNNTTGAGCCTGGGTGGAATTAAA
LysEFm5-Illumina-p5-Fwd1-N1	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGNCAGGCTAGTATGGTTGAGATTATT
LysEFm5-Illumina-p5-Fwd1-N2	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGNNCAGGCTAGTATGGTTGAGATTATT
LysEFm5-Illumina-p5-Fwd1-N3	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGNNNCAGGCTAGTATGGTTGAGATTATT
LysEFm5-Illumina-p145-Rev-N1	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGNACATCAAGGGAACGTTTTGGG
LysEFm5-Illumina-p145-Rev-N2	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGNNACATCAAGGGAACGTTTTGGG
LysEFm5-Illumina-p145-Rev-N3	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGNNNACATCAAGGGAACGTTTTGGG
LysEFm5-Illumina-p40-Fwd1-N1	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGNCGTGAACTCGCCTTGCGG
LysEFm5-Illumina-p40-Fwd1-N2	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGNNCGTGAACTCGCCTTGCGG
LysEFm5-Illumina-p40-Fwd1-N3	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGNNNCGTGAACTCGCCTTGCGG
LysEFm5-Illumina-p185-Rev-N1	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGNTCAGAAATAAGCTTTTGTTC
LysEFm5-Illumina-p185-Rev-N2	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGNNTCAGAAATAAGCTTTTGTTC
LysEFm5-Illumina-p185-Rev-N3	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGNNNTCAGAAATAAGCTTTTGTTC
Ni5N501	AATGATACGGCGACCACCGAGATCTACACTAGATCGCTCGTCGGCAGCGTC
Ni5N502	AATGATACGGCGACCACCGAGATCTACACCTCTCTATTCGTCGGCAGCGTC
Ni5N503	AATGATACGGCGACCACCGAGATCTACACTATCCTCTTCGTCGGCAGCGTC
Ni5N504	AATGATACGGCGACCACCGAGATCTACACAGAGTAGATCGTCGGCAGCGTC
Ni5N505	AATGATACGGCGACCACCGAGATCTACACGTAAGGAGTCGTCGGCAGCGTC
Ni5N506	AATGATACGGCGACCACCGAGATCTACACACTGCATATCGTCGGCAGCGTC
Ni5N507	AATGATACGGCGACCACCGAGATCTACACAAGGAGTATCGTCGGCAGCGTC
Ni5N508	AATGATACGGCGACCACCGAGATCTACACCTAAGCCTTCGTCGGCAGCGTC
Ni7N701	CAAGCAGAAGACGGCATACGAGATTCGCCTTAGTCTCGTGGGCTCGG

Ni7N702	CAAGCAGAAGACGGCATACGAGATCTAGTACGGTCTCGTGGGCTCGG
Ni7N703	CAAGCAGAAGACGGCATACGAGATTTCTGCCTGTCTCGTGGGCTCGG
Ni7N704	CAAGCAGAAGACGGCATACGAGATGCTCAGGAGTCTCGTGGGCTCGG
Ni7N705	CAAGCAGAAGACGGCATACGAGATAGGAGTCCGTCTCGTGGGCTCGG
Ni7N706	CAAGCAGAAGACGGCATACGAGATCATGCCTAGTCTCGTGGGCTCGG
Ni7N707	CAAGCAGAAGACGGCATACGAGATGTAGAGAGGTCTCGTGGGCTCG G
Ni7N708	CAAGCAGAAGACGGCATACGAGATCCTCTCTGGTCTCGTGGGCTCGG
Ni7N709	CAAGCAGAAGACGGCATACGAGATAGCGTAGCGTCTCGTGGGCTCGG
Ni7N710	CAAGCAGAAGACGGCATACGAGATCAGCCTCGGTCTCGTGGGCTCGG
Ni7N711	CAAGCAGAAGACGGCATACGAGATTGCCTCTTGTCTCGTGGGCTCGG
Ni7N712	CAAGCAGAAGACGGCATACGAGATTCCTCTACGTCTCGTGGGCTCGG

Supplemental Table 4.12 Lysin variants from Designed, RM, and Triplet libraries

Most stable variant controls							
Library	**Variant**	**Mutation**	**Assay Scores**				
			Chym.	**Tryp.**	**Prot. K**	**Aver. Stab.**	**Depletion**
Designed	DC	T29N, S34T, V35A, T40A	0.921	0.885	0.742	0.849	-1.142
RM	RC	R17L, F71L	0.950	0.960	0.925	0.945	-1.789
Triplet	TC	E38V	0.911	0.684	0.975	0.857	-0.669
Most active and stable variants							
Library	**Variant**	**Mutation**	**Assay Scores**				
			Chym.	**Tryp.**	**Prot. K**	**Aver. Stab.**	**Depletion**
Designed	D1	W30G	0.900	0.825	0.819	0.848	-2.146
Designed	D2	S34T, V35A, F54V	0.693	0.596	0.599	0.629	-1.803
RM	R1	N94S, T97A, L99D, A105N, V124P	0.860	0.551	0.514	0.642	-3.188
RM	R2	V35A	0.641	0.755	0.550	0.649	-3.895

RM	R3	N62D	0.732	0.410	0.673	0.605	-4.603
RM	R4	K100R	0.774	0.732	0.459	0.655	-4.732
RM	R5	N32S, V35A	0.925	0.947	0.783	0.885	-2.502
Triplet	T1	G82S, N83T, M84G	0.797	0.656	0.763	0.739	-7.239
Triplet	T2	Y58R, I59P, D60S	0.909	0.938	0.786	0.878	-9.188
Triplet	T3	D81W, G82R, N83L	0.778	0.653	0.688	0.706	-8.992
Triplet	T4	E90W, V91F, C92N	0.933	0.569	0.831	0.778	-6.987

Chapter 5 – Concluding Remarks

As discussed throughout this thesis, antimicrobial resistance is a critical global health threat, and the application of protein engineering strategies to antimicrobial proteins offers a compelling platform for the continued development of new antimicrobial therapies. The identification of performant class IIa variants from a genome-mined library (Chapter 2) and multiple lysin catalytic domains with enhanced activity and stability profiles (Chapters 3 and 4) demonstrates the utility of these approaches, while our development of high throughput methods for functionally evaluating lysin catalytic domains empowers future engineering efforts (Chapter 4). Further, the investigation of bacterial susceptibility and modes of resistance provide mechanistic insights to guide our use of antimicrobials and limit development of resistance (Chapter 2 and Appendix 7.1).

Our successful identification of class IIa bacteriocin variants from genome-mined sequences highlights the breadth of information available across natural homology data and the potential of class IIa bacteriocins. Yet, the rapid development of resistance towards class IIa bacteriocins and their relatively poor proteolytic stability limit their potential as individual antimicrobial therapies. Moving forward, class IIa bacteriocins are most likely to succeed translationally as combinatorial therapies with other antimicrobials with distinct modes of action. Future studies should focus on assessing the potential of such combinatorial therapies in *Enterococcus* and *Listeria in vivo* models. To combat the poor proteolytic stability of class IIa bacteriocins, beyond engineering strategies

presented in Chapter 2, probiotic bacteria capable of secreting AMPs *in vivo* are particularly interesting for future studies, as these can enable increased localization of AMPs in therapeutic environments. While we suspect combinatorial therapies will help slow the rapid development of resistance against class IIa bacteriocins *in vivo*, we have also demonstrated that this resistance enables efficient study and elucidation of bacterial resistance mechanisms, as shown in Appendix 7.1, making class IIa bacteriocins a strong test case for future work in this space.

While lysins have received much attention as compelling antimicrobials, only a handful have reached clinical trials, as natural lysins often require significant engineering to be therapeutically effective. In particular, lysin therapeutic efficacy would be enhanced by increased specificity and potency, to reduce development of resistance and off-target effects, as well as enhanced physiological stability. We hypothesize the lysin variant with greatest activity and stability following ongoing characterization efforts described in Chapter 4 will be a compelling lead molecule for *in vivo* treatment of *E. faecium* infections. Importantly, evaluation of such molecules *in vivo* is critical to define benchmarks for lysin development, as limited such benchmarks are available in literature. Moreover, our identification of lysin variants with an array of activities and stabilities offers molecules for the rigorous assessment of how these properties impact lysin therapeutic efficacy *in vivo*.

Equally impactful, this thesis work demonstrates an efficient process for engineering lysin function to address their shortcomings, most notably with the

development of a high throughput approach for assessing lysin catalytic activity. Future work in this space should seek to further validate this assay's performance by precise characterization of several catalytic domain clones to identify which molecular factors impact depletion. Additionally, continued engineering of alternative high throughput assays for screening antimicrobial activity is necessary to further empower AMP engineering. While depletion-based assays have enabled orders of magnitude improvement in the screening of antimicrobial proteins, high throughput assays which enable enrichment of improved variants (i.e. positive sorting strategies) are more desirable as iterative rounds of screening could yield highly improved variants, as has been demonstrated for protein affinity maturation[35,36,165]. Thus, while an immediate approach is not obvious, future efforts should aim to design such a system. Additionally, the application of our process demonstrated in Chapter 4 towards lysins targeting different bacteria, including those capable of targeting Gram-negative bacteria, which have proven particularly difficult in development of new antimicrobial treatments, should yield additional compelling lysins for therapeutic applications. Engineering efforts should also be expanded to the cell wall-binding domain to further enhance lysin stability and affinity for bacterial targets of interest.

Notably, this work used an array of protein library design, diversification, and modeling strategies to aid discovery of compelling proteins, broaden our search of protein sequence space, and learn sequence-function relationships. Our results demonstrate that moderate levels of protein mutagenesis elucidated

epistatic relationships via modeling. When possible, use of EVCouplings predictions should provide strong library performance. However, in the absence of large sets of homologous sequences, future work should implement a combination of error-prone PCR and saturation mutagenesis at all three-amino acid sets to balance protein diversification and functional performance.

Natural homologous sequence information was invaluable in both identification of compelling class IIa bacteriocins and design of improved lysin catalytic domains. Further, this work demonstrates a new efficient approach for constructing libraries of chimeric proteins from genome-mined sequences in the context of lysin catalytic domains. This method allows the search of vast protein sequence space and future work should implement this library design both for the identification of new functional lysins and proteins more broadly. Sequence-function modeling in this work was largely done via ridge regression to minimize overfitting and enable straightforward identification and interpretation of sitewise amino acid preferences and epistasis. While these models were predictive and yielded design of compelling variants in Chapter 4, more sophisticated modeling frameworks and machine learning approaches could enable better prediction of lysin function and are compelling to learn additional sequence-function relationships within our data. In particular, neural networks and other nonlinear models, which have proven successful at extracting fundamental protein features to predict more complicated protein functions[30,51,200], should be tested to elucidate lysin sequence features.

Moving forward, the continued development of high throughput screening methods, improving the quality and quantity of functional data available, coupled with use of protein sequence modeling to guide protein design and map the sequence-function landscape will empower protein engineering, both in the context of antimicrobial proteins and other relevant protein families.

Chapter 6 – References

1. Lewis, K. 2020. The Science of Antibiotic Discovery. *Cell* **181**, 29–45.
2. Lewis, K. 2013. Platforms for antibiotic discovery. *Nat. Rev. Drug Discov.* **12**, 371–387.
3. Schatz, A., Bugie, E. & Waksman, S. A. 1944. Streptomycin, a substance exhibiting antibiotic activity against Gram-positive and Gram-negative bacteria. *Proc. Soc. Exp. Biol. Med.* **55**, 66–69.
4. Macfarlane, G. 1984. *Alexander Fleming: The man and the myth*. **1**,
5. Hutchings, M., Truman, A. & Wilkinson, B. 2019. Antibiotics: past, present and future. *Curr. Opin. Microbiol.* **51**, 72–80.
6. Blair, J. M. A., Webber, M. A., Baylay, A. J., Ogbolu, D. O. & Piddock, L. J. V. 2015. Molecular mechanisms of antibiotic resistance. *Nat. Rev. Microbiol.* **13**, 42–51.
7. Wozniak, A., Villagra, N. A., Undabarrena, A., Gallardo, N., Keller, N., Moraga, M., Román, J. C., Mora, G. C. & García, P. 2012. Porin alterations present in non-carbapenemaseproducing Enterobacteriaceae with high and intermediate levels of carbapenem resistance in Chile. *J. Med. Microbiol.* **61**, 1270–1279.
8. Pagès, J. M., James, C. E. & Winterhalter, M. 2008. The porin and the permeating antibiotic: A selective diffusion barrier in Gram-negative bacteria. *Nat. Rev. Microbiol.* **6**, 893–903.
9. Doumith, M., Ellington, M. J., Livermore, D. M. & Woodford, N. 2009. Molecular mechanisms disrupting porin expression in ertapenem-resistant Klebsiella and Enterobacter spp. clinical isolates from the UK. *J. Antimicrob. Chemother.* **63**, 659–667.
10. Geldart, K. & Kaznessis, Y. N. 2017. Characterization of Class IIa Bacteriocin Resistance in *Enterococcus faecium*. *Antimicrob. Agents Chemother.* **61**, 1–17.
11. Gold, H. S. 2001. Vancomycin-resistant enterococci: Mechanisms and clinical observations. *Clin. Infect. Dis.* **33**, 210–219.
12. Levine, D. P. 2006. Vancomycin : A History. *Clin. Infect. Dis.* **42**, 5–12.
13. Melnyk, A. H., Wong, A. & Kassen, R. 2015. The fitness costs of antibiotic resistance mutations. *Evol. Appl.* **8**, 273–283.
14. O'Neill, J. 2016. *Tackling Drug-Resistant Infections Globally: Final Report and Recommendations*. The Review on Antimicrobial Resistance, London, United Kingdom.
15. Ochman, Howard, Lawrence, Jeffrey G. & Groisman, Eduardo A. 2000. Lateral gene transfer and the nature of bacterial innovation. *Nature* **405**, 299–304.
16. Centers for Disease Control and Prevention. 2019. *Antibiotic resistance threats in the United States*.
17. Emberger, J., Tassone, D., Stevens, M. P. & Markley, J. D. 2018. The Current State of Antimicrobial Stewardship: Challenges, Successes, and

Future Directions. *Curr. Infect. Dis. Rep.* **20**,
18. Cotter, P. D., Ross, R. P. & Hill, C. 2013. Bacteriocins-a viable alternative to antibiotics? *Nat. Rev. Microbiol.* **11**, 95–105.
19. Fischetti, V. A. 2008. Bacteriophage lysins as effective antibacterials. *Curr. Opin. Microbiol.* **11**, 393–400.
20. Nelson, D. C., Schmelcher, M., Rodriguez-Rubio, L., Klumpp, J., Pritchard, D. G., Dong, S. & Donovan, D. M. 2012. *Endolysins as Antimicrobials*. Elsevier Inc., **83**,
21. Wang, G. 2014. Human antimicrobial peptides and proteins. *Pharmaceuticals* **7**, 545–594.
22. Ghosh, C., Sarkar, P., Issa, R. & Haldar, J. 2018. Alternatives to Conventional Antibiotics in the Era of Antimicrobial Resistance. *Trends Microbiol.* **xx**, 1–16.
23. Ageitos, J. M., Sánchez-Pérez, A., Calo-Mata, P. & Villa, T. G. 2017. Antimicrobial peptides (AMPs): Ancient compounds that represent novel weapons in the fight against bacteria. *Biochem. Pharmacol.* **133**, 117–138.
24. Cui, Y., Zhang, C., Wang, Y., Shi, J., Zhang, L., Ding, Z., Qu, X. & Cui, H. 2012. Class IIa bacteriocins: Diversity and new developments. *Int. J. Mol. Sci.* **13**, 16668–16707.
25. Balandin, S. V., Sheremeteva, E. V. & Ovchinnikova, T. V. 2019. Pediocin-Like Antimicrobial Peptides of Bacteria. *Biochem.* **84**, 464–478.
26. Fischetti, V. A. 2005. Bacteriophage lytic enzymes: Novel anti-infectives. *Trends Microbiol.* **13**, 491–496.
27. De Maesschalck, V., Gutiérrez, D., Paeshuyse, J., Lavigne, R. & Briers, Y. 2020. Advanced engineering of third-generation lysins and formulation strategies for clinical applications. *Crit. Rev. Microbiol.* **46**, 548–564.
28. Rocklin, G. J., Chidyausiku, T. M., Goreshnik, I., Ford, A., Houliston, S., Lemak, A., Carter, L., Ravichandran, R., Mulligan, V. K., Chevalier, A., Arrowsmith, C. H. & Baker, D. 2017. Global analysis of protein folding using massively parallel design, synthesis, and testing. *Science (80-.).* **357**, 168–175.
29. Nisthal, A., Wang, C. Y., Ary, M. L. & Mayo, S. L. 2019. Protein stability engineering insights revealed by domain-wide comprehensive mutagenesis. *Proc. Natl. Acad. Sci. U. S. A.* **116**, 16367–16377.
30. Golinski, A. W., Mischler, K. M., Laxminarayan, S., Neurock, N. L., Fossing, M., Pichman, H., Martiniani, S. & Hackel, B. J. 2021. High-throughput developability assays enable library-scale identification of producible protein scaffold variants. *Proc. Natl. Acad. Sci. U. S. A.* **118**, 1–11.
31. Dejong, M. P., Ritter, S. C., Fransen, K. A., Tresnak, D. T., Golinski, A. W. & Hackel, B. J. 2021. A Platform for Deep Sequence-Activity Mapping and Engineering Antimicrobial Peptides. *ACS Synth. Biol.*
32. Cantor, A. J., Shah, N. H. & Kuriyan, J. 2018. Deep mutational analysis reveals functional trade-offs in the sequences of EGFR autophosphorylation sites. *Proc. Natl. Acad. Sci.* **115**, 201803598.

33. Manojlović, L. M. 2015. Photometry-based estimation of the total number of stars in the Universe. *Appl. Opt.* **54**, 6589.
34. Romero, P. A. & Arnold, F. H. 2009. Exploring protein fitness landscapes by directed evolution. *Nat. Rev. Mol. Cell Biol.* **10**, 866–876.
35. Boder, E. & Wittrup, K. D. 1997. Yeast surface display for screening combinatorial polypeptide libraries. *Nat. Biotechnol.* **15**, 553–557.
36. Hackel, B. J., Kapila, A. & Dane Wittrup, K. 2008. Picomolar Affinity Fibronectin Domains Engineered Utilizing Loop Length Diversity, Recursive Mutagenesis, and Loop Shuffling. *J. Mol. Biol.* **381**, 1238–1252.
37. Rader, C. & Barbas, C. F. 1997. Phage display of combinatorial antibody libraries. *Curr. Opin. Biotechnol.* **8**, 503–508.
38. Tresnak, D. T. & Hackel, B. J. 2020. Mining and statistical modeling of natural and variant class IIa bacteriocins elucidates activity and selectivity profiles across species. *Appl. Environ. Microbiol.* **86**, 1–18.
39. Lai, P.-K., Tresnak, D. T. & Hackel, B. J. 2019. Identification and elucidation of proline-rich antimicrobial peptides with enhanced potency and delivery. *Biotechnol. Bioeng.* **116**, 2439–2450.
40. Baryakova, T. H., Ritter, S. C., Tresnak, D. T. & Hackel, B. J. 2020. Computationally Aided Discovery of LysEFm5 Variants with Improved Catalytic Activity and Stability. *Appl. Environ. Microbiol.* **86**, 1–21.
41. Ritter, S. C., Yang, M. L., Kaznessis, Y. N. & Hackel, B. J. 2018. Multispecies activity screening of microcin J25 mutants yields antimicrobials with increased specificity toward pathogenic *Salmonella* species relative to human commensal *Escherichia coli*. *Biotechnol. Bioeng.* **115**, 2394–2404.
42. Scanlon, T. C., Dostal, S. M. & Griswold, K. E. 2014. A high-throughput screen for antibiotic drug discovery. *Biotechnol. Bioeng.* **111**, 232–243.
43. Tucker, A. T., Leonard, S. P., DuBois, C. D., Knauf, G. A., Cunningham, A. L., Wilke, C. O., Trent, M. S. & Davies, B. W. 2018. Discovery of Next-Generation Antimicrobials through Bacterial Self-Screening of Surface-Displayed Peptide Libraries. *Cell* **172**, 1–11.
44. Sharma, P., Marada, V. V. V. R., Cai, Q., Kizerwetter, M., He, Y., Wolf, S. P., Schreiber, K., Clausen, H., Schreiber, H. & Kranz, D. M. 2020. Structure-guided engineering of the affinity and specificity of CARs against Tn-glycopeptides. *Proc. Natl. Acad. Sci. U. S. A.* **117**, 15148–15159.
45. Koenig, P.-A., Das, H., Liu, H., Kümmerer, B. M., Gohr, F. N., Jenster, L.-M., Schiffelers, L. D. J., Tesfamariam, Y. M., Uchima, M., Wuerth, J. D., Gatterdam, K., Ruetalo, N., Christensen, M. H., Fandrey, C. I., Normann, S., Tödtmann, J. M. P., Pritzl, S., Hanke, L., Boos, J., *et al.* 2021. Structure-guided multivalent nanobodies block SARS-CoV-2 infection and suppress mutational escape. *Science (80-.).* **6230**, eabe6230.
46. Alford, R. F., Leaver-Fay, A., Jeliazkov, J. R., O'Meara, M. J., DiMaio, F. P., Park, H., Shapovalov, M. V., Renfrew, P. D., Mulligan, V. K., Kappel, K., Labonte, J. W., Pacella, M. S., Bonneau, R., Bradley, P., Dunbrack, R. L., Das, R., Baker, D., Kuhlman, B., Kortemme, T., *et al.* 2017. The Rosetta

All-Atom Energy Function for Macromolecular Modeling and Design. *J. Chem. Theory Comput.* **13**, 3031–3048.
47. Buß, O., Rudat, J. & Ochsenreither, K. 2018. FoldX as Protein Engineering Tool: Better Than Random Based Approaches? *Comput. Struct. Biotechnol. J.* **16**, 25–33.
48. Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A., Bridgland, A., Meyer, C., Kohl, S. A. A., Ballard, A. J., Cowie, A., Romera-Paredes, B., Nikolov, S., Jain, R., Adler, J., *et al.* 2021. Highly accurate protein structure prediction with AlphaFold. *Nature*
49. Bhardwaj, G., Mulligan, V. K., Bahl, C. D., Gilmore, J. M., Harvey, P. J., Cheneval, O., Buchko, G. W., Pulavarti, S. V. S. R. K., Kaas, Q., Eletsky, A., Huang, P. S., Johnsen, W. A., Greisen, P. J., Rocklin, G. J., Song, Y., Linsky, T. W., Watkins, A., Rettie, S. A., Xu, X., *et al.* 2016. Accurate de novo design of hyperstable constrained peptides. *Nature* **538**,
50. Ritter, S. C. & Hackel, B. J. 2019. Validation and Stabilization of a Prophage Lysin of Clostridium perfringens by Using Yeast Surface Display and Coevolutionary Models . *Appl. Environ. Microbiol.* **85**, 1–18.
51. Wu, Z., Jennifer Kan, S. B., Lewis, R. D., Wittmann, B. J. & Arnold, F. H. 2019. Machine learning-assisted directed protein evolution with combinatorial libraries. *Proc. Natl. Acad. Sci. U. S. A.* **116**, 8852–8858.
52. Centers for Disease Control and Prevention. 2013. *Antibiotic Resistance Threats in the United States*. U.S. Department of Health and Human Services, Washington, D.C.
53. Van Boeckel, T. P., Brower, C., Gilbert, M., Grenfell, B. T., Levin, S. A., Robinson, T. P., Teillant, A. & Laxminarayan, R. 2015. Global trends in antimicrobial use in food animals. *Proc. Natl. Acad. Sci. U. S. A.* **112**, 5649–5654.
54. Van Boeckel, T. P., Pires, J., Silvester, R., Zhao, C., Song, J., Criscuolo, N. G., Gilbert, M., Bonhoeffer, S. & Laxminarayan, R. 2019. Global trends in antimicrobial resistance in animals in low- and middle-income countries. *Science (80-.).* **365**, eaaw1944.
55. Lopez, F. E., Vincent, P. A., Zenoff, A. M., Salomón, R. A. & Farías, R. N. 2007. Efficacy of microcin J25 in biomatrices and in a mouse model of *Salmonella* infection. *J. Antimicrob. Chemother.* **59**, 676–680.
56. Wu, X., Wang, Z., Li, X., Fan, Y., He, G., Wan, Y., Yu, C., Tang, J., Li, M., Zhang, X., Zhang, H., Xiang, R., Pan, Y., Liu, Y., Lu, L. & Yang, L. 2014. *In vitro* and *in vivo* activities of antimicrobial peptides developed using an amino acid-based activity prediction method. *Antimicrob. Agents Chemother.* **58**, 5342–5349.
57. Dabour, N., Zihler, A., Kheadr, E., Lacroix, C. & Fliss, I. 2009. *In vivo* study on the effectiveness of pediocin PA-1 and *Pediococcus acidilactici* UL5 at inhibiting *Listeria monocytogenes*. *Int. J. Food Microbiol.* **133**, 225–233.
58. Geldart, K. G., Kommineni, S., Forbes, M., Dunny, G. M., Salzman, N. H. & Kaznessis, Y. N. 2018. Engineered *E. coli* Nissle 1917 for the reduction of

Vancomycin-resistant *Enterococcus* in the intestinal tract. *Bioeng. Transl. Med.* 197–208.
59. Forkus, B., Ritter, S., Vlysidis, M., Geldart, K. & Kaznessis, Y. N. 2017. Antimicrobial Probiotics Reduce *Salmonella enterica* in Turkey Gastrointestinal Tracts. *Sci. Rep.* **7**, 1–9.
60. Hwang, I. Y., Koh, E., Wong, A., March, J. C., Bentley, W. E., Lee, Y. S. & Chang, M. W. 2017. Engineered probiotic *Escherichia coli* can eliminate and prevent *Pseudomonas aeruginosa* gut infection in animal models. *Nat. Commun.* **8**, 1–11.
61. Volzing, K., Borrero, J., Sadowsky, M. J. & Kaznessis, Y. N. 2013. Antimicrobial peptides targeting gram-negative pathogens, produced and delivered by lactic acid bacteria. *ACS Synth. Biol.* **2**, 643–650.
62. Francino, M. P. 2016. Antibiotics and the human gut microbiome: Dysbioses and accumulation of resistances. *Front. Microbiol.* **6**, 1–11.
63. Guinane, C. M. & Cotter, P. D. 2013. Role of the gut microbiota in health and chronic gastrointestinal disease: Understanding a hidden metabolic organ. *Therap. Adv. Gastroenterol.* **6**, 295–308.
64. Rea, M. C., Dobson, A., O'Sullivan, O., Crispie, F., Fouhy, F., Cotter, P. D., Shanahan, F., Kiely, B., Hill, C. & Ross, R. P. 2011. Effect of broad- and narrow-spectrum antimicrobials on *Clostridium difficile* and microbial diversity in a model of the distal colon. *Proc. Natl. Acad. Sci.* **108**, 4639–4644.
65. Langdon, A., Crook, N. & Dantas, G. 2016. The effects of antibiotics on the microbiome throughout development and alternative approaches for therapeutic modulation. *Genome Med.* **8**,
66. Field, D., Connor, P. M. O., Cotter, P. D., Hill, C. & Ross, R. P. 2008. The generation of nisin variants with enhanced activity against specific Gram-positive pathogens. *Mol. Microbiol.* **69**, 218–230.
67. Eckert, R., Qi, F., Yarbrough, D. K., He, J., Anderson, M. H. & Shi, W. 2006. Adding Selectivity to Antimicrobial Peptides: Rational Design of a Multidomain Peptide against *Pseudomonas spp. Antimicrob. Agents Chemother.* **50**, 1480–1488.
68. Fimland, G., Johnsen, L., Dalhus, B. & Nissen-Meyer, J. 2005. Pediocin-like antimicrobial peptides (class IIa bacteriocins) and their immunity proteins: Biosynthesis, structure, and mode of action. *J. Pept. Sci.* **11**, 688–696.
69. Drider, D., Fimland, G., Hechard, Y., McMullen, L. M. & Prevost, H. 2006. The Continuing Story of Class IIa Bacteriocins. *Microbiol. Mol. Biol. Rev.* **70**, 564–582.
70. Ríos Colombo, N. S., Chalón, M. C., Navarro, S. A. & Bellomio, A. 2018. Pediocin-like bacteriocins: new perspectives on mechanism of action and immunity. *Curr. Genet.* **64**, 345–351.
71. Kjos, M., Salehian, Z., Nes, I. F. & Diep, D. B. 2010. An extracellular loop of the mannose phosphotransferase system component IIC is responsible for specific targeting by class IIa bacteriocins. *J. Bacteriol.* **192**, 5906–

5913.
72. Zhou, W., Wang, G., Wang, C., Ren, F. & Hao, Y. 2016. Both IIC and IID components of mannose phosphotransferase system are involved in the specific recognition between immunity protein PedB and bacteriocin-receptor complex. *PLoS One* **11**, 1–12.
73. Kjos, M., Nes, I. F. & Diep, D. B. 2009. Class II one-peptide bacteriocins target a phylogenetically defined subgroup of mannose phosphotransferase systems on sensitive cells. *Microbiology* **155**, 2949–2961.
74. Bernbom, N., Jelle, B., Brogren, C. H., Vogensen, F. K., Nørrung, B. & Licht, T. R. 2009. Pediocin PA-1 and a pediocin producing *Lactobacillus plantarum* strain do not change the HMA rat microbiota. *Int. J. Food Microbiol.* **130**, 251–257.
75. Tominaga, T. & Hatakeyama, Y. 2006. Determination of essential and variable residues in pediocin PA-1 by NNK scanning. *Appl. Environ. Microbiol.* **72**, 1141–1147.
76. Kazazic, M., Nissen-Meyer, J. & Fimland, G. 2002. 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.
77. Mcclintock, M. K., Kaznessis, Y. N. & Hackel, B. J. 2016. Enterocin A mutants identified by saturation mutagenesis enhance potency towards vancomycin-resistant enterococci. *Biotechnol. Bioeng.* **113**, 414–423.
78. Tominaga, T. & Hatakeyama, Y. 2007. Development of innovative pediocin PA-1 by DNA shuffling among class IIa bacteriocins. *Appl. Environ. Microbiol.* **73**, 5292–5299.
79. Johnsen, L., Fimland, G., Eijsink, V. & Nissen-Meyer, J. 2000. Engineering increased stability in the antimicrobial peptide pediocin PA-1. *Appl. Environ. Microbiol.* **66**, 4798–4802.
80. Fimland, G., Johnsen, L., Axelsson, L., Brurberg, M. B., Nes, I. F., Eijsink, V. G. H. & Nissen-Meyer, J. 2000. A C-terminal disulfide bridge in pediocin-like bacteriocins renders bacteriocin activity less temperature dependent and is a major determinant of the antimicrobial spectrum. *J. Bacteriol.* **182**, 2643–2648.
81. Kjos, M., Nes, I. F. & Diep, D. B. 2011. Mechanisms of resistance to bacteriocins targeting the mannose phosphotransferase system. *Appl. Environ. Microbiol.* **77**, 3335–3342.
82. Vadyvaloo, V., Hastings, J. W., Van der Merwe, M. J. & Rautenbach, M. 2002. Membranes of class IIa bacteriocin-resistant *Listeria monocytogenes* cells contain increased levels of desaturated and short-acyl-chain phosphatidylglycerols. *Appl. Environ. Microbiol.* **68**, 5223–5230.
83. Yu, G., Baeder, D. Y. & Regoes, R. R. 2016. Combination Effects of Antimicrobial Peptides. *Antimicrob. Agents Chemother.* **60**, 1717–1724.
84. Gelman, D., Beyth, S., Lerer, V., Adler, K., Poradosu-Cohen, R., Coppenhagen-Glazer, S. & Hazan, R. 2018. Combined bacteriophages and antibiotics as an efficient therapy against VRE *Enterococcus faecalis*

in a mouse model. *Res. Microbiol.* **169**, 531–539.
85. Mohamed, M. F., Abdelkhalek, A. & Seleem, M. N. 2016. Evaluation of short synthetic antimicrobial peptides for treatment of drug-resistant and intracellular *Staphylococcus aureus*. *Sci. Rep.* **6**, 2–15.
86. Bradshaw, J. P. 2003. Cationic antimicrobial peptides: Issues for potential clinical use. *BioDrugs* **17**, 233–240.
87. Sato, A. K., Viswanathan, M., Kent, R. B. & Wood, C. R. 2006. Therapeutic peptides: technological advances driving peptides into development. *Curr. Opin. Biotechnol.* **17**, 638–642.
88. The UniProt Consortium. 2017. UniProt: The universal protein knowledgebase. *Nucleic Acids Res.* **45**, D158–D169.
89. Gasteiger, E., Hoogland, C., Gattiker, A., Duvaud, S., Wilkins, M. R., Appel, R. D. & Bairoch, A. 2005. Protein identification and analysis tools in the ExPASy server. in 571–607.
90. Arbulu, S., Jiménez, J. J., Gútiez, L., Feito, J., Cintas, L. M., Herranz, C. & Hernández, P. E. 2019. Cloning and expression of synthetic genes encoding native, hybrid- and bacteriocin-derived chimeras from mature class IIa bacteriocins, by *Pichia pastoris* (syn. *Komagataella* spp.). *Food Res. Int.* **121**, 888–899.
91. O'Shea, E. F., O'Connor, P. M., Cotter, P. D., Ross, R. P. & Hill, C. 2010. Synthesis of Trypsin-Resistant Variants of the *Listeria*-Active Bacteriocin Salivaricin P. *Appl. Environ. Microbiol.* **76**, 5356–5362.
92. Himeno, K., Fujita, K., Zendo, T., Wilaipun, P., Ishibashi, N., Masuda, Y., Yoneyama, F., Leelawatcharamas, V., Nakayama, J. & Sonomoto, K. 2012. Identification of Enterocin NKR-5-3C, a Novel Class IIa Bacteriocin Produced by a Multiple Bacteriocin Producer, *Enterococcus faecium* NKR-5-3. *Biosci. Biotechnol. Biochem.* **76**, 1245–1247.
93. Cintas, L. M., Casaus, P., Håvarstein, L. S., Hernández, P. E. & Nes, I. F. 1997. Biochemical and genetic characterization of enterocin P, a novel sec- dependent bacteriocin from *Enterococcus faecium* P13 with a broad antimicrobial spectrum. *Appl. Environ. Microbiol.* **63**, 4321–4330.
94. Fimland, G., Blingsmo, O. R., Sletten, K., Jung, G., Nes, I. F. & Nissen-Meyer, J. 1996. New biologically active hybrid bacteriocins constructed by combining regions from various pediocin-like bacteriocins: The C-terminal region is important for determining specificity. *Appl. Environ. Microbiol.* **62**, 3313–3318.
95. Eijsink, V. G. H., Skeie, M., Middelhoven, P. H., Brurberg, M. B. & Nes, I. F. 1998. Comparative studies of class IIa bacteriocins of lactic acid bacteria. *Appl. Environ. Microbiol.* **64**, 3275–3281.
96. Steipe, B., Schiller, B., Pluckthun, A. & Steinbacher, S. 1994. Sequence Statistics Reliably Predict Stabilizing Mutations in a Protein Domain. *J. Mol. Biol.* **240**, 188–192.
97. Steipe, B. 2004. Consensus-based engineering of protein stability: From intrabodies to thermostable enzymes. *Methods Enzymol.* **388**, 176–186.
98. Cochran, J. R., Kim, Y. S., Lippow, S. M., Rao, B. & Wittrup, K. D. 2006.

Improved mutants from directed evolution are biased to orthologous substitutions. *Protein Eng. Des. Sel.* **19**, 245–253.
99. Geldart, K., Borrero, J. & Kaznessis, Y. N. 2015. Chloride-inducible expression vector for delivery of antimicrobial peptides targeting antibiotic-resistant *Enterococcus faecium*. *Appl. Environ. Microbiol.* **81**, 3889–3897.
100. van Asseldonk, M., de Vos, W. M. & Simons, G. 1993. Functional analysis of the *Lactococcus lactis* usp45 secretion signal in the secretion of a homologous proteinase and a heterologous α-amylase. *MGG Mol. Gen. Genet.* **240**, 428–434.
101. Steidler, L., Hans, W., Schotte, L., Neirynck, S., Obermeier, F., Falk, W., Fiers, W. & Remaut, E. 2000. Treatment of murine colitis by *Lactococcus lactis* secreting interleukin-10. *Science (80-.).* **289**, 1352–1355.
102. Steidler, L., Neirynck, S., Huyghebaert, N., Snoeck, V., Vermeire, A., Goddeeris, B., Cox, E., Remon, J. P. & Remaut, E. 2003. Biological containment of genetically modified *Lactococcus lactis* for intestinal delivery of human interleukin 10. *Nat. Biotechnol.* **21**, 785–789.
103. Song, A. A. L., In, L. L. A., Lim, S. H. E. & Rahim, R. A. 2017. A review on *Lactococcus lactis*: From food to factory. *Microb. Cell Fact.* **16**, 1–15.
104. Suzuki, M., Christians, F. C., Kim, B., Skandalis, A., Black, M. E. & Loeb, L. A. 1996. Tolerance of different proteins for amino acid diversity. *Mol. Divers.* **2**, 111–118.
105. Guo, H. H., Choe, J. & Loeb, L. A. 2004. Protein tolerance to random amino acid change. *Proc. Natl. Acad. Sci. U. S. A.* **101**, 9205–9210.
106. Johnsen, L., Fimland, G. & Nissen-Meyer, J. 2005. The C-terminal domain of pediocin-like antimicrobial peptides (class IIa bacteriocins) is involved in specific recognition of the C-terminal part of cognate immunity proteins and in determining the antimicrobial spectrum. *J. Biol. Chem.* **280**, 9243–9250.
107. Hur, S. J., Lim, B. O., Decker, E. A. & McClements, D. J. 2011. *In vitro* human digestion models for food applications. *Food Chem.* **125**, 1–12.
108. Acuña, L., Morero, R. D. & Bellomio, A. 2011. Development of Wide-Spectrum Hybrid Bacteriocins for Food Biopreservation. *Food Bioprocess Technol.* **4**, 1029–1049.
109. Acuña, L., Picariello, G., Sesma, F., Morero, R. D. & Bellomio, A. 2012. A new hybrid bacteriocin, Ent35-MccV, displays antimicrobial activity against pathogenic Gram-positive and Gram-negative bacteria. *FEBS Open Bio* **2**, 12–19.
110. El-Gebali, S., Mistry, J., Bateman, A., Eddy, S. R., Luciani, A., Potter, S. C., Qureshi, M., Richardson, L. J., Salazar, G. A., Smart, A., Sonnhammer, E. L. L., Hirsh, L., Paladin, L., Piovesan, D., Tosatto, S. C. E. & Finn, R. D. 2019. The Pfam protein families database in 2019. *Nucleic Acids Res.* **47**, D427–D432.
111. Potter, S. C., Luciani, A., Eddy, S. R., Park, Y., Lopez, R. & Finn, R. D. 2018. HMMER web server: 2018 update. *Nucleic Acids Res.* **46**, W200–W204.
112. Richard, C., Drider, D., Elmorjani, K., Marion, D. & Prévost, H. 2004.

Heterologous expression and purification of active divercin V41, a class IIa bacteriocin encoded by a synthetic gene in *Escherichia coli*. *J. Bacteriol.* **186**, 4276–84.
113. Holo, H. & Nes, I. F. 1989. High-Frequency Transformation, by Electroporation, of *Lactococcus Lactis* subsp. *cremoris* Grown with Glycine in Osmotically Stabilized Media. *Appl. Environ. Microbiol.* **55**, 3119–3123.
114. Edgar, R. C. 2010. Search and clustering orders of magnitude faster than BLAST. *Bioinformatics* **26**, 2460–2461.
115. Edgar, R. C. 2016. UNOISE2: improved error-correction for Illumina 16S and ITS amplicon sequencing. *bioRxiv* 081257.
116. Lai, P.-K., Geldart, K., Ritter, S., Kaznessis, Y. N. & Hackel, B. J. 2018. Systematic mutagenesis of oncocin reveals enhanced activity and insights into the mechanisms of antimicrobial activity. *Mol. Syst. Des. Eng.* **3**, 930–941.
117. Agarwala, R., Barrett, T., Beck, J., Benson, D. A., Bollin, C., Bolton, E., Bourexis, D., Brister, J. R., Bryant, S. H., Canese, K., Charowhas, C., Clark, K., Dicuccio, M., Dondoshansky, I., Federhen, S., Feolo, M., Funk, K., Geer, L. Y., Gorelenkov, V., *et al.* 2016. Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.* **44**, D7–D19.
118. Barriere, S. L. 2015. Clinical, economic and societal impact of antibiotic resistance. *Expert Opin. Pharmacother.* **16**, 151–153.
119. Young, R. 1992. Bacteriophage lysis: mechanism and regulation. *Microbiol. Rev.* **56**, 430–481.
120. São-José, C. 2018. Engineering of phage-derived lytic enzymes: Improving their potential as antimicrobials. *Antibiotics* **7**,
121. Loeffler, J. M., Nelson, D. & Fischetti, V. A. 2001. Rapid killing of Streptococcus pneumoniae with a bacteriophage cell wall hydrolase. *Science (80-.).* **294**, 2170–2172.
122. Schuch, R., Nelson, D. & Fischetti, V. A. 2002. A bacteriolytic agent that detects and kills. *Nature* **418**, 884–889.
123. Schmelcher, M., Donovan, D. M. & Loessner, M. J. 2012. Bacteriophage endolysins as novel antimicrobials. *Future Microbiol.* **7**, 1147–1171.
124. Nelson, D., Loomis, L. & Fischetti, V. A. 2001. 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.
125. Wang, Q., Euler, C. W., Delaune, A. & Fischetti, V. A. 2015. Using a novel lysin to help control Clostridium difficile infections. *Antimicrob. Agents Chemother.* **59**, 7447–7457.
126. Marr, A. K., Gooderham, W. J. & Hancock, R. E. 2006. Antibacterial peptides for therapeutic use: obstacles and realistic outlook. *Curr. Opin. Pharmacol.* **6**, 468–472.
127. Loessner, M. J. 2005. Bacteriophage endolysins - Current state of research and applications. *Curr. Opin. Microbiol.* **8**, 480–487.
128. Leonard, E., Ajikumar, P. K., Thayer, K., Xiao, W. H., Mo, J. D., Tidor, B.,

Stephanopoulos, G. & Prather, K. L. J. 2010. Combining metabolic and protein engineering of a terpenoid biosynthetic pathway for overproduction and selectivity control. *Proc. Natl. Acad. Sci. U. S. A.* **107**, 13654–13659.
129. Vermassen, A., Leroy, S., Talon, R., Provot, C., Popowska, M. & Desvaux, M. 2019. Cell wall hydrolases in bacteria: Insight on the diversity of cell wall amidases, glycosidases and peptidases toward peptidoglycan. *Front. Microbiol.* **10**,
130. Croux, C., Ronda, C., Lopez, R. & Garcia, J. L. 1993. Interchange of functional domains switches enzyme specificity: construction of a chimeric pneumococcal-clostridial cell wall lytic enzyme. *Mol. Microbiol.* **9**, 1019–1025.
131. Díez-Martínez, R., De Paz, H. D., García-Fernández, E., Bustamante, N., Euler, C. W., Fischetti, V. A., Menendez, M. & García, P. 2014. A novel chimeric phage lysin with high in vitro and in vivo bactericidal activity against Streptococcus pneumoniae. *J. Antimicrob. Chemother.* **70**, 1763–1773.
132. Gong, P., Cheng, M., Li, X., Jiang, H., Yu, C., Kahaer, N., Li, J., Zhang, L., Xia, F., Hu, L., Sun, C., Feng, X., Lei, L., Han, W. & Gu, J. 2016. Characterization of Enterococcus faecium bacteriophage IME-EFm5 and its endolysin LysEFm5. *Virology* **492**, 11–20.
133. Patterson, J. E., Sweeney, A. H., Simms, M., Carley, N., Mangi, R., Sabetta, J. & Lyons, R. W. 1995. An Analysis of 110 Serious Enterococcal Infections. *Medicine (Baltimore).* **74**, 191–200.
134. DiazGranados, C. A., Zimmer, S. M., Klein, M. & Jernigan, J. A. 2005. Comparison of mortality associated with vancomycin-resistant and vancomycin-susceptible enterococcal bloodstream infections: A meta-analysis. *Clin. Infect. Dis.* **41**, 327–333.
135. Edmond, M. B., Ober, J. F., Dawson, J. D., Weinbaum, D. L. & Wenzel, R. P. 1996. Vancomycin-resistant enterococcal bacteremia: Natural history and attributable mortality. *Clin. Infect. Dis.* **23**, 1234–1239.
136. Low, L. Y., Yang, C., Perego, M., Osterman, A. & Liddington, R. C. 2005. Structure and lytic activity of a Bacillus anthracis prophage endolysin. *J. Biol. Chem.* **280**, 35433–35439.
137. Zoll, S., Pätzold, B., Schlag, M., Götz, F., Kalbacher, H. & Stehle, T. 2010. Structural basis of cell wall cleavage by a staphylococcal autolysin. *PLoS Pathog.* **6**,
138. Yoong, P., Schuch, R., Nelson, D. & Fischetti, V. A. 2004. Identification of a Broadly Active Phage Lytic Enzyme with Lethal Activity against Antibiotic-Resistant Enterococcus faecalis and Enterococcus faecium. *J. Bacteriol.* **186**, 4808–4812.
139. Midelfort, K. S., Kumar, R., Han, S., Karmilowicz, M. J., McConnell, K., Gehlhaar, D. K., Mistry, A., Chang, J. S., Anderson, M., Villalobos, A., Minshull, J., Govindarajan, S. & Wong, J. W. 2013. Redesigning and characterizing the substrate specificity and activity of Vibrio fluvialis

aminotransferase for the synthesis of imagabalin. *Protein Eng. Des. Sel.* **26**, 25–33.
140. Miklos, A. E., Kluwe, C., Der, B. S., Pai, S., Sircar, A., Hughes, R. A., Berrondo, M., Xu, J., Codrea, V., Buckley, P. E., Calm, A. M., Welsh, H. S., Warner, C. R., Zacharko, M. A., Carney, J. P., Gray, J. J., Georgiou, G., Kuhlman, B. & Ellington, A. D. 2012. Structure-based design of supercharged, highly thermoresistant antibodies. *Chem. Biol.* **19**, 449–455.
141. Cocco, S., Feinauer, C., Figliuzzi, M., Monasson, R. & Weigt, M. 2018. Inverse Statistical Physics of Protein Sequences: A Key Issues Review.
142. Hopf, T. A., Ingraham, J. B., Poelwijk, F. J., Schärfe, C. P. I., Springer, M., Sander, C. & Marks, D. S. 2017. Mutation effects predicted from sequence co-variation. *Nat. Biotechnol.* **35**, 128–135.
143. Figliuzzi, M., Barrat-Charlaix, P. & Weigt, M. 2018. How pairwise coevolutionary models capture the collective residue variability in proteins? *Mol. Biol. Evol.* **35**, 1018–1027.
144. Miton, C. M. & Tokuriki, N. 2016. How mutational epistasis impairs predictability in protein evolution and design. *Protein Sci.* 1260–1272.
145. Anishchenko, I., Ovchinnikov, S., Kamisetty, H. & Baker, D. 2017. Origins of coevolution between residues distant in protein 3D structures. *Proc. Natl. Acad. Sci.* **114**, 9122–9127.
146. Levy, R. M., Haldane, A. & Flynn, W. F. 2017. Potts Hamiltonian models of protein co-variation, free energy landscapes, and evolutionary fitness. *Curr. Opin. Struct. Biol.* **43**, 55–62.
147. Miyazawa, S. 2017. Selection originating from protein stability/foldability: Relationships between protein folding free energy, sequence ensemble, and fitness. *J. Theor. Biol.* **433**, 21–38.
148. Jacquin, H., Gilson, A., Shakhnovich, E., Cocco, S. & Monasson, R. 2016. Benchmarking Inverse Statistical Approaches for Protein Structure and Design with Exactly Solvable Models. *PLoS Comput. Biol.* **12**, 1–18.
149. Ekeberg, M., Lövkvist, C., Lan, Y., Weigt, M. & Aurell, E. 2013. Improved contact prediction in proteins: Using pseudolikelihoods to infer Potts models. *Phys. Rev. E - Stat. Nonlinear, Soft Matter Phys.* **87**, 1–16.
150. Reetz, M. T., Wang, L. W. & Bocola, M. 2006. Directed evolution of enantioselective enzymes: Iterative cycles of CASTing for probing protein-sequence space. *Angew. Chemie - Int. Ed.* **45**, 1236–1241.
151. Sandalova, T., Lee, M., Henriques-Normark, B., Hesek, D., Mobashery, S., Mellroth, P. & Achour, A. 2016. 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.
152. Finn, R. D., Clements, J. & Eddy, S. R. 2011. HMMER web server: Interactive sequence similarity searching. *Nucleic Acids Res.* **39**, 29–37.
153. Pei, J., Kim, B. H. & Grishin, N. V. 2008. PROMALS3D: A tool for multiple protein sequence and structure alignments. *Nucleic Acids Res.* **36**, 2295–2300.

154. Jacobs, T. M., Yumerefendi, H., Kuhlman, B. & Leaver-Fay, A. 2015. SwiftLib: Rapid degenerate-codon-library optimization through dynamic programming. *Nucleic Acids Res.* **43**, 1–10.
155. Gibson, D. 2009. One-step enzymatic assembly of DNA molecules up to several hundred kilobases in size. *Protoc. Exch.* 1–3.
156. Clark, D. P. & Pazdernik, N. J. 2009. Improving protein secretion. in Elsevier Academic Press, Burlington, MA. 312–315.
157. Cheng, Q. & Fischetti, V. A. 2007. Mutagenesis of a bacteriophage lytic enzyme PlyGBS significantly increases its antibacterial activity against group B streptococci. *Appl. Microbiol. Biotechnol.* **74**, 1284–1291.
158. Proença, D., Fernandes, S., Leandro, C., Silva, F. A., Santos, S., Lopes, F., Mato, R., Cavaco-Silva, P., Pimentel, M. & São-José, C. 2012. Phage endolysins with broad antimicrobial activity against Enterococcus faecalis clinical strains. *Microb. Drug Resist.* **18**, 322–332.
159. Guan, R., Roychowdhury, A., Ember, B., Kumar, S., Boons, G. J. & Mariuzza, R. A. 2004. Structural basis for peptidoglycan binding by peptidoglycan recognition proteins. *Proc. Natl. Acad. Sci. U. S. A.* **101**, 17168–17173.
160. Zhang, L., Li, D., Li, X., Hu, L., Cheng, M., Xia, F., Gong, P., Wang, B., Ge, J., Zhang, H., Cai, R., Wang, Y., Sun, C., Feng, X., Lei, L., Han, W. & Gu, J. 2016. LysGH15 kills Staphylococcus aureus without being affected by the humoral immune response or inducing inflammation. *Sci. Rep.* **6**, 1–9.
161. Liu, Q., Xun, G. & Feng, Y. 2019. The state-of-the-art strategies of protein engineering for enzyme stabilization. *Biotechnol. Adv.* **37**, 530–537.
162. Kosciolek, T. & Jones, D. T. 2016. Accurate contact predictions using covariation techniques and machine learning. *Proteins Struct. Funct. Bioinforma.* **84**, 145–151.
163. Marks, D. S., Colwell, L. J., Sheridan, R., Hopf, T. A., Pagnani, A., Zecchina, R. & Sander, C. 2011. Protein 3D structure computed from evolutionary sequence variation. *PLoS One* **6**,
164. Broendum, S. S., Buckle, A. M. & McGowan, S. 2018. Catalytic diversity and cell wall binding repeats in the phage-encoded endolysins. *Mol. Microbiol.* **110**, 879–896.
165. Woldring, D. R., Holec, P. V., Stern, L. A., Du, Y. & Hackel, B. J. 2017. A Gradient of Sitewise Diversity Promotes Evolutionary Fitness for Binder Discovery in a Three-Helix Bundle Protein Scaffold. *Biochemistry* **56**, 1656–1671.
166. Lepore, C. J., Kim, W., Thomas, J., Visi, D. & Talkington, K. 2019. *Longitudinal Analysis of the Antibiotics Clinical Pipeline*. Washington, D.C.
167. Kosikowska, P. & Lesner, A. 2016. Antimicrobial peptides (AMPs) as drug candidates: a patent review (2003–2015). **26**,
168. Czaplewski, L., Bax, R., Clokie, M., Dawson, M., Fairhead, H., Fischetti, V. A., Foster, S., Gilmore, B. F., Hancock, R. E. W., Harper, D., Henderson, I. R., Hilpert, K., Jones, B. V., Kadioglu, A., Knowles, D., Ólafsdóttir, S.,

Payne, D., Projan, S., Shaunak, S., *et al.* 2016. Alternatives to antibiotics-a pipeline portfolio review. *Lancet Infect. Dis.* **16**, 239–251.
169. Golinski, A. W., Holec, P. V., Mischler, K. M. & Hackel, B. J. 2019. Biophysical characterization platform informs protein scaffold evolvability. *ACS Comb. Sci.*
170. Sarkisyan, K. S., Bolotin, D. A., Meer, M. V., Usmanova, D. R., Mishin, A. S., Sharonov, G. V., Ivankov, D. N., Bozhanova, N. G., Baranov, M. S., Soylemez, O., Bogatyreva, N. S., Vlasov, P. K., Egorov, E. S., Logacheva, M. D., Kondrashov, A. S., Chudakov, D. M., Putintseva, E. V., Mamedov, I. Z., Tawfik, D. S., *et al.* 2016. Local fitness landscape of the green fluorescent protein. *Nature* **533**, 397–401.
171. Fellouse, F. A., Esaki, K., Birtalan, S., Raptis, D., Cancasci, V. J., Koide, A., Jhurani, P., Vasser, M., Wiesmann, C., Kossiakoff, A. A., Koide, S. & Sidhu, S. S. 2007. High-throughput Generation of Synthetic Antibodies from Highly Functional Minimalist Phage-displayed Libraries. *J. Mol. Biol.* **373**, 924–940.
172. Araya, C. L., Fowler, D. M., Chen, W., Muniez, I., Kelly, J. W. & Fields, S. 2012. A fundamental protein property, thermodynamic stability, revealed solely from large-scale measurements of protein function. *Proc. Natl. Acad. Sci. U. S. A.* **109**, 16858–16863.
173. Ren, H., Li, J., Zhang, N., Hu, L. A., Ma, Y., Tagari, P., Xu, J. & Zhang, M. Y. 2020. Function-based high-throughput screening for antibody antagonists and agonists against G protein-coupled receptors. *Commun. Biol.* **3**,
174. Gerstmans, H., Grimon, D., Gutiérrez, D., Lood, C., Rodríguez, A., van Noort, V., Lammertyn, J., Lavigne, R. & Briers, Y. 2020. A VersaTile-driven platform for rapid hit-to-lead development of engineered lysins. *Sci. Adv.* **6**, eaaz1136.
175. Kamisetty, H., Ovchinnikov, S. & Baker, D. 2013. Assessing the utility of coevolution-based residue-residue contact predictions in a sequence- and structure-rich era. *Proc. Natl. Acad. Sci. U. S. A.* **110**, 15674–15679.
176. Coucke, A., Uguzzoni, G., Oteri, F., Cocco, S., Monasson, R. & Weigt, M. 2016. Direct coevolutionary couplings reflect biophysical residue interactions in proteins. *J. Chem. Phys.* **145**,
177. Croce, G., Gueudré, T., Ruiz Cuevas, M. V., Keidel, V., Figliuzzi, M., Szurmant, H. & Weigt, M. 2019. A multi-scale coevolutionary approach to predict interactions between protein domains. *PLoS Comput. Biol.* **15**, e1006891.
178. Cong, Q., Anishchenko, I., Ovchinnikov, S. & Baker, D. 2019. Protein interaction networks revealed by proteome coevolution. *Science (80-.).* **365**, 185–189.
179. Figliuzzi, M., Jacquier, H., Schug, A., Tenaillon, O. & Weigt, M. 2016. Coevolutionary landscape inference and the context-dependence of mutations in beta-lactamase tem-1. *Mol. Biol. Evol.* **33**, 268–280.
180. Saraf, M. C., Moore, G. L. & Maranas, C. D. 2003. Using multiple

sequence correlation analysis to characterize functionally important protein regions. *Protein Eng.* **16**, 397–406.
181. Hopf, T. A., Green, A. G., Schubert, B., Mersmann, S., Schärfe, C. P. I., Ingraham, J. B., Toth-Petroczy, A., Brock, K., Riesselman, A. J., Palmedo, P., Kang, C., Sheridan, R., Draizen, E. J., Dallago, C., Sander, C. & Marks, D. S. 2019. The EVcouplings Python framework for coevolutionary sequence analysis. *Bioinformatics* **35**, 1582–1584.
182. Wang, I. N., Smith, D. L. & Young, R. 2000. Holins: The protein clocks of bacteriophage infections. *Annu. Rev. Microbiol.* **54**, 799–825.
183. Loessner, M. J., Kramer, K., Ebel, F. & Scherer, S. 2002. C-terminal domains of Listeria monocytogenes bacteriophage murein hydrolases determine specific recognition and high-affinity binding to bacterial cell wall carbohydrates. *Mol. Microbiol.* **44**, 335–349.
184. Loessner, M. J. & Wendlinger, G. 1995. Heterogeneous endolysins in Listeria monocytogenes bacteriophages : a new class of enzymes and evidence for conserved holin genes within the siphoviral lysis cassettes. **16**, 1231–1241.
185. Freudl, R. 2018. Signal peptides for recombinant protein secretion in bacterial expression systems. *Microb. Cell Fact.* **17**, 1–10.
186. Yoon, S. H., Kim, S. K. & Kim, J. F. 2010. Secretory production of recombinant proteins in Escherichia coli. *Recent Pat. Biotechnol.* **4**, 23–29.
187. Thie, H., Schirrmann, T., Paschke, M., Dübel, S. & Hust, M. 2008. SRP and Sec pathway leader peptides for antibody phage display and antibody fragment production in E. coli. *N. Biotechnol.* **25**, 49–54.
188. Pechsrichuang, P., Songsiriritthigul, C., Haltrich, D., Roytrakul, S., Namvijtr, P., Bonaparte, N. & Yamabhai, M. 2016. OmpA signal peptide leads to heterogenous secretion of B. subtilis chitosanase enzyme from E. coli expression system. *Springerplus* **5**,
189. Zhang, L. H., Fath, M. J., Mahanty, H. K., Tai, P. C. & Kolter, R. 1995. Genetic Analysis of the Colicin V Secretion Pathway. *Genetics* **141**, 25–32.
190. Palmer, T. & Berks, B. C. 2012. The twin-arginine translocation (Tat) protein export pathway. *Nat. Rev. Microbiol.* **10**, 483–496.
191. Frain, K. M., Robinson, C. & van Dijl, J. M. 2019. Transport of Folded Proteins by the Tat System. *Protein J.* **38**, 377–388.
192. Goldenzweig, A., Goldsmith, M., Hill, S. E., Gertman, O., Laurino, P., Ashani, Y., Dym, O., Unger, T., Albeck, S., Prilusky, J., Lieberman, R. L., Aharoni, A., Silman, I., Sussman, J. L., Tawfik, D. S. & Fleishman, S. J. 2016. Automated Structure- and Sequence-Based Design of Proteins for High Bacterial Expression and Stability. *Mol. Cell* **63**, 337–346.
193. Senior, A. W., Evans, R., Jumper, J., Kirkpatrick, J., Sifre, L., Green, T., Qin, C., Žídek, A., Nelson, A. W. R., Bridgland, A., Penedones, H., Petersen, S., Simonyan, K., Crossan, S., Kohli, P., Jones, D. T., Silver, D., Kavukcuoglu, K. & Hassabis, D. 2020. Improved protein structure prediction using potentials from deep learning. *Nature* **577**,
194. Traxlmayr, M. W. & Obinger, C. 2012. Directed evolution of proteins for

increased stability and expression using yeast display. *Arch. Biochem. Biophys.* **526**, 174–180.
195. Chun, J., Bai, J. & Ryu, S. 2020. Yeast Surface Display System for Facilitated Production and Application of Phage Endolysin. *ACS Comb. Sci.* **9**, 508–516.
196. Woldring, D. R., Holec, P. V, Zhou, H. & Hackel, B. J. 2015. High-Throughput Ligand Discovery Reveals a Sitewise Gradient of Diversity in Broadly Evolved Hydrophilic Fibronectin Domains. *PLoS One* 1–29.
197. Boratyn, G. M., Camacho, C., Cooper, P. S., Coulouris, G., Fong, A., Ma, N., Madden, T. L., Matten, W. T., McGinnis, S. D., Merezhuk, Y., Raytselis, Y., Sayers, E. W., Tao, T., Ye, J. & Zaretskaya, I. 2013. BLAST: a more efficient report with usability improvements. *Nucleic Acids Res.* **41**, 29–33.
198. Sayers, E. W., Beck, J., Bolton, E. E., Bourexis, D., Brister, J. R., Canese, K., Comeau, D. C., Funk, K., Kim, S., Klimke, W., Marchler-Bauer, A., Landrum, M., Lathrop, S., Lu, Z., Madden, T. L., O'Leary, N., Phan, L., Rangwala, S. H., Schneider, V. A., *et al.* 2021. Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.* **49**, D10–D17.
199. Cadwell, R. C. & Joyce, G. F. 1992. Randomization of genes by PCR mutagenesis. *Genome Res.* **2**, 28–33.
200. Alley, E. C., Khimulya, G., Biswas, S., AlQuraishi, M. & Church, G. M. 2019. Unified rational protein engineering with sequence-based deep representation learning. *Nat. Methods* **16**, 1315–1322.
201. Kristich, C. J., Nguyen, V. T., Le, T., Barnes, A. M. T., Grindle, S. & Dunny, G. M. 2008. Development and use of an efficient system for random mariner transposon mutagenesis to identify novel genetic determinants of biofilm formation in the core Enterococcus faecalis genome. *Appl. Environ. Microbiol.* **74**, 3377–3386.
202. Dale, J. L., Beckman, K. B., Willett, J. L. E., Nilson, J. L., Palani, N. P., Baller, J. A., Hauge, A., Gohl, D. M., Erickson, R., Manias, D. A., Sadowsky, M. J. & Dunny, G. M. 2018. Comprehensive Functional Analysis of the Enterococcus faecalis Core Genome Using an Ordered, Sequence-Defined Collection of Insertional Mutations in Strain OG1RF. *mSystems* 1–19.
203. Chatterjee, A., Willett, J. L. E., Dunny, G. M. & Duerkop, B. A. 2021. Phage infection and sub-lethal antibiotic exposure mediate Enterococcus faecalis type VII secretion system dependent inhibition of bystander bacteria. *PLoS Genet.* **17**, 1–26.

Chapter 7 – Appendices

7.1 Functional investigation of *Enterococcus faecalis* response to antimicrobial protein treatment

Daniel T. Tresnak, Julia L. E. Willett, Ethan B. Robertson, Lucy M. Kwiatkowski, Gary M. Dunny, Benjamin J. Hackel

The work contained in this chapter, including analysis, experimental design and implementation, data interpretation, and composition of text was conducted by D.T.T., J.L.E.W., E.R.R., L.M.K., G.M.D., and B.J.H. In particular, D.T.T. contributed in part to generation of the original idea, production and testing of antimicrobial proteins, assembly of DNA constructs, analysis of results, and manuscript preparation. The work presented here is ongoing and will be submitted for publication upon completion of experiments and analysis outlined at the end of Appendix 7.1.

7.1.1 Motivation and Results

In Chapter 2, we demonstrated that factors beyond mannose phosphotransferase (manPTS) receptors impact enterococcal susceptibility to class IIa bacteriocins. Based on these findings, we were interested in further exploring enterococcal responses to treatment with class IIa bacteriocins. We hypothesized that treatment of a previously constructed library of *E. faecalis* OG1RF mutants containing single transposon insertions throughout the genome with class IIa bacteriocins would enable identification of genetic determinants impacting resistance, similar to a previous investigation of bile resistance[201,202]. These transposon insertions effectively “knock out” various genes by causing frameshift mutations, resulting in expression of incorrect or truncated gene products. By deep sequencing of the

initial library and the library after treatment with a class IIa bacteriocin, we can identify gene knockouts which provide a functional benefit or detriment, thus elucidating determinants impacting resistance. In addition to this primary goal, it has been documented that enterococcal biofilms respond and restructure when treated with antibiotics or other stressors which target the cell envelope[203]. Given class IIa bacteriocin targeting and mode of action (pore formation in the cell membrane), we were curious whether class IIa bacteriocins also induced such a response in biofilms and would assess this via analysis of biofilm growth and architecture following treatment with the peptides. To test these hypotheses, hiracin JM79 (HJ79) and divercin V41 (DV41) were selected from work in Chapter 2 as especially potent class IIa bacteriocins.

We first sought to quantify the inhibitory activity of HJ79 and DV41 against *E. faecalis* OG1RF to ensure they possessed sufficient potency to induce a bacterial response. Both peptides were individually produced by expression in *L. lactis* and concentrated via ammonium sulfate (AS) precipitation as detailed in Chapter 2, serially diluted, and tested for inhibitory activity to *E. faecalis* OG1RF and *E. faecium* strain 8E9. *E. faecium* 8E9 was included in this evaluation as a positive control, as we demonstrated HJ79 and DV41 inhibitory activity against this strain in Chapter 2. Notably, OG1RF shows significantly lower susceptibility to class IIa bacteriocins than 8E9, similar to prior results seen for treatment of other *E. faecalis* strains in Chapter 2 (Figure 7.1). While DV41 only shows mild inhibitory

activity to OG1RF, HJ79 displays noticeably higher growth inhibition, supporting its use for treating the transposon insertion library.

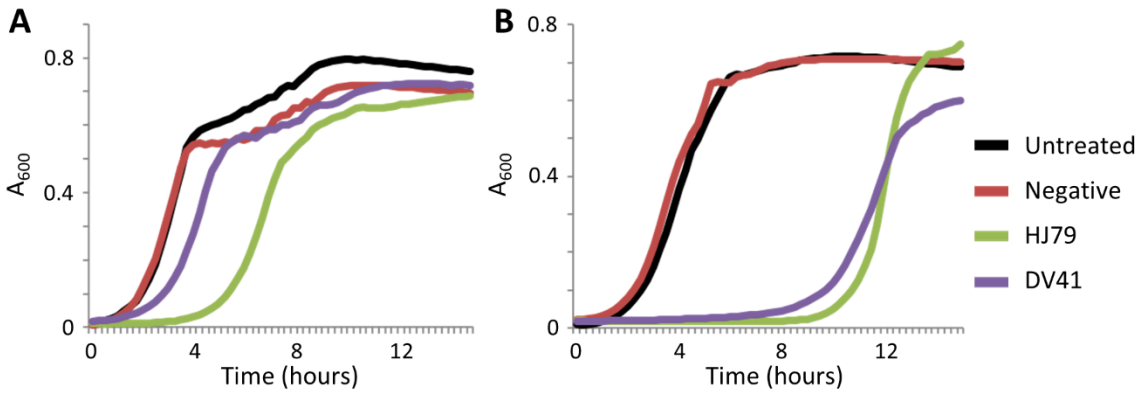


Figure 7.1 Inhibitory activity of HJ79 and DV41 against enterococcus strains. (AB) Cell density dynamics ("growth curves") of OG1RF (A) and 8E9 (B) when treated with HJ79 and DV41. Cell density is measured via absorbance at 600 nm for cells grown with an eightfold dilution of each peptide or a comparable volume of negative AS solution (negative) or growth media (untreated).

Following confirmation of HJ79 inhibitory activity against OG1RF, we wanted to ensure that transposon insertions in known class IIa bacteriocin targets resulted in resistance. It is well documented that class IIa bacteriocins interact with the manPTS EIIC and EIID subunits to achieve pore formation. Thus, we hypothesized that transposon insertions in these subunits would result in resistance to class IIa bacteriocins. Further, as manPTS is responsible for mannose uptake, we suspected knockout of these subunits would also hinder growth on media containing mannose as the only nutrient source. To test this, we identified two OG1RF mutants with insertions in either manPTS EIIC or EIID and monitored their growth on minimal media containing mannose or glucose as the only nutrient source. We found that both mutants grew significantly worse when

mannose was the only nutrient source as assessed by OD_{600} measurements after 4 hours of growth (Figure 7.2, p-value = 1.1×10^{-5} and 0.002 for the EIIC and EIID mutants, respectively). We also quantified whether addition of HJ79 resulted in growth inhibition of these mutants. Contrary to parental OG1RF (Figure 7.1), the transposon mutants displayed clear resistance to HJ79 with nearly identical OD_{600} values at 4 hours (Figure 7.2, p-value = 0.89 and 0.32 for the EIIC and EIID mutants, respectively). These results confirm the ability of the transposon insertion library to assess genetic determinants which impact resistance to class IIa bacteriocins.

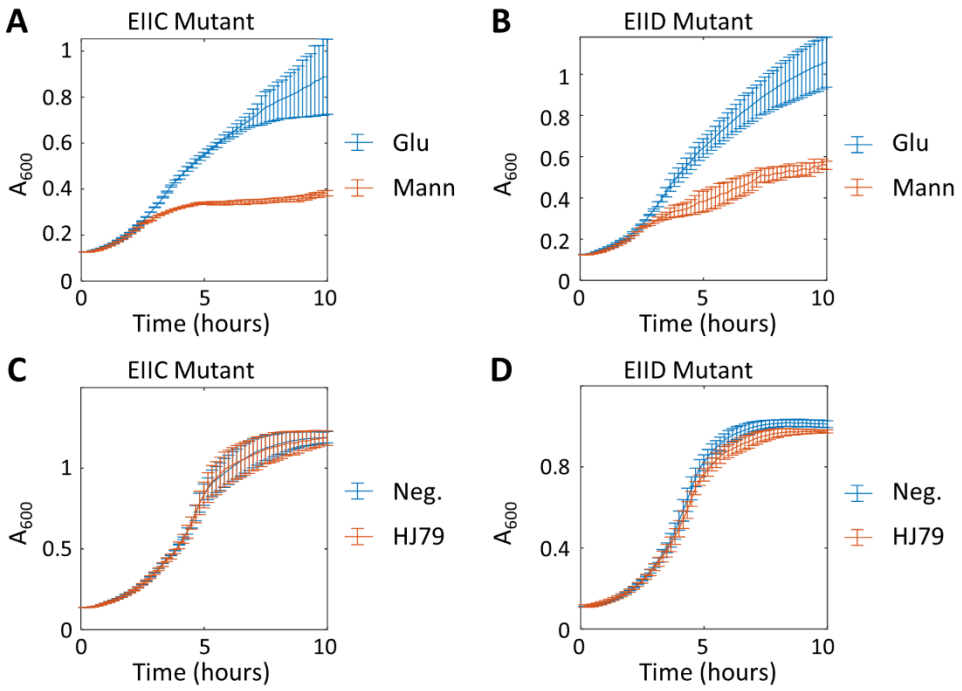


Figure 7.2 Transposon insertions oblate manPTS function in EIIC and EIID mutants.

(AB) Growth curves of manPTS EIIC (A) and EIID (B) mutants when grown on minimal media containing glucose or mannose as the only nutrient. (CD) Growth

curves of manPTS EIIC (C) and EIID (D) mutants when grown in the presence of HJ79 or negative AS precipitate solution. Growth was measured via monitoring the absorbance at 600 nm (A_{600}) for several hours. Error bars represent the standard deviation across triplicate measurements. Data shown are for mutants transformed with empty pCIE plasmids for comparison to Figures 7.3-7.4. Growth curves of mutants prior to transformation showed similar results (Data not shown).

To assess whether transposon insertion-induced knockout of the manPTS receptors was recoverable, we inserted genes encoding either the parental OG1RF or 8E9 manPTS EIIC or EIID receptors into the nisin-inducible pCIE expression vector and transformed them into the EIIC and EIID transposon mutants of OG1RF. To assess recovery of manPTS function, we induced expression of the manPTS EIIC or EIID receptor and repeated growth experiments on minimal media containing glucose or mannose. We hypothesized that only strains displaying full recovery of manPTS function would be able to efficiently grow on mannose, whereas all strains should grow well on glucose. Analysis of strain growth showed that induced expression of the OG1RF manPTS EIID subunit fully recovered normal growth behavior, but expression of the OG1RF manPTS EIIC subunit did not (Figure 7.3, p-value = 0.01 and 0.52 for induced expression of EIIC and EIID subunits in mutants compared to uninduced controls, respectively). Additionally, expression of the 8E9 manPTS EIIC subunit did not recover function, while the 8E9 manPTS EIID subunit marginally improved growth rate of the EIID OG1RF mutant (p-value = 0.002 and 0.24 for induced expression of 8E9 EIIC and EIID subunits compared to uninduced controls, respectively). These results suggest that these subunits are sufficiently distinct between the two species,

though the mild recovery of function from expression of the 8E9 EIID subunit may suggest OG1RF manPTS subunits are more efficient that those of 8E9.

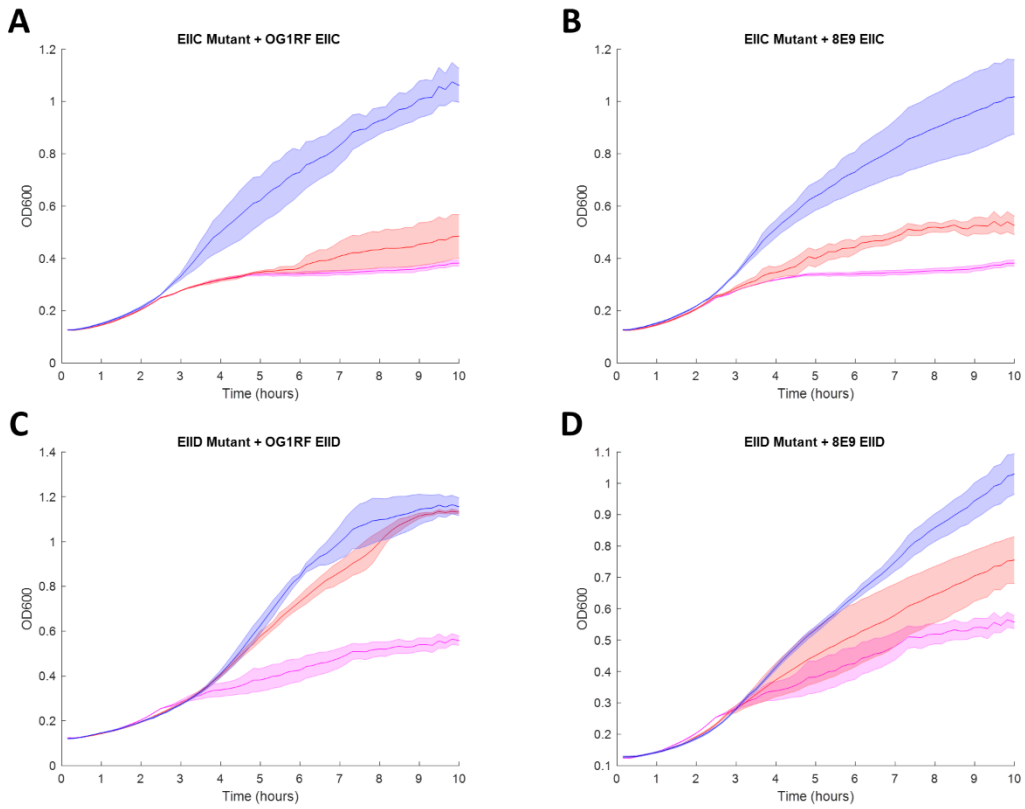


Figure 7.3 Expression of manPTS EIID subunits recovers function via growth on mannose.

(AB) Growth curves of manPTS EIIC mutant expressing the OG1RF (A) and 8E9 (B) manPTS EIIC subunits when grown on minimal media containing glucose or mannose as the only nutrient. (CD) Growth curves of manPTS EIID mutant expressing the OG1RF (C) and 8E9 (D) manPTS EIID subunits when grown on minimal media containing glucose or mannose as the only nutrient. Red curves represent cells grown on mannose, blue curves represent cells grown on glucose, and magenta curves represent cells grown on mannose with no expression of any manPTS subunits (same data as mannose growth curves in Figure 7.2 AB). Growth was measured via monitoring the absorbance at 600 nm (A_{600}) for several hours. Error bars represent the standard deviation across triplicate measurements.

As a second assessment of manPTS function, we tested whether induced expression of manPTS receptors resulted in recovery of susceptibility to HJ79 and observed similar trends, as only expression of OG1RF manPTS EIID in the EIID mutant led to growth inhibition (Figure 7.4, p-value = 0.49 and 7.9×10^{-4} for induced expression of OG1RF EIIC and EIID subunits compared to uninduced controls, respectively). Expression of 8E9 manPTS subunits did not significantly impact susceptibility to HJ79, with p-values = 0.90 and 0.17 for expression of the EIIC and EIID subunits, respectively. While it is surprising that induced expression of both OG1RF and 8E9 manPTS EIIC did not recover manPTS function, it is possible that transposon insertion in manPTS EIIC also caused a frameshift in expression of the EIID subunit, since it is downstream in the genome. This would result in errors in both manPTS EIIC and EIID subunits. Thus, induced expression of only the manPTS EIIC subunit would then not be expected to regain full function. Ongoing work will assess this hypothesis by inducing expression of both manPTS subunits in the EIIC mutant.

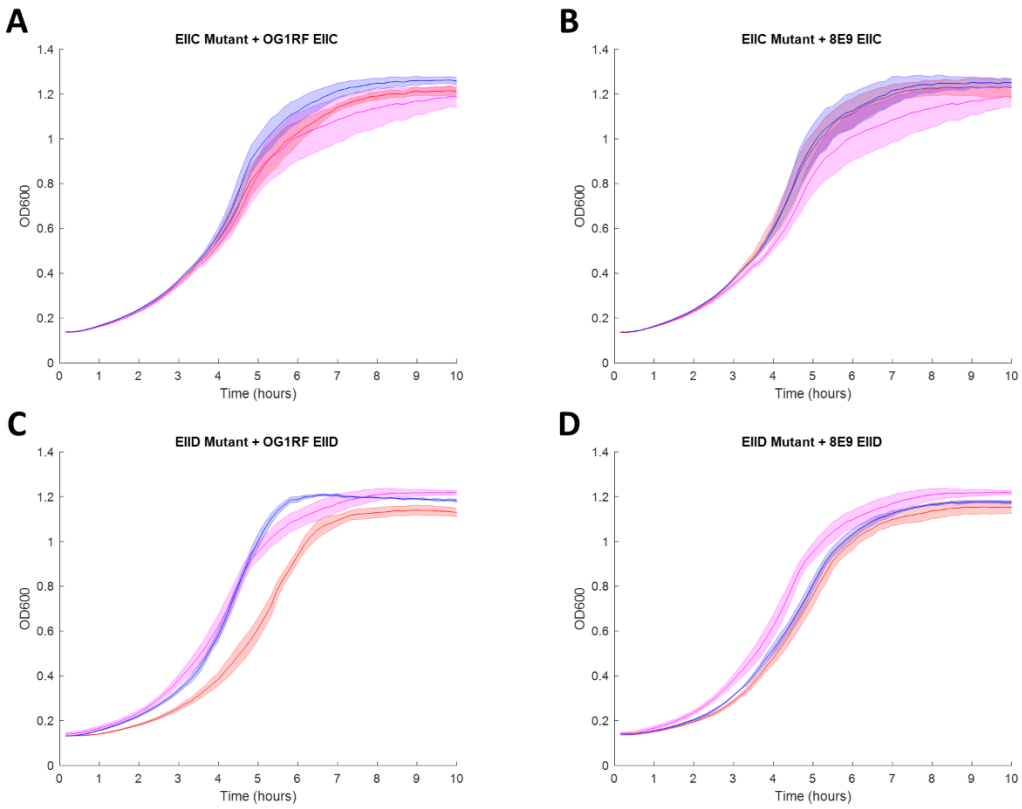


Figure 7.4 Only expression of OG1RF manPTS EIID subunits recovers function via growth inhibition by HJ79.

(AB) Growth curves of transposon insertion EIIC mutants expressing the OG1RF (A) and 8E9 (B) manPTS EIIC subunits when grown in the presence of HJ79. (CD) Growth curves of transposon insertion EIID mutants expressing the OG1RF (C) and 8E9 (D) manPTS EIID subunits when grown in the presence of HJ79. Red curves represent cells grown in the presence of HJ79, blue curves represent cells grown in the presence of negative, and magenta curves represent cells grown in the presence of HJ79 with no expression of any manPTS subunits (same data as HJ79 growth curves in Figure 7.2 CD). Growth was measured via monitoring the absorbance at 600 nm (A_{600}) for several hours. Error bars represent the standard deviation across triplicate measurements.

Ongoing work in this space encompasses two aims: (1) screening the transposon insertion library in a competition experiment when treated with HJ79; and (2) assessing the impact of HJ79 on biofilm formation and morphology. We expect that completion of aim 1 will help identify genetic determinants which impact

enterococcal susceptibility to class IIa bacteriocins while completion of Aim 2 will provide mechanistic insights into the enterococcal stress response which causes biofilm restructuring.

7.1.2 Materials and Methods

7.1.2.1 Bacterial cultures

E. coli cells were grown in lysogeny broth (LB; Fisher BioReagents) which contained 1.6% (vol/vol) agar in the case of solid-phase growth at 37 °C, with shaking of liquid cultures at 250 rpm. When specified, LB cultures were supplemented with 50 µg/mL kanamycin or 10 µg/mL tetracycline. *E. faecium* strain 8E9 (generously provided by Prof. Patricia Ferrieri of the University of Minnesota) and *E. faecalis* strain OG1RF (generously provided by Prof. Gary Dunny of the University of Minnesota) were grown in liquid brain heart infusion (BHI, include brand supplier) medium at 37 °C with shaking at 250 rpm or on solid BHI which contained 1.6% (vol/vol) agar at 37 °C unless otherwise noted. When noted, minimal M9 media was used supplemented with 0.4% of either glucose or mannose (10 mL 10X M9 salts, 1 mL 0.1 M $MgSO_4$, 1 mL 0.01 M $CaCl_2$, 84 mL M9 medium, 2 mL 20% glucose/mannose). M9 medium consisted of 3 g yeast extract and 10 g casamino acids in 860 mL dH_20. 10X M9 salts consisted of 60 g Na_2HPO_4, 30 g KH_2PO_4, 5 g NaCl, and 10 g NH_4Cl in 1 L dH_2O.

7.1.2.2 Production of hiracin JM79 and divercin V41

Glycerol stocks of *L. lactis* cultures were obtained from work completed in Chapter 2 and concentrated samples of the peptides HJ79 and DV41 were generated

following similar protocols described in Chapter 2. To produce individual AMP solutions, 40 mL of BHI medium was inoculated from glycerol stocks of peptide-producing or negative control, empty pNZC- containing *L. lactis* and incubated under stationary conditions overnight at 30 °C. The culture was centrifuged at 3,500 × g for 5 minutes, the supernatant was discarded, and the cells were resuspended in an equal volume of fresh BHI medium. The culture was incubated under stationary conditions for 4 hours at 30 °C. Following incubation, the culture was centrifuged at 3,500 × g for 5 minutes, and the supernatant was sterile filtered. Ammonium sulfate (AS) was added to the supernatant at 45% (wt/vol) to achieve a 70% saturated AS solution, which was rotated for ~18 hours at 4°C. The AS solution was centrifuged for 10 min at 11,000 × g, and the pellet was resuspended in 1 mL of ultrapure water and heat sterilized at 98 °C for 10 minutes. The resulting AS precipitation solutions were stored at –20 °C until further use.

7.1.2.3 96-well plate growth assays

Cultures of the *E. faecium* or *E. faecalis* indicator strains were grown overnight in BHI medium at 37 °C with shaking at 250 rpm. Overnight cultures were diluted 1000-fold in fresh BHI media and 150 µL of cell dilution was plated in each well of a clear 96-well plate. AS precipitation solutions were two-fold serially diluted and 50 µL of each AS dilution was added to wells of the indicator strain. Growth of indicator cells was then monitored via measuring the optical density at 600 nm (OD_{600}) in a Biotek Synergy H1 plate reader for 18-24 hours with shaking at 37 °C.

7.1.2.4 Construction of OG1RF manPTS expression plasmids

DNA stocks of the cCF10-inducible pCIE expression vector were obtained from previous work completed by the Dunny Lab. pCIE vector was digested with BamHI-HF and NheI-HF restriction enzymes (New England Biolabs) according to manufacturer's protocols and isolated via gel electrophoresis. Genes encoding for the parental manPTS EIIC or EIID receptors, from both *E. faecalis* OG1RF and *E. faecium* 8E9, were amplified with primers designed from sequenced genomes with sufficient overlap for insertion into pCIE (primers included in Table 7.1). Individual manPTS EIIC or EIID genes were inserted into pCIE via HiFi DNA Assembly (New England Biolabs) and transformed into *E. Coli* (New England Biolabs, C2987H) grown on solid LB supplemented with tetracycline. Individual colonies were grown overnight and constructs were sequence verified for further use. Following sequence confirmation, constructs were transformed into parental *E. faecalis* OG1RF (for empty pCIE vector control) or OG1RF mutants with transposons inserted into reading frame 10020 (for manPTS EIIC genes) or 10021(for manPTS EIID genes). The reading frame 10020 encodes for manPTS EIIC and the reading frame 10021 encodes for manPTS EIID. These specific reading frame transposon-insertion mutants are referred to as the manPTS EIIC and manPTS EIID mutants in the main text. Empty vector controls were also transformed into transposon insertion mutants. Individual colonies of transformed *E. faecalis* OG1RF strains were grown overnight and glycerol stocks were created and used for all experiments.

Table 7.1 Primers used for pCIE-manPTS construction

Primer Name	DNA Sequence
OG1RF manPTS EIIC Fwd	TTTTGTTGTCTGTTGGGGGATCCTCGCCAAATACAAATCAAAATAGG
OG1RF manPTS EIIC Rev	ACATGGTTACTTCTTTAGGGCTAGCTTTCTTTGCTCCTCCTTCAG
OG1RF manPTS EIID Fwd	TTTTGTTGTCTGTTGGGGGATCCAACGACTACTAATTCTGAAGG
OG1RF manPTS EIID Rev	ACATGGTTACTTCTTTAGGGCTAGCTTCGTCATTCTTATAATAAGCCG
8E9 manPTS EIIC Fwd	TTTTGTTGTCTGTTGGGGGATCCATGTCTATTATTTCAATAATTTTAGTCG
8E9 manPTS EIIC Rev	ACATGGTTACTTCTTTAGGGCTAGCTTAATAGTCATTCAAAATGTCGCCTA
8E9 manPTS EIID Fwd	TTTTGTTGTCTGTTGGGGGATCCATGGCAGAAGAAAAAATCAAATTAT
8E9 manPTS EIID Rev	ACATGGTTACTTCTTTAGGGCTAGCTTATAAAAGTCCGATGACGTGTC

7.1.2.5 ManPTS Functional Analysis Experiments

OG1RF strains containing pCIE constructs were grown overnight in liquid BHI supplemented with tetracycline. Overnight cultures were diluted 1000-fold in fresh BHI or minimal M9 media supplemented with glucose or mannose. Cultures diluted in mannose or glucose were grown as uninduced controls or induced with 25 ng/μL cCF10 and growth was monitored in a plate reader as described previously. Cultures diluted in BHI were plated with a 32-fold dilution of either negative control or HJ79 AS precipitate solution as uninduced controls or induced with 25 ng/μL cCF10 and growth was monitored as described previously.

7.1.2.6 Statistical Analysis

All statistical tests between growth conditions were done using *ttest2* function in Matlab. Comparisons were made between triplicate OD_{600} values from the first measurement after 4 hours of growth in the plate reader experiments.