

Protease Engineering to Enable Noninvasive Disease Detection

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

Brian Michael Mikolajczyk

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

Benjamin J. Hackel

March 2020

## **Acknowledgements**

---

I want to first thank my adviser, Benjamin Hackel, for the years of guidance and mentorship. I can recall numerous times when specific experiments, and sometimes entire projects, were not quite panning out as planned. While these roadblocks were oftentimes frustrating, to put it mildly, Ben's calm, cool and collected demeanor and approach was always a reassuring presence when navigating choppy waters. I greatly appreciated his willingness to meet with me as quickly as possible to address these obstacles, and I always, without exception, left our meetings much more optimistic about the path forward. Ben pushed me to think critically and consider all angles when designing experiments and analyzing data, and I feel fortunate to have had an advisor who combines his insight, wisdom, work ethic, creativity and expertise with a strong sense of ethics, purpose and leadership. The opportunity to learn from Ben these past five and a half years has made me a more complete engineer, scientist and person. Finally, Ben cared for my professional development but also for me as a person, and his sincere concern for the health and well-being for the members of his lab, and their family members, was always deeply appreciated.

I also want to acknowledge the members of the Hackel Lab that made daily life (when I was not working nights) so enjoyable. Danny, Brett and Larry were the exemplary elder statesmen of the lab when I joined, and I always valued their friendship, as well as the scientific and professional wisdom they imparted. Others labmates during my time here, such as Max, Sadie, Seth, Justin and Andrew, were likewise appreciated. I would be remiss if I did not also mention the delightful non-scientific discussions we had in lab, whether that pertained to culture or politics or current events. The relentlessly

positive, fun, open, intellectually engaging and collaborative environment fostered by these individuals, in addition to the lab's scientific focus and the aforementioned points regarding Ben, made joining this lab years ago a no-brainer, and I consider myself incredibly fortunate to have been welcomed with open arms. This lab culture is being propagated moving forward with the likes of Alex, Patrick, Suvam, Abby and everyone else, and I look forward to hearing about the exciting research being conducted by the lab in the future.

The MSTP leadership at the University of Minnesota, particularly Yoji and Bryce, have additionally been instrumental in helping me reach this milestone. Completing a PhD in chemical engineering as part of the MD/PhD training program at this institution presented unique obstacles, especially so for someone like myself who did not major in chemical engineering as an undergrad. Yoji worked with me from the beginning to make this work, and I want to thank both him and Bryce for their unwavering support and confidence in me. I will be eternally grateful for their investment in terms of time and energy, and I want to thank both Yoji and Bryce and the entire MSTP leadership for all of the insight and reassurance over the years.

To my parents and siblings in Chicago, thank you for your support and guidance during my years in Minneapolis. I am blessed to have been raised by Michael and Rebecca Mikolajczyk, who taught me to treat others with kindness and respect and to think boldly, as well as the importance of family and the value of grit and determination in accomplishing goals.

To my incredible wife Kayley, thank you for your love and support during the last five and a half years, and as long as we have known each other. I am not quite sure how

this family would function without your preternatural ability to juggle everything, but I am thankful and deeply appreciative that you do everything that you do. I would not have been able to get to this point without your dedication, diligence, determination, kindness, empathy and thoughtfulness. We married almost eight years ago, right before moving to Minneapolis to start my MD/PhD training. It has been quite a ride thus far, and I look forward with great enthusiasm to continuing our journey together. To my two amazing boys, Charlie and Clark, thank you for your unceasing ability to make me laugh and smile. While there have been some chaotic and sleepless moments during this journey, your constant joy and zest for life has been, and will continue to be, an inspiration to me to always appreciate life and see the sunny side.

## **Abstract**

---

Proteases are proteolytic enzymes with a wide range of industrial, biotechnological, and medical applications. Due to their importance, proteases have been the subject of many attempts to engineer improved performance, but campaigns to improve activity via directed evolution have been hindered by inefficient analytical techniques and insufficient understanding of sequence-function relationships. Tobacco etch virus protease (TEVp) has mostly been engineered for attributes other than catalytic activity, and most of the past efforts have employed random mutagenesis methods such as error-prone PCR as opposed to targeted mutagenesis. We developed a novel and seemingly generalizable yeast surface display approach that co-displays protease mutants adjacent to substrate on the same Aga2 anchor protein. Enhanced activity mutants are identified by protease cleavage of tethered substrate removing an epitope tag, which empowers flow cytometric isolation of cells with a decrease in anti-epitope antibody signal. This platform was shown to quantitatively differentiate catalytic activity at the single-cell level for TEVp and sortase A.

We leveraged this display platform to perform high throughput screens on seven structure-based active site combinatorial libraries created via saturation mutagenesis, and then screened a second-generation library combining the resultant beneficial mutations. Deep sequencing of functional mutants elucidated sequence-function relationships across 34 sites and identified improved multi-mutants. Clonal analysis of a host of recombinant TEVp multi-mutants with purified substrate demonstrated up to 2.9-fold improvement in catalytic efficiency, generally via decreased  $K_M$ . The novel yeast surface protease/substrate co-display system and the insights gleaned on rational active site

library design and the TEVp sequence-function map will aid future protease engineering efforts, and the collection of improved multi-mutants will benefit the biotechnological community in utilizing TEVp in its multitude of applications.

One class of application for engineered proteases is physiological release of diagnostic or therapeutic moieties. We introduced a novel extension of synthetic reporters to noninvasively detect abnormal receptor expression. Synthetic reporters have been demonstrated to noninvasively detect a host of diseases via nanoparticles conjugated to reporters via substrate linkers; biomarkers are generated dependent upon a disease-specific enzyme and filtered into the urine. This approach is limited by its reliance on upregulation of disease-specific proteases, but many diseases are characterized by abnormal expression of cell-surface receptors. The new approach harnesses ligand-enzyme fusion proteins to impart exogenous enzymatic activity to tissue with aberrant receptor expression. A mathematical model for epidermal growth factor receptor (EGFR) tumor xenografts in mice demonstrated feasibility of this approach with TEVp-based fusions, suggesting detection down to tumor diameters of 0.28 mm at standard substrate concentrations. Multiple fusions were produced using different enzymes, ligands, and orientations, and binding and catalytic activity was generally well preserved, indicating a modular fusion framework. Demonstrating feasibility with anti-EGFR TEVp-based fusions in an *in vitro* cellular assay was not consistently successful. However, the following limitations were identified for improvement: high substrate lability, and insufficient fusion-specific product generation due to inadequate catalytic activity – which would motivate protease engineering – or suboptimal fusion linker design that resulted in ineffective projection of receptor-bound fusion’s enzyme component to

engage soluble substrate. Together, this work introduced a novel extension of the synthetic reporter concept to quantify receptor expression, and we have demonstrated theoretical *in vivo* feasibility as well as empirical functionality of the required ligand-enzyme fusions. We have also introduced a novel display platform that can be harnessed for screening combinatorial protease libraries to find mutants with improved catalytic efficiency, which will aid the synthetic reporter approach.

## **Table of Contents**

---

Acknowledgements.....	i
Abstract.....	iv
Table of Contents.....	vii
List of Tables.....	ix
List of Figures.....	x
Chapter 1 – Introduction.....	1
1.1 Engineering protease activity.....	1
1.2 Genotype-phenotype linkage strategy for selecting improved protease activity.....	2
1.3 Tools to constrain diversification of sequence space.....	4
1.4 Applying improved proteases to synthetic urinary biomarkers to detect aberrant receptor expression.....	8
1.5 Summary of approach and results in this thesis.....	10
Chapter 2 - Engineering tobacco etch virus protease activity by screening rational active site combinatorial libraries using a novel yeast surface display construct.....	12
2.1 Abstract.....	12
2.2 Introduction.....	13
2.3 Methods.....	16
2.3.1 Creation of TEVp DNA constructs.....	17
2.3.2 Yeast surface display to validate display construct and optimize induction time.....	18
2.3.3 TEVp first-generation library construction.....	19
2.3.4 Fluorescence-activated cell sorting of first-generation library.....	22
2.3.5 Deep sequencing of first-generation libraries.....	23
2.3.6 TEVp second-generation library construction.....	24
2.3.7 Fluorescence-activated cell sorting of second-generation library.....	26
2.3.8 Sanger sequencing of <i>top</i> and <i>middle</i> gates from first sort of second-generation TEVp library.....	27
2.3.9 Sanger sequencing of unsorted second-generation TEVp library.....	27
2.3.10 Deep sequencing of second-generation TEVp library.....	28
2.3.11 TEVp mutant production and characterization.....	28
2.3.12 Creation of Srt7 DNA constructs.....	32
2.3.13 Estimation of solvent-accessible surface area.....	32
2.4 Results.....	33
2.4.1 Design of TEVp/substrate fusion display construct.....	33
2.4.2 Validation of display construct to differentiate enzymatic activity.....	34
2.4.3 TEVp mutational analysis.....	37
2.4.3.1 <i>First-generation TEVp library</i> .....	37
2.4.3.2 <i>Second-generation TEVp library</i> .....	42
2.4.4 TEVp mutant production and characterization from sitewise analysis.....	47
2.4.5 TEVp mutant production and characterization from epistatic analysis.....	50
2.4.6 Investigation of potential auxiliary cysteine protease activity at sites	



30-32.....	53
2.4.7 Enzyme-substrate co-display system successfully stratifies protease mutants by activity.....	54
2.4.8 Extension of the enzyme-substrate co-display system to sortase enzyme.....	55
2.5 Discussion.....	57
Chapter 3 - Applying ligand-enzyme fusion proteins to synthetic urinary biomarkers for noninvasive detection of abnormal disease-related receptor expression.....	68
3.1 Abstract.....	68
3.2 Introduction.....	69
3.3 Methods.....	80
3.3.1 Creation of fusion protein-encoding DNA constructs.....	80
3.3.2 Fusion protein production.....	81
3.3.3 Cell culturing.....	82
3.3.4 Affinity titration of fusion proteins.....	82
3.3.5 Fusion protein catalytic analysis.....	82
3.3.6 <i>In vitro</i> cell assay.....	83
3.4 Results.....	84
3.4.1 Mathematical modeling of ligand-enzyme-reporter system.....	84
3.4.2 Fusion protein production.....	88
3.4.3 Affinity titration of fusion proteins.....	89
3.4.4 Fusion protein catalytic analysis.....	90
3.4.5 <i>In vitro</i> cellular assay.....	91
3.5 Discussion.....	94
Chapter 4 – Concluding Remarks and Future Work.....	101
References.....	111
Supplemental Tables.....	122
Supplemental Figures.....	125

## List of Tables

---

Table 2-1. Summary statistics of flow cytometry-assisted TEVp activity profiling.....	36
Table 2-2. Naïve library analysis.....	38
Table 2-3. FACS data from the first-generation TEVp libraries.....	39
Table 2-4. Deep sequencing reads for unsorted and doubly sorted populations from first-generation TEVp libraries A-G.....	40
Table 2-5. List of beneficial mutations from the first-generation TEVp libraries.....	41
Table 2-6. Second-generation TEVp library design.....	43
Table 2-7. FACS data from the second-generation TEVp library.....	46
Table 2-8. Deep sequencing reads for the BIB and BBB populations from the second-generation TEVp library.....	46
Table 2-9. List of TEVp mutants produced in <i>E. coli</i> .....	47
Table 2-10. Summary of the effect on enzymatic parameters of TEVp2 single-mutant reversions to wild-type.....	50
Table 2-11. Summary statistics of flow cytometry-assisted Srt7 activity profiling.....	56
Table 2-12. Results of homologous protein sequence analysis.....	62
Table 3-1. The sequences, epitopes and affinities of EGFR-binding Fn mutants.....	79
Table 3-2. Catalytic performance of fusion proteins FnD-TEVp1 and FnNB-TEVp1.....	91
Supplemental Table 2-1. List of gBlocks and oligonucleotide primers used.....	122
Supplemental Table 3-1. List of oligonucleotide primers used for PCR amplification.....	124

## List of Figures

---

Figure 1-1. Sequence of events in directed evolution.....	4
Figure 1-2. Fitness landscape.....	5
Figure 1-3. Rough fitness landscape.....	5
Figure 1-4. Noninvasive urinary detection of disease using synthetic reporters linked to nanoparticles by a protease-sensitive linkage.....	10
Figure 2-1. Mutant library design.....	20
Figure 2-2. Schematic of library assembly via fusion PCR.....	22
Figure 2-3. Schematic of library assembly via fusion PCR for the second-generation TEVp library.....	26
Figure 2-4. Standard curve of 2-Abz.....	31
Figure 2-5. Michaelis-Menten fit to kinetic data.....	31
Figure 2-6. Yeast displayed protease activity design.....	34
Figure 2-7. Protease activity profiling of TEVp via flow cytometry.....	36
Figure 2-8. Flow cytometric analysis of the first (a) and second (b) rounds of sorting of the first-generation TEVp library.....	39
Figure 2-9. Heat map of sitewise enrichment for first-generation TEVp libraries.....	41
Figure 2-10. Candidate sites for second-generation TEVp library.....	41
Figure 2-11. Top 10 most enriched clones for each of the 7 TEVp first-generation libraries.....	43
Figure 2-12. Initial heat map for library G using a limited number of reads.....	43
Figure 2-13. Flow cytometric analysis of first (a), second (b), and third (c) rounds of sorting of second-generation TEVp library.....	44
Figure 2-14. Second-generation naive library analysis.....	45
Figure 2-15. Heat map of sitewise enrichment for second-generation TEVp library.....	46
Figure 2-16. SDS-PAGE analysis to quantify TEVp production yield.....	48
Figure 2-17. Comparison of enzymatic parameters of TEVp 2-15 to wild-type enzyme (TEVp1).....	49
Figure 2-18. Illustration of sites mutated in optimal variant TEVp2.....	50
Figure 2-19. Epistatic analysis.....	51
Figure 2-20. Comparison of enzymatic parameters of TEVp 18-23 to wild-type enzyme (TEVp1).....	51
Figure 2-21. Summary of the effect on enzymatic parameters of singular mutations when evaluated in the context of a various TEVp clones.....	53
Figure 2-22. Illustration of proximity of active site cysteine (C151) to sites 30-32.....	54
Figure 2-23. Comparison of enzymatic activity of TEVp 24-27 to wild-type enzyme (TEVp1).....	54
Figure 2-24. Comparison of enzymatic activity of <i>top</i> and <i>middle</i> gate mutants to wild-type enzyme (TEVp1).....	55
Figure 2-25. Protease activity profiling of Srt7 via flow cytometry.....	57
Figure 2-26. Relative solvent accessibility is a poor predictor of mutational tolerance in the sites around TEVp's active site.....	58
Figure 2-27. Distance to the substrate cut site is a poor predictor of mutational tolerance in the sites around TEVp's active site.....	59
Figure 2-28. FoldX-predicted stability changes from point mutations is a poor predictor of mutational tolerance in the sites around TEVp's active site.....	59

Figure 2-29. BLOSUM62 score is a poor predictor of mutational tolerance in the sites around TEVp's active site.....	61
Figure 2-30. Relationship between first-generation enrichment and second-generation enrichment.....	63
Figure 2-31. Schematic of yeast surface display construct.....	65
Figure 3-1. Biopsy of a heterogenous tumor.....	71
Figure 3-2. Illustration of how synthetic reporters can non-invasively detect disease.....	73
Figure 3-3. Signaling pathways of EGFR.....	75
Figure 3-4. Synthetic reporters to detect or quantify receptor expression.....	77
Figure 3-5. The wild-type tenth type III domain of fibronectin (PDB: 1TTF).....	78
Figure 3-6. Schematic illustrating feasibility of our extension of the <i>in vivo</i> synthetic reporter approach that leverages ligand-enzyme fusions to remotely detect aberrant receptor expression.....	86
Figure 3-7. Urinary reporter concentration after one hour as a function of EGFR-expressing tumor burden at various substrate concentrations.....	86
Figure 3-8. Schematic illustrating the sequence of the synthetic reporter <i>in vitro</i> assay using ligand-enzyme fusion proteins.....	87
Figure 3-9. SDS-PAGE analysis to quantify fusion production yield.....	88
Figure 3-10. Production yield of various fusions.....	89
Figure 3-11. Affinity estimations for various fusions compared to literature $K_D$ values for ligand only.....	90
Figure 3-12. Michaelis-Menten fit to kinetic data.....	91
Figure 3-13. Results of <i>in vitro</i> cellular assay.....	93
Figure 3-14. FnD in <i>E. coli</i> lysate generates minimal background fluorescence signal compared to TEVp mutants in lysate.....	94
Supplemental Figure 3-1. Mathematical model for the extension of the <i>in vivo</i> synthetic reporter approach to detecting aberrant receptor expression.....	125

## **Chapter 1 – Introduction**

---

### **1.1 Engineering protease activity**

Proteases, also known as peptidases or proteinases, are proteolytic enzymes that hydrolyze peptide bonds in proteins<sup>1</sup>. They are categorized as aspartic proteases, cysteine proteases, glutamic proteases, metalloproteases, asparagine proteases, serine proteases, or threonine proteases based on the presence of specific amino acid residues at the active site that participate in the catalytic mechanism<sup>2</sup>. The applications of proteases, from industrial and biotechnological to medical and pharmaceutical, are legion<sup>3-11</sup>. The global industrial enzyme market, of which the largest segment relates to proteolytic enzymes, has a multi-billion-dollar valuation alone<sup>4</sup>. Proteases have been the focus of many medical and pharmaceutical investigations because they are both found in all living organisms and have myriad functions in the biological processes of various hosts and pathogens, including apoptosis, signal transduction, and the processing of polypeptide hormones<sup>5-6</sup>. Additionally, their applications in industry and biotechnology include peptide synthesis, diagnostics and therapy, materials processing, and the removal of fusion tags from recombinant proteins<sup>6-8</sup>. Due to the large-scale importance of proteases, they have been widely investigated to better understand their structure-function relationships and interactions with substrates and inhibitors; because high activity and specificity are generally essential for optimal enzymatic performance in any of the aforementioned applications, many protein engineering efforts have been undertaken to improve or alter these enzymes in terms of proteolytic or thermal stability, specificity or catalytic efficiency<sup>5-6</sup>. Display technologies that tether libraries of protein mutants to a specific platform (e.g., ribosomes<sup>12</sup> or the surface of phage<sup>13</sup>, mammalian<sup>14</sup> or yeast<sup>15</sup> cells) have been quite successful in forming a genotype-phenotype linkage that facilitates

high throughput selections to identify protein variants with improvements in specific functions (e.g., ligand binding, stability)<sup>16-17</sup>. However, there are unique challenges in forming this genotype-phenotype linkage for enzymes such as proteases, as the use of soluble substrate will not alter surface-bound signal<sup>16</sup>. Further, regardless of the specific protein function to be improved, the challenges of immensity, sparsity and ruggedness of the sequence space, or all possible sequences for that protein, impact a protein engineer's ability to find beneficial mutations<sup>18</sup>. Finally, within protein engineering, engineering enhanced activity in proteases is uniquely challenging due to the multifunctional mechanism of substrate binding and enzymatic cleavage, as the performance of each can be significantly reduced by even minor structural changes<sup>18</sup>. In the following subsections of this thesis introduction, these challenges, as well as the relevant literature and advancements heretofore made, will be explained in detail.

## **1.2 Genotype-phenotype linkage strategy for selecting improved protease activity**

Proteins generally are very evolvable in that they respond and adapt to selection pressure<sup>18</sup>. The protein engineering community has put this trait to productive use by utilizing directed evolution – a vital protein engineering technique that cycles between gene diversification and selection of improved variants – to improve the function of countless proteins<sup>18-20</sup> (Figure 1-1). Directed evolution, however, is a process which necessitates a link between genotype and phenotype<sup>20</sup>. This connection oftentimes manifests in display of plasmid-encoded protein variants that allow for flow cytometry-based screening to identify cells with an increase in surface-bound fluorescence. When the displayed protein is an enzyme such as a protease, however, change in surface-bound signal will not materialize if soluble substrate is used, necessitating a different

approach<sup>16</sup>. The use of fluorescently-labeled soluble substrate that is converted by cell-bound enzyme to a fluorescent product that can bind the surface of yeast cells has been explored<sup>21-25</sup>, but this approach requires a substrate that is amenable to modification (e.g., adding an affinity handle or fluorescent probe), which dramatically narrows the list of enzymes that can be studied. A protease-specific generalized approach involves sequestering the enzyme and substrate in the endoplasmic reticulum (ER), with enzyme-specific cleavage of substrate leading to removal of the substrate's ER retention sequence and subsequent cell surface display of the remaining substrate peptide<sup>26</sup>. Another approach, seemingly for bond-forming enzymes like the cysteine transpeptidase *Staphylococcus aureus* sortase A (SrtA)<sup>27-28</sup>, presents both enzyme and substrate on the same Aga2 display protein; enzyme conjugates a soluble molecule to tethered substrate, with soluble molecule conjugation, easily detected by fluorescently-tagged streptavidin labeling and then collecting cells with increased fluorescence via flow cytometry, indicative of enzymatic activity. These fairly recent advancements provide several potential platforms for efficient screening of enzyme libraries, but up to this point cell-surface display constructs have not been utilized for screening protease libraries. In this work, we employed the yeast surface display platform put forth by Cochran *et al.*<sup>27</sup> and validated this construct for this precise purpose.

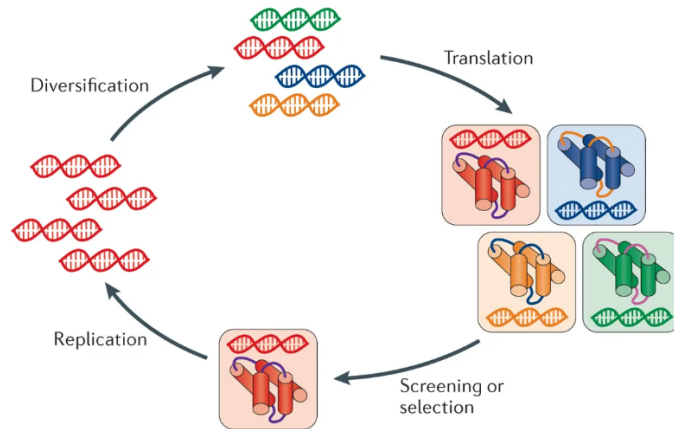


Figure 1-1. **Sequence of events in directed evolution.** An expressed gene library is first screened for a particular function by exploiting a genotype-phenotype linkage, and desirable genes are replicated and used for the next round of diversification. Figure reproduced from {20}.

### 1.3 Tools to constrain diversification of sequence space

In a protein's sequence space, adjacent protein sequences are differentiated by a single missense mutation, with greater distance separating two proteins indicative of greater heterogeneity between the two protein sequences. Different proteins in the sequence space will possess different levels of 'fitness', which is a property defined by the experimenter (e.g., binding affinity for a target of interest when considering ligands, catalytic efficiency against a particular substrate when considering enzymes)<sup>18,29</sup>. Three characteristics of sequence space, namely immensity, sparsity of functional mutants, and ruggedness around elevated fitness mutants, challenge combinatorial efforts to engineer improved protein fitness<sup>18</sup> (Figure 1-2). The reason for the immensity can be illustrated with basic combinatorics, as a protein of length  $N$  would have  $20^N$  possibly protein sequences because the genetic code is able to produce 20 different amino acid monomers at a given site<sup>29</sup>. Even for modest protein lengths, we can only practically test a miniscule percentage of sequence space<sup>18</sup>. The sparsity of the sequence space refers to the fact that the vast majority of proteins will lack fitness and be nonfunctional<sup>30-33</sup>. However, it is



also true that the remaining functional proteins, albeit a minute fraction of the overall sequence space, are likely to be in relatively close proximity to one another in sequence space<sup>34-36</sup>. Unfortunately, this area of the fitness landscape may in fact be rugged, which makes the search for greater fitness problematic: if starting at a local maxima in sequence space, protein evolution by point mutation, which can be understood as moving to nearby territory in the sequence space, can easily lead to a decrease, before an increase, in fitness if the topography is rugged<sup>18</sup> (Figure 1-3). In reality, however, the likelihood of a true local maxima, in which any possible point mutation is harmful to fitness, seems unlikely except in the event of compromised stability, which can motivate prioritizing stability engineering so the protein's amenability to mutation improves and the landscape theoretically becomes smoother and less rugged<sup>37-38</sup>.

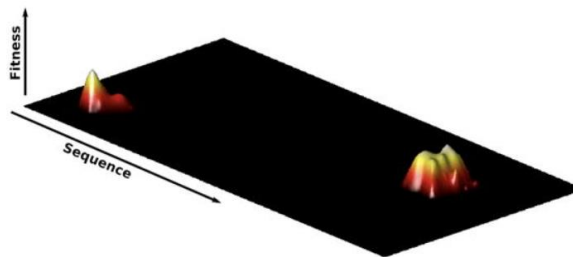


Figure 1-2: **Fitness landscape.** Black and yellow regions represent low and high fitness, respectively. Figure reproduced from {18}.

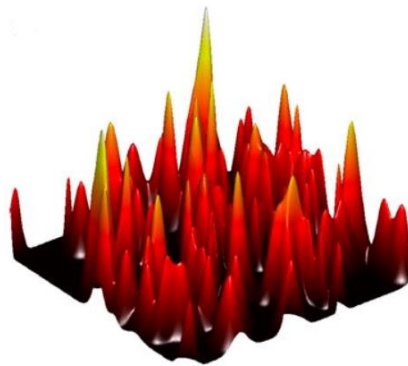


Figure 1-3: **Rough fitness landscape.** Evolution with this landscape is particularly challenging as any small mutational changes may lead to dramatic reductions in fitness. Figure reproduced from {18}.

While random mutagenesis could be combined with directed evolution, these challenges associated with a protein's sequence space motivate the utilization of tools to help focus library diversity to sites and specific residues most likely to confer increased fitness<sup>17,20</sup>. While it still can be difficult to exactly predict which amino acid sequence changes will lead to a specific desired behavior, a rational design approach that harnesses available structural, functional and phylogenetic information has been shown to be of tremendous use in beneficially targeting library diversity<sup>18,20</sup>. In recent years, a semi-rational approach, which is a hybrid of rational design and directed evolution, has emerged as a way to harness the strengths of each<sup>39-45</sup>. Structural information, such as from a solved atomic-scale crystal structure of an enzyme complexed with its substrate, can be utilized to target mutations to a particular region, in this example the residues in the substrate binding pocket and the area around the active site. Several studies have demonstrated that some beneficial mutations may actually reside outside the area directly around the enzymatic active site<sup>46-48</sup>. Kagamiyama *et al.* demonstrated that all but one of the 17 beneficial mutations incorporated into a mutant of aspartate aminotransferase, which was created by directed evolution to possess a million-fold increase in catalytic efficiency for valine, were in fact not in the vicinity of the active site<sup>48</sup>. Nevertheless, a plethora of research indicates that the active site region is still a prime target to find mutations that improve protein fitness<sup>18-19,49-52</sup>. In investigating *Pseudomonas fluorescens* esterase, Kazlauskas *et al.* showed that active site-centric mutations were far more likely to increase selectivity towards methyl 3-bromo-2-methylpropionate substrate than distant mutations, a result that led them to advocate for random mutagenesis that focused on the substrate binding pocket specifically when trying to efficiently engineer enhanced

catalytic properties<sup>51</sup>. Phylogenetic consensus sequence information is oftentimes gathered by probing databases to identify multiple protein homologues of the protein of interest, followed by aligning the homolog sequences to best identify the most frequent residue(s) at every position<sup>53</sup>. By leveraging evolutionary information, we can identify these consensus residues, which typically lend to improved fitness of the protein<sup>39</sup>. Wyss *et al.*<sup>54</sup> showed that consensus residues can improve stability and enzymatic activity of phytases, which are phosphatase enzymes that catalyze hydrolysis of phytic acid. Improvements can be dramatic, as they improved the phytase melting temperature ( $T_m$ ) by approximately 27-34 °C. In this sequence-homology approach, residue preferences will already also be suggested. In the structural approach mentioned above, either saturation mutagenesis should be employed to test every single amino acid option at the narrowed list of sites, or diversity at these positions should be restricted by specific tools. Computational models such as Rosetta<sup>55</sup> and FoldX<sup>56</sup> evaluate steric effects and intermolecular interactions (e.g., hydrogen bonding, electrostatic and hydrophobic interactions) in order to compute free energies that can predict which point mutations are likely to be beneficial for protein stability. Janssen *et al.*<sup>57</sup> utilized this algorithmic approach to identify 10-12 beneficial mutations in the enzyme limonene epoxide hydrolase that, when combined, boosted the melting temperature by 35 °C, improved catalytic activity and led to an increase in half-life over 250-fold. Christensen and Kepp<sup>58</sup> utilized a modified FoldX protocol to demonstrate a strong correlation between decreases in FoldX-predicted mutational  $\Delta\Delta G$  and an increase in  $T_m$  in a fungal laccase. Various indices that describe amino acid substitutability/homology have also been utilized to predict beneficial mutations<sup>59-64</sup>. These concepts and techniques can be used to restrict

diversity at particular sites to those amino acids most likely to improve fitness. By focusing our libraries on positions and amino acids identified by these rational structural- or sequence-based methods, directed evolution experiments can remove large swaths of low-fitness sequence space and thus improve sampling of more desirable, high-fitness sequence space<sup>18,20</sup>. By narrowing our focus to the most relevant regions of sequence space, we can better study and elucidate sequence-function relationships for a variety of proteins, our understanding of which is still oftentimes lacking<sup>18</sup>. In this thesis, we advanced the understanding of sequence-function mapping of a particular protease: tobacco etch virus protease (TEVp).

#### **1.4 Applying improved proteases to synthetic urinary biomarkers to detect aberrant receptor expression**

The use of naturally-occurring blood-based molecular biomarkers for diagnosis of a multitude of disease states is beset by several issues such as low concentration<sup>65</sup>, rapid degradation<sup>66</sup>, or an inability to detect the biomarker in a complex biological fluid<sup>67</sup>. Bhatia and colleagues conceived synthetic biomarkers to address these shortcomings, which have demonstrated potential to detect multiple diseases including thrombosis, liver fibrosis, and colorectal cancer<sup>68-70</sup>. In this approach, a nanoscale probe is administered intravenously to traverse the local disease environment and generate small reporter molecules in accordance with protease activity, which is correlative with disease severity (Figure 1-4). The reporters are then filtered into the urine for noninvasive disease detection. To achieve this, nanoscale agents have consisted of reporter peptides connected via a cleavable linkage to nanoparticles, and this linkage is sensitive to proteases upregulated in the local disease site. One limitation of this technology is its dependence on a local dysregulated enzymatic profile as well as the absence of the

equivalent enzymatic activity elsewhere in the otherwise healthy body. A useful translation would be to achieve a similar ability to noninvasively diagnose disease based on aberrant cell-surface protein expression. Present techniques for characterizing abnormal receptor expression possess their own drawbacks such as requiring an invasive and risky procedure such as a biopsy<sup>71-77</sup> or the cost and radiation exposure from PET imaging<sup>78-80</sup>. As the synthetic reporter approach has relied on soluble enzymatic upregulation in the local disease site, a method to link aberrant receptor expression, such as overexpression, to an increased enzymatic presence is needed. A possible means to achieve this is with fusion proteins that combine exogenous enzymatic activity with high-affinity binding for the receptor being overexpressed. This links receptor overexpression with enzymatic upregulation in the disease site, which can be exploited by reporter-bearing nanoparticles with a reporter-nanoparticle cleavable linkage now sensitive to this exogenous enzyme. The work presented in this thesis represents an advancement of this novel concept, from mathematical modeling to demonstrate feasibility to empirical validation of ligand-enzyme fusion functionality to elucidating the various factors needed to optimize this system *in vitro*.

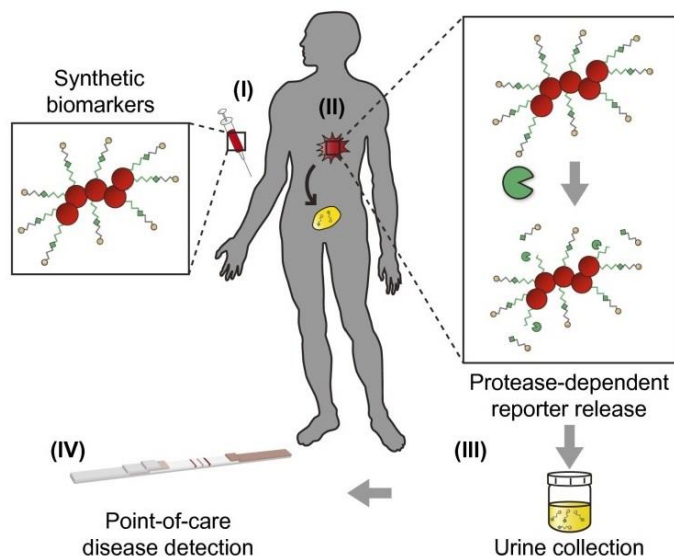


Figure 1-4. **Noninvasive urinary detection of disease using synthetic reporters linked to nanoparticles by a protease-sensitive linkage.** Intravenously-administered nanoparticle-reporter conjugates probe disease site, where upregulated proteases cut reporter from nanoparticle. Filtration of small reporters into urine leads to disease detection. Figure reproduced from {68}.

### 1.5 Summary of approach and results in this thesis

The work in this thesis sought to harness the co-display of protease and substrate on the yeast surface to screen seven rational design-focused saturation mutagenesis libraries around the active site of TEVp to identify beneficial mutations that improve catalytic efficiency. This strategy allowed us to better elucidate the relationship between sequence space and function for this protease, and we successfully identified many multi-mutants that improved catalytic efficiency in a statistically significant manner, up to 2.8- to 2.9-fold in the case of three specific multi-mutants. The application of ligand-enzyme fusions to the synthetic reporter approach was mathematically modeled to test for feasibility. Our analysis suggested the feasibility of this framework as the reporter generation in urine one hour after substrate administration is predicted to be four orders of magnitude above the ELISA detection limit. We also demonstrated the ability to produce ligand-enzyme fusion proteins, with a multitude of ligands and proteases, that

preserve most, if not all, of the catalytic and binding functionality of their components.

While we encountered limitations of these fusions in applying them *in vitro* to the synthetic reporter technology to generate signal from the cell surface in congruence with receptor overexpression, we have identified concrete areas – including fusion linker design, enzymatic activity, substrate lability – that can be enhanced in future work to demonstrate this system's feasibility and lead to further optimization.

## **Chapter 2 – Engineering tobacco etch virus protease activity by screening rational active site combinatorial libraries using a novel yeast surface display construct**

---

### **2.1 Abstract**

Engineering improved activity, either increased  $k_{\text{cat}}$  or decreased  $K_M$ , into proteases using directed evolution remains a challenge due to a relative lack of well-studied display constructs for library screening and a need to elucidate optimal library design. Tobacco etch virus protease (TEVp), a 27-kDa viral cysteine protease commonly used for specific cleavage of recombinant protein affinity tags, has been the subject of past protease engineering efforts, but historically related to attributes other than catalytic activity. Further, most of the past efforts have employed random mutagenesis methods such as error-prone PCR to synthesize libraries. In this study, we developed a novel and seemingly generalizable yeast surface display construct that co-displays protease mutants with substrate on the same Aga2 protein. Identification of enhanced activity mutants is enabled by protease cleavage of tethered substrate removing a hemagglutinin (HA) epitope tag, which empowers FACS-facilitated isolation of cells with a decrease in signal from fluorophore-linked anti-HA antibodies. This platform was shown to quantitatively differentiate catalytic activity at the single-cell level for TEVp and sortase A. Further, utilizing the fact that TEVp is well-studied and has a solved crystal structure, we leveraged this display platform to perform high throughput screens on seven structure-based active site combinatorial libraries created via saturation mutagenesis. Beneficial mutations from these first-generation libraries were incorporated into a single second-generation library, which was screened to identify which individual beneficial mutations



performed optimally in a multi-mutant context. Using enrichment data from the second-generation library screen, we produced a host of TEVp multi-mutants, the vast majority of which demonstrated statistically-significant improvement in catalytic efficiency, generally with a decreased  $K_M$  as opposed to an increased  $k_{cat}$ . These improved multi-mutants will be of benefit to the biotechnological community, and our novel yeast surface protease/substrate co-display system and the insights gleaned on rational active site library design and the TEVp sequence-function map will aid future protease engineering efforts.

## **2.2 Introduction**

Engineering enhanced protein fitness using current methodologies and technologies is beset by a variety of challenges relating to the protein's fitness landscape, most prominently the immensity of the possible variants, sparsity of functional variants, and potential ruggedness around regions of elevated fitness<sup>18</sup>. Thus, rational exploration of sequence space and/or thorough mutant sampling is needed to identify improved function. Moreover, engineering enhanced activity into enzymes such as proteases presents further unique hurdles given their multifunctional mechanism involving substrate binding and proteolysis, which can oftentimes be substantially hindered by even minor structural changes<sup>18</sup>. Sequence space can be explored efficiently, if sparsely, by coupling protein function to the encoding gene sequence to empower high throughput selections without abiotic spatial isolation<sup>16</sup>. When yeast surface display, a prominent example of a genotype-phenotype linkage strategy, is combined with current flow cytometry technologies,  $10^8$  protein variants can be screened. Basic combinatoric calculations suggest the diversification of a limited number of amino acid residues in the enzyme to ensure a relatively thorough sampling of any combinatorial library<sup>16</sup>. Random

mutagenesis, such as by error-prone PCR, can achieve the desired level of diversity, but the probability of finding improved mutants is substandard because sequence space is being targeted via random deviations from the starting sequence as opposed to focusing on those regions that are most likely to impact fitness<sup>17</sup>. The identification of residues most suited for mutation to improve fitness has been the subject of thorough study<sup>41,81-84</sup>. The two emergent major strategies use either sequence-based methods, such as multiple sequence alignment of homologous proteins<sup>39</sup>, or structure-based methods, which leverage three-dimensional models, such as a solved atomic-scale crystal structure of the enzyme complexed with its substrate, and our understanding of substrate-binding within the active site of the enzyme to determine residues most important to reaction progression<sup>17</sup>. While there is a growing body of research suggesting that non-active site mutations, including those on the protein surface, can be beneficial for activity<sup>48,85-88</sup>, mutations made in and around the active site and substrate-binding pocket remain a prime area to find potentially beneficial mutations<sup>30,89</sup>. Therefore, an example of a rational structure-based approach involves identifying a narrowed list of residues (e.g., four or five per library) in this region for site-directed saturation mutagenesis, or an expanded list of sites if diversity is limited (e.g., by chemical homology).

Directed evolution has proven to be a vital technology in protein engineering, and the use of display technologies to screen genetic libraries in a high-throughput manner to identify improved fitness mutants has been invaluable when studying functionality such as ligand binding or protein stability<sup>16-17</sup>. However, there are distinct challenges in linking the genotype of protease mutants with catalytic activity. Most flow cytometric display-based screens, for example as with ligand binding, sort for increased surface-bound

fluorescence; if using displayed enzyme mutants and soluble substrate, an alternative sorting strategy must be devised as the enzymatic conversion of substrate to product will not change surface-bound signal<sup>16</sup>. One strategy that addresses this challenge uses fluorescently-labeled substrate that is enzymatically converted to a fluorescent product that binds to the surface of yeast cells<sup>21-22</sup>. Other similar methods have been developed to identify enhanced enzyme mutants using yeast display and high-throughput fluorescence-activated cell sorting (FACS) technology<sup>23-25</sup>, but in general these methods are reliant on the ability to modify the substrate via an affinity handle or fluorescent probe, and thus these approaches are limited to a subset of enzymes whose substrate can tolerate such alterations<sup>16</sup>. A clever and more generalized approach sequesters the enzyme variant and antibody epitope-flanked substrate as separate entities in the endoplasmic reticulum (ER), with proteolysis of the substrate by the enzyme variant leading to removal of the substrate's ER retention sequence and subsequent cell surface display of the remaining substrate peptide<sup>26</sup>. A particularly interesting new approach has been implemented for bond-forming enzymes like the cysteine transpeptidase *Staphylococcus aureus* sortase A (SrtA)<sup>27-28</sup>. This technique employs dual-display of both the enzyme variant and its substrate on the same Aga2p display protein, with the enzyme variant responsible for conjugating a soluble molecule to the tethered substrate; fluorescently-tagged streptavidin labeling of this newly-added cell-bound molecule, and subsequent collection of yeast cells with increased fluorescence, isolates enhanced enzyme mutants. This dual-display approach could be expanded for enzymes other than transpeptidases, and we sought to utilize this framework to identify improved protease mutants by sorting for cells with decreased fluorescence after proteolytic cutting of substrate removed an epitope tag.

Here we validate the use of a yeast-based protease display system that successfully allows for the quantitative differentiation of enzymatic activity of multiple proteases, chiefly tobacco etch virus protease (TEVp) but also SrtA when acting as a protease. TEVp is a viral cysteine protease commonly used for the specific removal of affinity tags from recombinant proteins<sup>6,90</sup>. It consists of the 27-kDa C-terminal domain of the 49 kDa viral polyprotein processing product nuclear inclusion protein Ia<sup>91-92</sup>. Wild-type TEVp is subject to significant autolysis, so the S219P variant is often used as it largely eliminates autolysis while only reducing the catalytic efficiency by approximately two-fold<sup>93-95</sup>. Previous attempts at screening error-prone PCR random mutagenesis libraries for TEVp using the yeast ER sequestration approach yielded a triple mutant (G79E/T173A/S219V) that increases catalytic efficiency by approximately four-fold compared to the S219P variant<sup>26</sup>, and incorporates the S219V mutation that was also previously proven to prevent autolysis and slightly aid catalytic efficiency (approximately a 1.77-fold improvement compared to wild-type)<sup>93</sup>. This TEVp triple mutant was used as the starting point in our library design, and our display system was leveraged to screen seven rational design-focused saturation mutagenesis libraries (4-5 amino acids per library, 34 residues in total) around the active site of TEVp to identify the mutations in each library that improve catalytic efficiency. A second-generation library evaluated combining beneficial mutations across each first-generation library while also allowing maintenance of wild-type. Selection of the most active mutants from the second-generation library yielded a collection of compelling multi-mutants for catalytic analysis, several of which improved the catalytic efficiency by nearly three-fold.

### **2.3 Methods**

### 2.3.1 Creation of TEVp DNA constructs

Plasmid pCT-SAT (substrate-Aga2p-TEVp) for *S. cerevisiae* cell surface display was prepared through several steps. First, gBlock 1 was PCR amplified using primers 1 and 2 (Supplemental Table 2-1). All PCRs used Phusion DNA polymerase, Phusion HF buffer, and deoxynucleoside triphosphate (dNTP) mixture from New England Biolabs (NEB, Ipswich, MA), oligonucleotide primers from Integrated DNA Technologies (IDT, Coralville, IA), and gene fragments from either Twist Bioscience (San Francisco, CA) or as gBlocks from IDT. For every PCR, a 50  $\mu$ L volume contained 10 ng gene fragment, 1X Phusion HF Buffer, 0.2 mM dNTPs, 0.5  $\mu$ M of each primer, and 1 U Phusion DNA polymerase. The sample was denatured at 98  $^{\circ}$ C for 30 s, followed by 30 cycles of 98  $^{\circ}$ C for 10 s, a unique annealing temperature – determined using the  $T_m$  calculator on the NEB website for specific gene fragment/primer pairings – for 30 s, and 72  $^{\circ}$ C for 30 s per kb, with a final extension at 72  $^{\circ}$ C for 10 min. pCT acceptor vector was digested at 37  $^{\circ}$ C for 8-16 hrs using CutSmart buffer and DNA restriction enzymes NheI and XhoI from NEB. PCR products and digested vector were isolated by 1% w/v agarose gel electrophoresis and purified using the GenCatch Gel Extraction Kit from Epoch Life Sciences (Missouri City, TX). PCR product was inserted into pCT acceptor vector via the Gibson Assembly Protocol<sup>96</sup> (E5510, NEB). A 20  $\mu$ L volume containing 1X Gibson Assembly Master Mix from NEB, 50-100 ng vector, and 3-fold molar excess of insert was incubated in a thermocycler for 20 minutes at 50  $^{\circ}$ C, followed by a 5 min incubation at 4  $^{\circ}$ C. 2  $\mu$ L of reaction mixture was transformed into NEB5 $\alpha$  competent *Escherichia coli* from NEB. Plasmid DNA was harvested from *E. coli* cultures using the GenCatch Plasmid DNA Mini-prep Kit from Epoch Life Sciences. pCT-SAT<sub>C151A</sub>, containing the cysteine-to-alanine inactivated version of TEVp in the display construct, was prepared

using pCT-SAT, primers 3 and 4, and the QuikChange Site-Directed Mutagenesis Kit from Agilent Technologies (Santa Clara, CA). pCT-S<sub>sc</sub>AT contained the scrambled substrate of LSQEFYN in the display construct. To create this particular construct, single-stranded oligonucleotides 5a and 5b were first combined via a single thermocycler step; in 50  $\mu$ L volume with both oligonucleotides at 50  $\mu$ M, the temperature was first raised to 72 °C for 1 min and then lowered to 66 °C for 3 min. The double-stranded resulting DNA was then gel-recovered and digested, along with pCT-SAT, with PstI and NcoI. Gel-recovered vector and insert were then ligated together; specifically, a 20  $\mu$ L volume containing 1X T4 DNA ligase buffer and 100 U T4 DNA ligase from NEB, 100 ng vector, and 4-fold molar excess of insert was incubated at room temperature for 10 min. Two  $\mu$ L of reaction mixture was used for transformation with NEB5 $\alpha$  competent *E. coli* and plasmid DNA was harvested from *E. coli* cultures via miniprep. For each plasmid, DNA Sanger sequencing was performed with oligonucleotide primers from IDT at the University of Minnesota Genomics Center.

### 2.3.2 Yeast surface display to validate display construct and optimize induction time

*S. cerevisiae* EBY100 cells were transformed with one of the aforementioned three pCT plasmids for TEVp/substrate fusion display via the Frozen-EZ Yeast Transformation II Kit from Zymo Research (Irvine, CA). Transformants were cultured in 5 mL SD-CAA (0.07 M sodium citrate, 20 g/L glucose, 6.7 g/L yeast nitrogen base, 5 g/L casamino acids) at 30 °C for 16-24 hours. Before reaching an OD<sub>600</sub> of 6, cells were pelleted at 1000 g for 3 min and resuspended in SG-CAA (0.1 M sodium phosphate, 19 g/L galactose, 1 g/L glucose, 6.7 g/L yeast nitrogen base, and 5 g/L casamino acids) to yield an OD<sub>600</sub> less than 1, with subsequent SG-CAA incubation times of either 20 min,

2 hrs, 10 hrs, or 24 hrs.  $5 \times 10^5$  yeast cells were centrifuged at 1000 g for 3 min, and supernatant was removed by pipetting. Cells were suspended in 1 mL PBS with 1% BSA (PBSA), followed by centrifugation at 1000 g for 3 min and supernatant removal. To quantify intact and total substrate display levels, via HA and c-Myc epitopes, cells were suspended in 50  $\mu$ L of PBSA with 67 nM polyclonal chicken anti-HA antibody (ab9111, Abcam, Cambridge, United Kingdom) and 67 nM monoclonal mouse anti-c-Myc antibody (9E10, Thermo Fisher Scientific, Waltham, MA) for 30 min at room temperature. Cells were then washed once with 1 mL PBSA, followed by resuspension in 50  $\mu$ L of PBSA with 67 nM of both polyclonal goat anti-chicken antibody conjugated with Alexa Fluor 488 (ab150169, Abcam) and polyclonal goat anti-mouse antibody conjugated with Alexa Fluor 647 (ab150115, Abcam) for 30 min at 4 °C. Cells were washed once more with 1 mL PBSA, followed by resuspension in 100  $\mu$ L PBSA for analysis with the Accuri C6 Plus instrument and software (Becton Dickinson, Franklin Lakes, NJ). Reduction in HA-derived signal (AF488) with preservation of c-Myc-derived signal (AF647) indicated enzymatic cutting of tethered substrate.

### 2.3.3 TEVp first-generation library construction

The protein crystal structure from the Protein Data Bank (1LVB) of catalytically inactivated TEVp (C151A) complexed with its substrate (ENLYFQ↓S) was analyzed by PyMOL and used to identify residues within 6 Angstroms of the substrate for mutation. Four residues were exempted from mutation: the three residues in the catalytic triad (H46, D81, C151) and position 219 due to its importance in preventing autocatalysis. The remaining residues were organized into seven libraries, of 4 or 5 amino acids each, for saturation mutagenesis. Individual library sizes of 4 or 5 residues were chosen because

NNK codon-facilitated construction of libraries of this size will yield a theoretical genetic diversity of 1.1 or 34 million, respectively, which can be reasonably sampled using conventional FACS methodologies. Residue groupings in each library were designed so as to mostly maximize three-dimensional spatial proximity; Figure 2-1 shows that L204 is a notable exception. Among those residues within 6 Angstroms of the substrate, L204 is one of the more isolated and distant sites, but it was included in library E because it was one of the few libraries that could accommodate such an addition (*i.e.*, less than 5 residues) and its incorporation allowed for easy assembly.

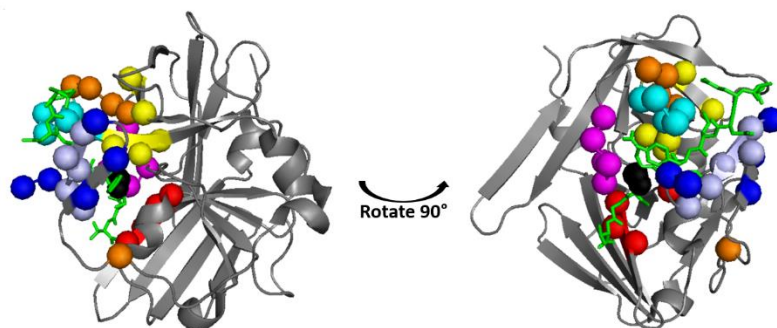


Figure 2-1. **Mutant library design.** TEVp (C151A) is shown complexed with its substrate ENLYFQ↓S (green), with the  $\alpha$ -carbon of the amino acid residues in the seven TEVp libraries color coded: A (red: T30, S31, L32, N44), B (yellow: F139, K141, H167, S168, A169), C (pink: T146, K147, D148, G149, Q150), D (cyan: S170, N171, F172, A173, N174), E (orange: T175, N176, N177, Y178, L204), F (dark blue: V209, W211, G213, K220, P221), and G (light blue: H214, K215, V216, F217, M218). In the substrate, the carbonyl carbon of Q and the amide nitrogen of S are represented as black spheres.

The individual libraries were constructed by fusion PCR via overlap extension of two DNA fragments (Figure 2-2). For libraries A, C, D, F, and G, the first DNA fragment was created by a single PCR amplification using pCT-SAT, a common forward oligonucleotide 6 that annealed at the C-terminal region of Aga2p, and a library-specific reverse oligonucleotide (7-11) that annealed in the TEVp gene. For libraries B and E, two sequential PCR steps were necessary to sufficiently extend the first DNA fragment in the 3' direction using a pair of library-specific reverse oligonucleotides (12-13 for library



B; 14-15 for library E). The second DNA fragment for libraries A-G was analogously created by a single PCR amplification, but with a common reverse oligonucleotide 16 that annealed immediately downstream of TEVp and a library-specific forward oligonucleotide annealing in the TEVp gene (17-23). Library-specific oligonucleotides were designed such that the two amplified DNA fragments would have sufficient overlap to enable their fusion by PCR with oligonucleotides 6 and 16. In this fusion PCR step, 5  $\mu$ L of each gel-purified DNA fragment was added to the 50  $\mu$ L PCR reaction volume in place of the template. The acceptor vector pCT was digested with Sall, BamHI and SacII to remove the TEVp gene. Gel-recovered library inserts and digested vector were both concentrated with PelletPaint, and resuspended in dH<sub>2</sub>O for electroporation of competent yeast cells. For all electroporation procedures, 300  $\mu$ L of competent yeast were combined with 6  $\mu$ g plasmid vector and ~120  $\mu$ g library insert (or no insert for a negative control), and electroporated at 1.2 kV and 25  $\mu$ F to combine vector and insert into whole plasmid via homologous recombination<sup>97</sup>. Cells were incubated in YPD (10 g/L yeast extract, 20 g/L bacto peptone, 20 g/L glucose) at 30 °C, 250 rpm for 1 hr, after which cells were centrifuged at 1300 g for 1 min and resuspended in SD-CAA for a >16 hr incubation at 30 °C, 250 rpm to propagate successful transformants. Dilutions plated on SD-CAA plates allowed for the calculation of the number of unique transformants for libraries.

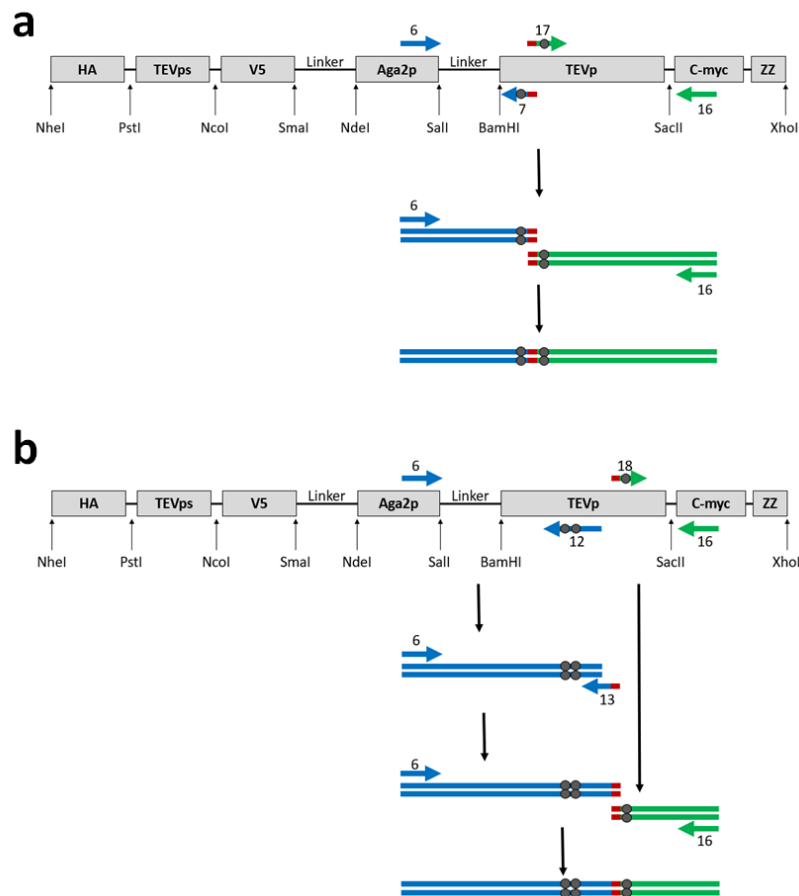


Figure 2-2. **Schematic of library assembly via fusion PCR.** (a) Library A. (b) Library B. Libraries C, D, F and G were constructed via a similar process to (A) and library E was constructed similarly to (B), but different oligonucleotide primers annealing to the TEVp gene were used. Area of overlap between the two DNA fragments is denoted by red color, and areas of diversity are illustrated with grey circles. Oligonucleotide primer identification numbers are found in Supplemental Table 2-1.

### 2.3.4 Fluorescence-activated cell sorting of first-generation library

For each sorting experiment, a positive control (pCT-SAT) and two negative controls (pCT-SAT<sub>C151A</sub> and pCT-S<sub>sc</sub>AT) were analyzed before sorting.  $5 \times 10^5$  yeast cells for each control were induced and labeled as before.

Prior to reaching an OD600 of 6 in SD-CAA, electroporated yeast cells – specifically, 10-fold the diversity for library A (~10 million) and 3-fold the diversity for libraries B-G (~100 million) – were pelleted and resuspended in SG-CAA at an OD600 of 1.0 for a 4-hr incubation at 30 °C, 250 rpm to induce protein expression and surface

display. Yeast cells – the same number as before, 10 and 100 million, respectively – were pelleted and resuspended in 100  $\mu$ L of PBSA with ab9111 and 9E10 (both at 667 nM except for library A: 67 nM) for 30 min at room temperature. Cells were washed once with 1 mL PBSA, followed by resuspension in 100  $\mu$ L of PBSA with both ab150169 and ab150115 (both at 667 nM except for library A: 67 nM) for 30 min at 4 °C. Cells were washed once more with 1 mL PBSA, followed by resuspension in 100  $\mu$ L PBSA for fluorescence-activated cell sorting (FACS) using the FACS Aria II instrument (Becton Dickinson). Collected cells were propagated in SD-CAA at 30 °C, 250 rpm for a second round of sorting. For the second sort, 10 million yeast cells in SD-CAA at an OD600 < 6 from each library's first sort were pelleted and resuspended in SG-CAA at an OD600 of 1.0 for a 4-hr incubation at 30 °C, 250 rpm. 10 million yeast cells were pelleted and processed as before, but with antibody concentrations of 67 nM for ab9111, 9E10, ab150169 and ab150115. Collected cells from the second sort were then propagated in SD-CAA at 30 °C, 250 rpm for deep sequencing.

#### 2.3.5 Deep sequencing of first-generation libraries

For each of the 7 libraries, these second sort collections, as well as the unsorted libraries, were prepped for deep sequencing. From these 14 total samples, the plasmid DNA was harvested from yeast cells using Longlife Zymolyase from G-Biosciences (St. Louis, MO). In general,  $1 \times 10^8$  cells were incubated for 1 hr at 37 °C in 200  $\mu$ L of lysis solution (50 mM phosphate buffer, 1 M sorbitol, 10 mM  $\beta$ -mercaptoethanol) with 1:20 dilution of Zymolyase, with DNA subsequently purified using silica spin columns. The diversified regions underwent two sequential PCR amplifications. The first PCR amplification used oligonucleotides 24-25 (library A), 25-26 (libraries B-D), and 27-28

(libraries E-G). This amplified DNA fragment was then subjected to a subsequent round of PCR amplification using the common forward oligonucleotide primer 29 and a reverse oligonucleotide primer (30-36 and 37-43 for the unsorted and sorted populations, respectively, of libraries A-G) to add unique Illumina adapters for deep sequencing. Deep sequencing was performed using an iSeq 100 Sequencing System (Illumina, San Diego, CA) at the University of Minnesota Genomics Center.

The specific amino acid frequency at each site was calculated by summing reads containing that amino acid at that site and dividing by the total number of reads. Sitewise enrichment scores for each amino acid were calculated as the  $\log_2$  of the ratio of the frequency in the bottom gate to the frequency in the unsorted library. The enrichment range at a particular site was calculated by subtracting the smallest enrichment value at that site from the largest enrichment value at that site; if a specific single mutation at a site possessed an enrichment score greater than the wild-type enrichment value at that site plus ten percent of that site's enrichment range, that mutation was deemed beneficial.

### 2.3.6 TEVp second-generation library construction

Enrichment data from the 7 first-generation libraries helped create a single second-generation library with constrained diversity (Table 2-6). Beneficial mutations (Table 2-5) were always incorporated unless the entire wild-type sequence for an individual first-generation sublibrary was clearly the most enriched clone, in which case the wild-type sequence was exclusively used in the second-generation library regardless of any compelling individual mutations at a particular site in the sequence. Potentially compelling non-beneficial mutations, either possessing equal or close-to-equal enrichment compared to wild-type, were also considered for incorporation into the

second-generation library. Wild-type residues, whether highly enriched or not in the first-generation libraries, were also allowed in the second-generation library. The library insert was constructed by fusion PCR of three DNA fragments (Figure 2-3). The first DNA fragment was created by PCR using pCT-SAT, the aforementioned forward oligonucleotide 6, and a mixture of 12 reverse oligonucleotides (44-55) introducing targeted diversity at sites 30-32. The second DNA fragment was analogously created by PCR, but with a forward oligonucleotide 56 and a mixture of 3 reverse oligonucleotides (57-59) introducing targeted diversity at site 204. The third DNA fragment was created by using the aforementioned reverse oligonucleotide 16 and a mixture of 12 forward oligonucleotides (60-71), introducing targeted diversity at sites 211, 214-216, 218, and 220. Oligonucleotide incorporation was imbalanced, which later resulted in incomplete sampling of the second-generation library, but comparative sequencing analysis of pre- and post-sorting populations still allowed for an accurate assessment of sitewise enrichment. Oligonucleotides were designed such that the three DNA fragments would have sufficient overlap to enable their fusion by PCR as analogously done before with the first-generation libraries. For the second-generation library, the three fragments were connected in a piecemeal approach of two steps of overlap extension PCR. First, the second and third fragments were connected by fusion PCR with oligonucleotides 16 and 56; this product was then connected to the first fragment by fusion PCR with oligonucleotides 6 and 16. This library insert was combined via electroporation with digested vector as before to create intact plasmid. The number of unique transformants for the library was  $2.0 \times 10^7$ , with less than  $1 \times 10^6$  transformants for the vector-only negative control.

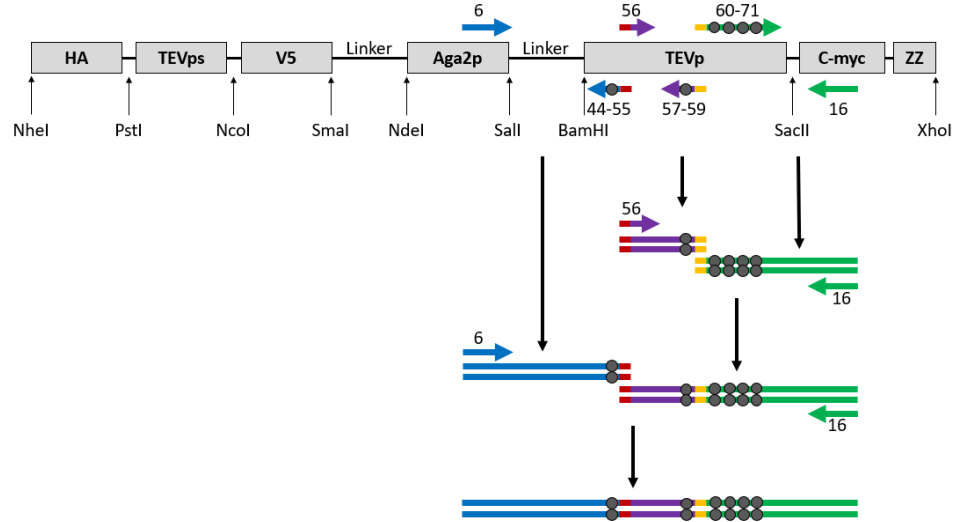


Figure 2-3. **Schematic of library assembly via fusion PCR for the second-generation TEVp library.** Three PCR fragments (blue, purple, and green) were created, and connected in a piecemeal fashion. The purple and green fragments were first connected via overlap extension PCR using the yellow area of overlap. This new fragment was then connected to the blue fragment via overlap extension PCR using the red area of overlap. Areas of diversity are illustrated with grey circles. Oligonucleotide primer identification numbers are found in Supplemental Table 2-1.

### 2.3.7 Fluorescence-activated cell sorting of second-generation library

100 million electroporated yeast cells in SD-CAA at an OD600 < 6 were pelleted and resuspended in SG-CAA at an OD600 of 1.0 for a 4-hr incubation at 30 °C, 250 rpm to induce protein expression and surface display. 100 million yeast cells were then pelleted and labeled in the same fashion as the first sort of libraries B-G. Cells with minimized HA:c-Myc ratio were collected via FACS in a *bottom* gate, and were propagated in SD-CAA at 30 °C, 250 rpm for a second round of sorting. 4 million yeast cells from this propagated population were then similarly induced, then pelleted and labeled in the same fashion as the first sort of library A. Cells were then collected via FACS in two gates: *improve*, with a moderate reduction in the HA:c-Myc ratio compared to wild-type TEVp, and *best*, which even further lowered the ratio compared to *improve*. These two populations, *bottom-improve* and *bottom-best*, were propagated, and 2 million yeast cells were similarly processed and prepped for FACS, where cells were collected in

a *best* gate with minimized HA:c-Myc ratio. These two populations, *bottom-improve-best* (BIB) and *bottom-best-best* (BBB), were propagated in SD-CAA for deep sequencing. The relevant FACS statistics for each of the three sorts of the second-generation library can be found in Table 2-8.

#### 2.3.8 Sanger sequencing of *top* and *middle* gates from first sort of second-generation TEVp library

To further validate the display construct's ability to stratify populations on FACS by enzymatic activity, yeast cells were also collected in *top* and *middle* gates during the second-generation TEVp library's first sort, and pCT plasmid DNA was extracted from  $1 \times 10^8$  cells from each gate. The plasmid DNA mixture was then used to transform NEB5 $\alpha$  competent *E. coli*, and transformants were plated on agar plates with 100  $\mu\text{g/mL}$  ampicillin; monoclonal colonies from agar plates (3 from the *top* gate, 8 from the *middle* gate) were then individually cultured and miniprepped. The TEVp gene sequence in the purified plasmid DNA was then amplified and elongated by PCR using oligonucleotides 72-73. pET vector was separately digested with NheI and BamHI. Intact pET plasmid was created via Gibson Assembly using gel-recovered cut vector and PCR product, and transformed NEB5 $\alpha$  competent *E. coli* cells were then transferred to agar plates with 50  $\mu\text{g/mL}$  kanamycin. Individual colonies were cultured and sequenced using Sanger sequencing before protein production.

#### 2.3.9 Sanger sequencing of unsorted second-generation TEVp library

Plasmid DNA was similarly extracted from  $1 \times 10^8$  yeast cells from the unsorted second-generation TEVp library, and used to transform NEB5 $\alpha$  competent *E. coli*. 30 transformants on ampicillin agar plates were cultured, Miniprepped, and Sanger sequenced. The empirical amino acid frequencies from these 30 sequences matched the

theoretical values well, so theoretical amino acid frequency values were used moving forward in calculating sitewise enrichment.

#### 2.3.10 Deep sequencing of second-generation TEVp library

For samples BBB and BIB, as well as the unsorted second-generation library, plasmid DNA of  $1 \times 10^8$  yeast cells was extracted, and the diversified regions were PCR amplified using oligonucleotides 24 and 28. These amplified DNA fragments were then subjected to a subsequent round of PCR amplification using the common forward oligonucleotide primer 29 and a reverse oligonucleotide primer (30, 31, and 32 for the BBB, BIB, and unsorted population, respectively) to add unique Illumina adapters for deep sequencing on an iSeq 100 Sequencing System. 274,288 reads for BIB and 140,054 reads for BBB were obtained, and values of sitewise  $\log_2$  enrichment versus unsorted library for these datasets were calculated as before with the first-generation library. Clones also were ordered by their clonal enrichment score; specifically, a  $\log_2$  function was applied to the ratio of the percentage of reads belonging to a particular clone divided by the theoretical frequency of that clone in the unsorted library. The top 100 enriched clones in the BIB and BBB gates were then used to determine sitewise enrichment. Finally, epistasis, or the context-dependence of mutational effects, was examined by calculating the  $\log_2$  enrichment of various double mutants. The combination of data for sitewise enrichment (for the entire set of reads as well as the top 100 enriched clones) and epistasis helped inform and narrow the list of TEVp mutants we sought to produce in *E. coli*.

#### 2.3.11 TEVp mutant production and characterization



Gene fragments for all compelling TEVp variants were synthesized by manufacturer with codons optimized for *E. coli* production. Fragments were PCR amplified and elongated for addition to digested pET vector via Gibson Assembly. Sanger sequencing validated proper assembly.

To produce TEVp clones, T7 Express competent *E. coli* (NEB) were transformed with TEVp mutant-containing pET plasmid and grown in 150 mL lysogeny broth (10 g/L tryptone, 5 g/L yeast extract, 10 g/L sodium chloride) cultures at 37 °C, 250 rpm. Cells were induced with 0.5 mM isopropyl- $\beta$ -D-thiogalactoside (IPTG) from Teknova (Hollister, CA) at an OD600 of 0.7-1.0 for 5 hrs at 30 °C, 250 rpm. Cells were pelleted by centrifugation at 3200 g for 15 min and lysed in 1.5 mL of lysis buffer (pH 7.5, 50 mM sodium phosphate, 0.5 M NaCl, 5% glycerol, 5 mM CHAPS detergent, and 25 mM imidazole). Cells were subjected to four cycles of freeze-thaw and centrifuged at 12,000 g for 10 min at 10 °C. Supernatant was filtered using a 0.45  $\mu$ m filter, and lysate was analyzed by sodium dodecyl sulfate polyacrylamide gel electrophoresis (SDS-PAGE), using NuPAGE reagents from Thermo, to identify and quantify TEVp concentration. For each TEVp mutant produced, 6  $\mu$ L of cell lysate [in 18  $\mu$ L total volume, 1X NuPAGE LDS Sample Buffer, and 10%  $\beta$ -mercaptoethanol (Sigma-Aldrich)] was run in an individual well on a NuPAGE 4-12% Bis-Tris gel. Additionally, an 11 kDa epidermal growth factor receptor (EGFR)-binding fibronectin (Fn) clone, given the alias D (FnD)<sup>98</sup>, was produced under the same conditions as TEVp variants to allow analysis of the 28 kDa region from a culture not expressing TEVp. 5  $\mu$ g each of bovine serum albumin (66 kDa) and carbonic anhydrase (30 kDa) were added in a single well as concentration standards. Gels were run at 200 V for 40 min in 1X NuPAGE MES SDS running buffer.

Gels were stained with SimplyBlue Safe Stain (Thermo) for 1 hr and washed in dH<sub>2</sub>O for 1 hr. Protein concentrations were analyzed via densitometry using a ChemiDoc MP Imaging System (Bio-Rad) and free GIMP 2 imaging software.

The custom peptide 2-Abz-ENLYFQSGTK-Dnp (United Peptide, Herndon, VA), in which Dnp quenches emission of the Abz fluorophore, allowed for examination of TEVp clone catalytic parameters. All TEVp clones were diluted to a clonal concentration of 0.2  $\mu$ M using TEVp-deficient lysate of *E. coli*-produced FnD. For a given TEVp clone, 40  $\mu$ L of this solution was added to five microplate wells. For TEVp 1-15 and 18-23, 40  $\mu$ L solution containing the peptide substrate in PBS at various concentrations (0, 3, 10, 30, and 100  $\mu$ M) was then added to the five wells to give a final TEVp concentration of 0.1  $\mu$ M and final substrate concentrations of 0, 1.5, 5, 15, and 50  $\mu$ M. For variants 24-38, only the 50  $\mu$ M substrate concentration was probed. All samples were run in triplicate, with fluorescence ( $\lambda_{\text{ex}} = 320$  nm,  $\lambda_{\text{em}} = 420$  nm) measurements taken every 15 s for 1 min total on a Synergy H1 microplate reader (BioTek Instruments, Winooski, VT). A TEVp-deficient negative control, consisting of FnD lysate, was also analyzed at these five substrate concentrations. At every substrate concentration, background fluorescence was subtracted using the FnD negative control, and graphs of background-subtracted fluorescence values versus time provided slopes in fluorescent units/min at different substrate concentrations. 2-aminobenzoic acid (2-Abz) was also obtained from Sigma-Aldrich (St. Louis, MO) to make a standard curve (Figure 2-4); this allowed for the conversion of graphical slopes in units of fluorescence units/min to initial reaction velocities in units of [unquenched 2-Abz]/min. For TEVp variants 1-15 and 18-23, these

reaction velocities at various substrate concentrations were then fit to a Michaelis-Menten model (Figure 2-5) to calculate  $k_{cat}$  and  $K_M$ .

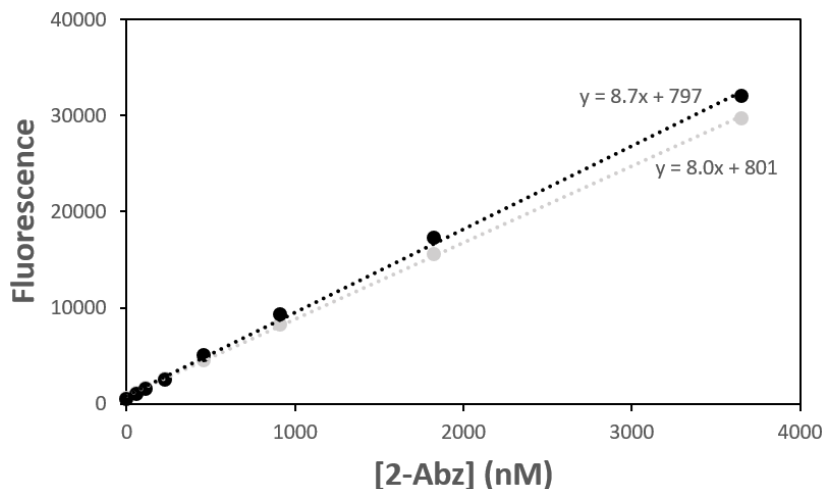


Figure 2-4. **Standard curve of 2-Abz.** Two trials were performed, with linear fit  $R^2$  values exceeding 0.99 in both cases. The average slope of 8.32 fluorescence units/[2-Abz] was used in calculations to convert fluorescence units/min to [2-Abz]/min.

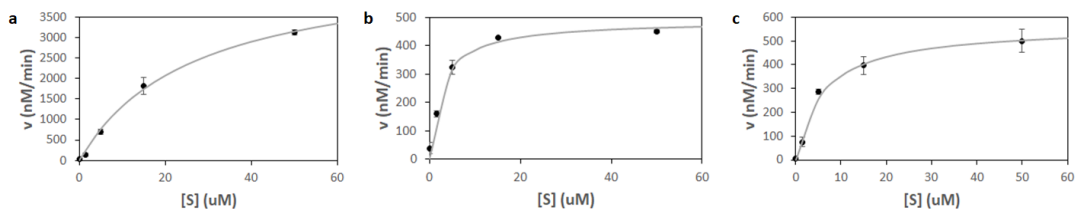


Figure 2-5. **Michaelis-Menten fit to kinetic data.** Three examples of TEVp variants are shown: a) TEVp3, b) TEVp21, and c) TEVp2 R215K reversion.

TEVp12, the hexamutant (T30I/S31T/L32V/L204M/W211I/K215R) clone determined to have optimal improvement in catalytic efficiency over parental TEVp1, was evaluated for the impact made by each individual mutation in the context of TEVp12. Specifically, six TEVp12 reversion clones were produced that left five of the mutations in place but mutated the remaining site back to its wild-type amino acid residue via PCR. These clones were analogously produced, quantified and analyzed as described above.

### 2.3.12 Creation of Srt7 DNA constructs

Plasmids pCT-SAS [substrate(LPETGG)-Aga2p-Srt7] and pCT-S<sub>sc</sub>AS [scrambled substrate(TEGLGP)-Aga2p-Srt7] for *S. cerevisiae* cell surface display were prepared through several steps. First, a 50 µL volume containing 50 µM each of primers 74a and 74b was incubated in a thermocycler at 72 °C for 90 s followed by an incubation at 67 °C for 3 min to anneal the complementary primers together and create a dsDNA version of the Srt7 substrate LPETGG flanked by N-terminal PstI cut site and C-terminal NcoI cut site. The identical procedure was performed using primers 75a and 75b to create an analogous dsDNA version of the scrambled substrate TEGLGP. These two dsDNA molecules, along with pCT-SAT plasmid, were then digested at 37 °C for 8-16 hrs using CutSmart buffer and DNA restriction enzymes PstI and NcoI. The digested vector and dsDNA substrate molecules were isolated by gel electrophoresis and purified using the GenCatch Gel Extraction Kit. Gel-recovered vector and insert then ligated together as described before.

gBlock 2, containing the gene sequence for Srt7 flanked by BamHI and SacII cut sites, was PCR amplified using primers 76 and 77. This PCR fragment was then digested with BamHI and SacII restriction enzymes. Also, the two newly created pCT plasmids, containing TEVp and either Srt7 substrate LPETGG or scrambled substrate TEGLGP, were similarly digested with BamHI and SacII to remove the TEVp gene. The digested vectors and PCR product were isolated and purified as before, and the PCR product containing the Srt7 sequence was ligated into both digested vectors to create pCT-SAS and pCT-S<sub>sc</sub>AS.

### 2.3.13 Estimation of solvent-accessible surface area

In PyMOL, the command “set dot\_solvent, 1” enabled calculation of solvent-accessible surface area, and the command “set dot\_density, 4” maximized the number of dots to enable optimal accuracy in area estimations. The get\_area command was used on select residues in the PDB file 1LVB<sup>36</sup>.

## 2.4 Results

### 2.4.1 Design of TEVp/substrate fusion display construct

In our display system, a yeast adhesion receptor subunit Aga2 domain facilitates surface display of a fusion protein containing both a TEVp canonical substrate and a TEVp variant (Figure 2-6). On the N-terminal side of Aga2p is the TEVp substrate (TENLYFQ↓SGTRRW<sup>26</sup>, with the specific canonical substrate underlined and the protease cleavage site indicated by ↓), itself flanked by N-terminal hemagglutinin (HA) and C-terminal V5 epitope tags; on the C-terminal side of Aga2p is the TEVp mutant followed by a C-terminal c-Myc tag. The Aga2p protein is connected to the N-terminal V5 tag and the C-terminal TEVp mutant by a linker of composition (G<sub>4</sub>S)<sub>2</sub>-ASASPAAPAPASPAAPAPSA-(G<sub>4</sub>S)<sub>2</sub> that utilizes a validated proline, alanine and serine (PAS) sequence<sup>99-100</sup>. This design allows for the sampling of tethered substrate by the tethered TEVp mutant, and mutants with greater enzymatic activity are responsible for a more robust removal of the HA tag. The resultant displayed protein can then be labeled with fluorophore-linked antibodies targeting HA and c-Myc, and increased proteolysis will manifest in a decreased HA-derived signal with preserved fluorescence from c-Myc.

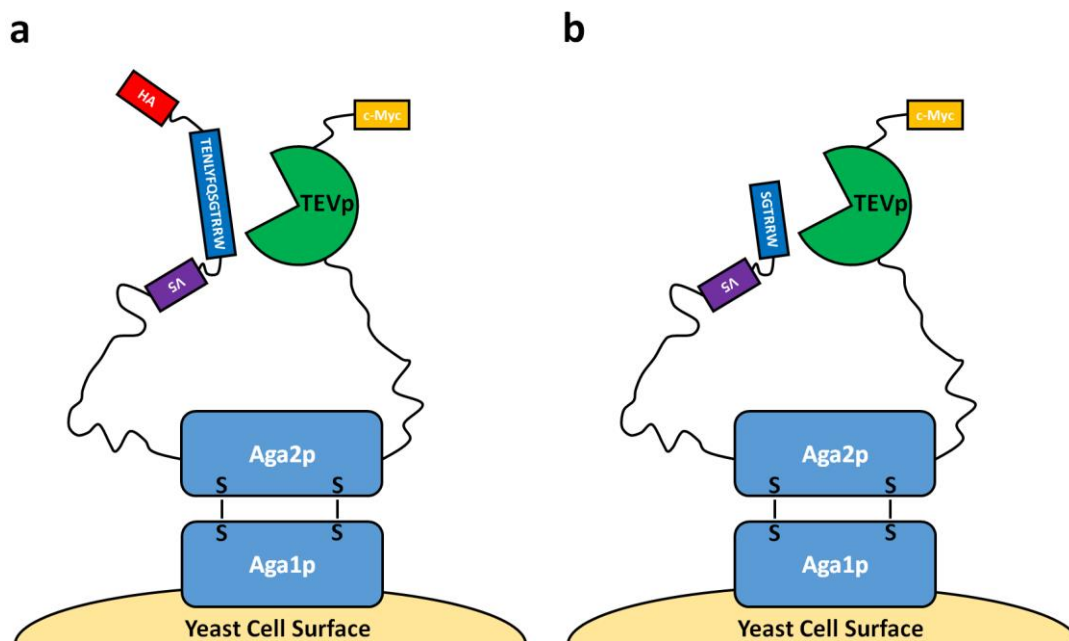


Figure 2-6. **Yeast displayed protease activity design.** Schematic of the yeast surface display construct, which allows for simultaneous extracellular presentation of a TEVp mutant and its substrate, in the a) uncleaved and b) cleaved substrate states, with the latter resulting in the removal of an HA epitope tag. The order of the construct from N-terminus to C-terminus is: HA epitope (red) – TEVp substrate (TENLYFQSGTRRW, dark blue) – V5 epitope (purple) – (G<sub>4</sub>S)<sub>2</sub>ASASPAAPAPASPAAPAPSA(G<sub>4</sub>S)<sub>2</sub> linker (long black line) – Aga2p (light blue) – (G<sub>4</sub>S)<sub>2</sub>ASASPAAPAPASPAAPAPSA(G<sub>4</sub>S)<sub>2</sub> linker (long black line) – TEVp (green) – c-Myc (yellow).

#### 2.4.2 Validation of display construct to differentiate enzymatic activity

We first aimed to assess the ability of the substrate-Aga2p-TEVp (SAT) fusion display system to differentiate enzymatic activity. To enable comparison, two negative control plasmids were created: S<sub>sc</sub>AT, which replaced the TEVp substrate with a scrambled version of the 7-amino acid canonical cut site (TLSQEFYNGTRRW), and SAT<sub>C151A</sub>, where the TEVp is inactivated with a C151A mutation to eliminate the catalytic triad cysteine<sup>94</sup>. Yeast transformed with one of these plasmids were induced for 0, 0.3, 2, 10 or 24 hrs. A sample of yeast transformed with SAT, but not induced to display SAT and rather incubated in galactose-free medium, was also included as a negative control. Cells were labeled with antibodies against the N-terminal HA and C-

terminal c-Myc tags and analyzed by flow cytometry (Figure 2-7). For the three time points between 2-24 hrs, there is a clear reduction in N-terminal epitope signal for the displayed population of SAT compared to the two negative controls SAT<sub>C151A</sub> and S<sub>sc</sub>AT. The loss in HA signal with preservation of the C-terminal c-Myc signal is consistent with TEVp-mediated cleavage of tethered substrate and release of HA epitope tag, thus demonstrating this fusion construct can differentiate between active and inactive TEVp mutants. As induction time increases, protein display increases (Table 2-1). Increased time also provides more opportunity for substrate cutting, which furthers differentiation between negative controls and SAT but also reduces the opportunity to identify mutants more active than wild-type (Figure 2-7). Thus, an induction time of 4 hrs was used moving forward to enable differentiation of wild-type activity versus either reduced or enhanced activity.

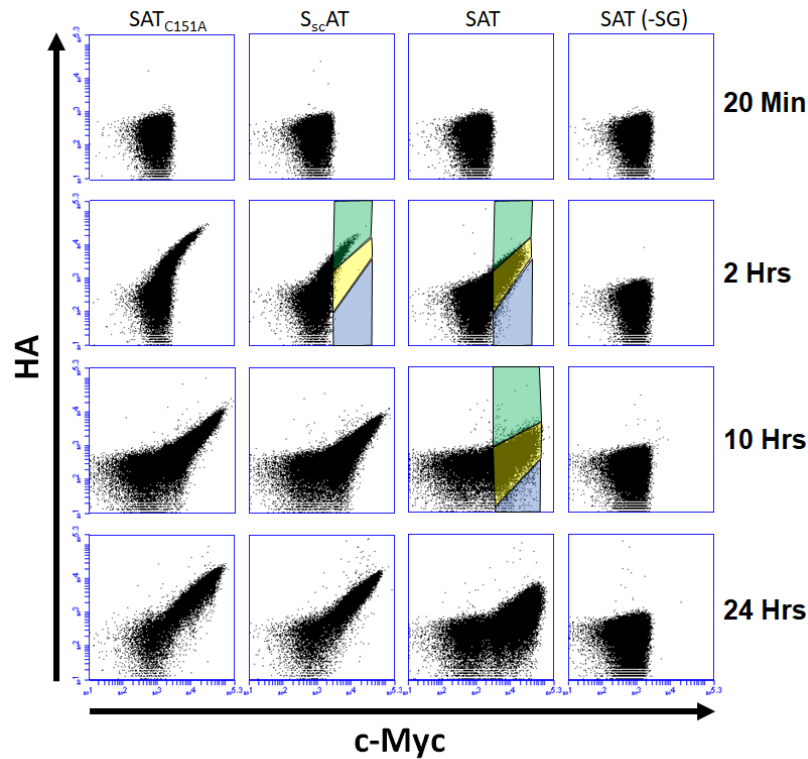


Figure 2-7. **Protease activity profiling of TEVp via flow cytometry.** SAT (-SG), the uninduced sample, was incubated in SD-CAA for 20 min, 2 hrs, 10 hrs, or 24 hrs. SAT (wild-type TEVp with the canonical ENLYFQ/S substrate), as well as negative controls SAT<sub>C151A</sub> (inactivating mutation of the active site cysteine of wild-type TEVp) and S<sub>sc</sub>AT (scrambled substrate, LSQEFYN, unrecognizable to TEVp), were induced for similar times. Cells were labeled with fluorescent antibodies to detect HA and c-Myc tags. X- and y-axes show up to  $2 \times 10^5$  fluorescence units. For the SAT 2 hr timepoint, three polygons are shown: yellow (captures wild-type TEVp population), green (decreased activity) and blue (improved activity). These exact polygons are also shown for the S<sub>sc</sub>AT 2 hr timepoint, illustrating how this negative control population falls almost entirely in the decreased activity green polygon. Analogous polygons are drawn for the SAT 10 hr timepoint, demonstrating the diminishing opportunity to collect improved mutants as induction time increases.

Plasmid	Induction Time (hrs)	c-Myc+ %	Median c-Myc Signal of c-Myc+ Cells
SAT	2	26	5,700
	10	41	9,652
	24	75	22,694
S <sub>sc</sub> AT	2	8.2	4,192
	10	69	10,139
	24	84	17,156
SAT <sub>C151A</sub>	2	30	4,743
	10	48	8,138
	24	79	18,696

Table 2-1. **Summary statistics of flow cytometry-assisted TEVp activity profiling.** Percentage of cells analyzed by flow cytometry that were c-Myc+, as well as the median c-Myc signal of this subpopulation, for SAT, S<sub>sc</sub>AT, and SAT<sub>C151A</sub> at different SG-CAA induction times. Induction times of 0 and 0.3 hrs are not shown as they resulted in negligible protein display.



### 2.4.3 TEVp mutational analysis

#### *2.4.3.1 First-generation TEVp library*

The validated SAT system was used for directed evolution of TEVp catalytic efficiency using rationally guided combinatorial libraries. We hypothesized that residues within the active site, with the exception of the catalytic triad (H46, D81, and C151)<sup>101</sup>, would provide the greatest opportunity for impactful evolution. Residues within 6 Å of the substrate ENLYFQS were identified for mutation using the crystal structure of catalytically inactivate TEVp (C151A) complexed with its substrate (PDBID: 1LVB)<sup>102</sup>. Two exceptions were conserved without mutation: the catalytic triad and V219 due to its importance in preventing autocatalysis<sup>93</sup>. Diversification of the remaining residues presents sequence space that is much too vast for full combinatorial exploration: 20<sup>34</sup>. Thus, we further hypothesized that focused libraries, clustered by location, could be used to identify beneficial mutations while preserving the potential to identify local epistasis. The sites were organized into seven saturation mutagenesis libraries comprising 4-5 residues each (Figure 2-1). Libraries were synthesized with degenerate oligonucleotides including NNK codons (encoding all 20 amino acids resulting in 1 and 34 million for the 4- and 5-site libraries, respectively) and homologously recombined into the pCT-SAT yeast display construct (Figure 2-2). The number of unique transformants generated by electroporation for libraries A-G were 780, 520, 560, 220, 340, 160, and 700 million, respectively. Deep sequencing performed on the libraries indicated amino acid representation at each site that was reasonably consistent with the usage of NNK codons (Table 2-2).

Mutation	Theoretical	Library A				Library B				Library C				Library D				Library E				Library F				Library G											
		T30	S31	L32	N44	F139	K141	H167	S168	A169	T146	K147	D148	G149	Q150	S170	N171	F172	A173	N174	T175	N176	N177	Y178	L204	V209	W211	G213	K220	P221	H214	K215	V216	F217	M218		
F	3%	3%	3%	3%	2%	4%	3%	2%	1%	4%	1%	4%	3%	2%	3%	4%	5%	5%	4%	4%	2%	3%	2%	2%	3%	2%	2%	2%	4%	4%	4%	4%	4%	5%	4%		
W	3%	2%	1%	2%	6%	2%	1%	5%	7%	6%	1%	2%	2%	3%	2%	4%	0%	0%	0%	0%	7%	1%	2%	1%	2%	5%	2%	2%	2%	4%	4%	4%	5%	5%	5%	5%	
Y	3%	4%	4%	4%	2%	7%	4%	3%	3%	1%	2%	4%	4%	4%	4%	3%	4%	4%	3%	2%	3%	3%	3%	3%	2%	2%	3%	3%	3%	3%	3%	3%	3%	3%	3%		
P	6%	1%	1%	1%	2%	1%	1%	2%	7%	1%	6%	5%	6%	5%	8%	5%	1%	2%	7%	7%	1%	1%	1%	1%	1%	5%	1%	1%	10%	4%	6%	2%	2%	2%	2%	2%	
M	3%	3%	2%	3%	3%	2%	1%	4%	5%	2%	1%	2%	1%	2%	2%	3%	3%	3%	3%	3%	3%	3%	2%	2%	3%	3%	3%	3%	3%	3%	3%	3%	3%	3%	3%	5%	
I	3%	3%	3%	4%	2%	4%	2%	3%	2%	4%	1%	1%	2%	2%	2%	4%	3%	3%	2%	2%	4%	3%	3%	2%	2%	2%	3%	2%	3%	3%	3%	3%	3%	3%	4%	3%	
L	9%	9%	9%	10%	6%	10%	12%	7%	8%	7%	5%	7%	7%	7%	7%	13%	8%	9%	8%	7%	11%	10%	10%	10%	11%	8%	8%	8%	10%	9%	8%	8%	8%	8%	8%		
V	6%	4%	4%	4%	10%	3%	3%	10%	10%	9%	1%	7%	7%	7%	5%	9%	1%	1%	1%	1%	4%	4%	4%	3%	9%	5%	4%	4%	8%	8%	10%	10%	10%	10%	10%	10%	
A	6%	5%	6%	6%	6%	5%	5%	6%	6%	7%	5%	7%	7%	7%	8%	3%	5%	5%	6%	6%	6%	6%	7%	7%	6%	6%	6%	6%	6%	6%	5%	5%	5%	5%	5%		
G	6%	3%	3%	2%	2%	4%	3%	3%	3%	3%	10%	7%	7%	6%	6%	9%	2%	2%	2%	2%	2%	2%	2%	2%	4%	4%	4%	3%	10%	11%	11%	11%	11%	11%	11%		
C	3%	2%	2%	2%	4%	3%	1%	6%	4%	5%	1%	4%	3%	4%	3%	3%	8%	6%	6%	6%	3%	3%	3%	2%	5%	2%	2%	2%	5%	4%	4%	4%	5%	5%	4%	4%	
S	3%	2%	2%	2%	4%	3%	1%	6%	4%	5%	1%	4%	3%	4%	3%	3%	8%	6%	6%	6%	3%	3%	3%	2%	5%	2%	2%	2%	5%	4%	4%	4%	5%	5%	4%	4%	
T	6%	9%	10%	7%	3%	10%	10%	3%	4%	4%	6%	5%	6%	6%	7%	2%	2%	2%	3%	3%	10%	11%	11%	11%	4%	9%	9%	3%	4%	4%	4%	4%	4%	4%	4%	4%	
N	3%	4%	4%	4%	2%	4%	5%	2%	2%	2%	1%	2%	2%	3%	3%	3%	3%	2%	1%	2%	4%	4%	4%	2%	1%	3%	4%	4%	2%	2%	2%	2%	2%	2%	2%	3%	
Q	3%	4%	4%	4%	2%	3%	2%	2%	2%	1%	2%	2%	3%	2%	5%	4%	1%	2%	2%	2%	3%	3%	4%	4%	2%	4%	4%	5%	2%	2%	2%	2%	2%	2%	2%	2%	
D	3%	3%	3%	3%	3%	3%	2%	4%	3%	4%	9%	7%	7%	6%	4%	3%	4%	4%	3%	3%	3%	3%	3%	3%	3%	2%	3%	3%	3%	3%	3%	3%	3%	3%	3%	3%	3%
E	3%	2%	2%	2%	4%	1%	2%	4%	5%	4%	8%	5%	5%	4%	5%	2%	4%	4%	4%	4%	1%	1%	1%	1%	3%	3%	3%	3%	3%	3%	3%	3%	3%	3%	3%	3%	3%
H	3%	6%	5%	5%	2%	6%	4%	2%	2%	2%	2%	3%	2%	3%	4%	4%	1%	1%	2%	1%	7%	6%	6%	5%	2%	4%	5%	5%	3%	3%	3%	2%	2%	2%	2%	2%	
K	3%	3%	3%	3%	2%	2%	2%	2%	2%	3%	1%	3%	2%	3%	3%	2%	2%	2%	2%	2%	2%	2%	2%	1%	4%	3%	4%	4%	2%	3%	5%	3%	3%	3%	3%	3%	
R	9%	7%	7%	7%	11%	7%	6%	12%	10%	8%	4%	5%	5%	8%	8%	4%	7%	7%	8%	10%	6%	7%	7%	8%	12%	10%	10%	10%	10%	10%	9%	9%	9%	9%	10%	10%	

Table 2-2. **Naïve library analysis.** Amino acid frequencies at every mutated site in the seven unsorted first-generation TEVp libraries were determined by deep sequencing, with theoretical amino acid frequencies as predicted by NNK codon usage shown for reference. The stop codon frequency ranged from 1.8-2.3% across the 34 sites. The amino acid frequencies shown represent the percentage representation of a particular amino acid with regards to all non-stop codon mutations. At a given site, dark red shading was assigned if the measured frequency was greater than the theoretical frequency by  $\geq 5\%$ ; similarly, dark blue shading was assigned if the measured frequency was less than the theoretical frequency by  $\geq 5\%$ , or equal to 0% in the case where theoretical frequencies were 3%. If a site occupied the exact predicted theoretical frequency, white color was assigned.

The libraries were analyzed for enzymatic activity (Figure 2-8a). Strikingly, a reasonably large fraction (27-55%) of each library exhibited activity relatively comparable to wild-type, which is consistent with relatively high tolerance to the extensive saturation mutagenesis at 4-5 sites near the active site. Each library also exhibited variants with increased activity relative to wild-type. Two sequential rounds of FACS were performed to enrich more active mutants. In the first round of sorting (Figure 2-8a) a conservative gate was used to improve capture of potentially beneficial mutants because the short induction time needed for functional differentiation resulted in moderate display (13-24% c-Myc+) and thereby moderate sampling of the large diversities (14-77%) (Table 2-3). In the second round of sorting, reduced diversities enabled cells to be sorted more stringently (Figure 2-8b, Table 2-3). The plasmid DNA of both the unsorted library and bottom gate collections was extracted and analyzed via deep sequencing.

Library	First Sort				Second Sort			
	Cells Evaluated (millions)	c-Myc+ %	WT Activity %	Library Sampling %	Cells Collected (millions)	Cells Evaluated (millions)	c-Myc+ %	Cells Collected (thousands)
A	4.6	17	27	77	0.1	1.5	55	1.9
B	45	13	32	17	1.2	10	46	10
C	39	17	47	20	0.9	4.6	52	34
D	26	18	53	14	0.5	5.6	57	13
E	46	18	48	24	1.2	10	51	26
F	48	24	55	34	1.0	7.7	52	28
G	38	16	42	18	0.8	7.1	49	23

Table 2-3. **FACS data from the first-generation TEVp libraries.** Wild-type (WT) activity % was calculated by taking the number of cells emulating wild-type (SAT) activity and dividing by the number of overall expressing cells. Library sampling was calculated by multiplying evaluated cell number by c-Myc+ percentage, and then dividing by theoretical genetic diversity (1.1 million and 34 million for library A and libraries B-G, respectively).

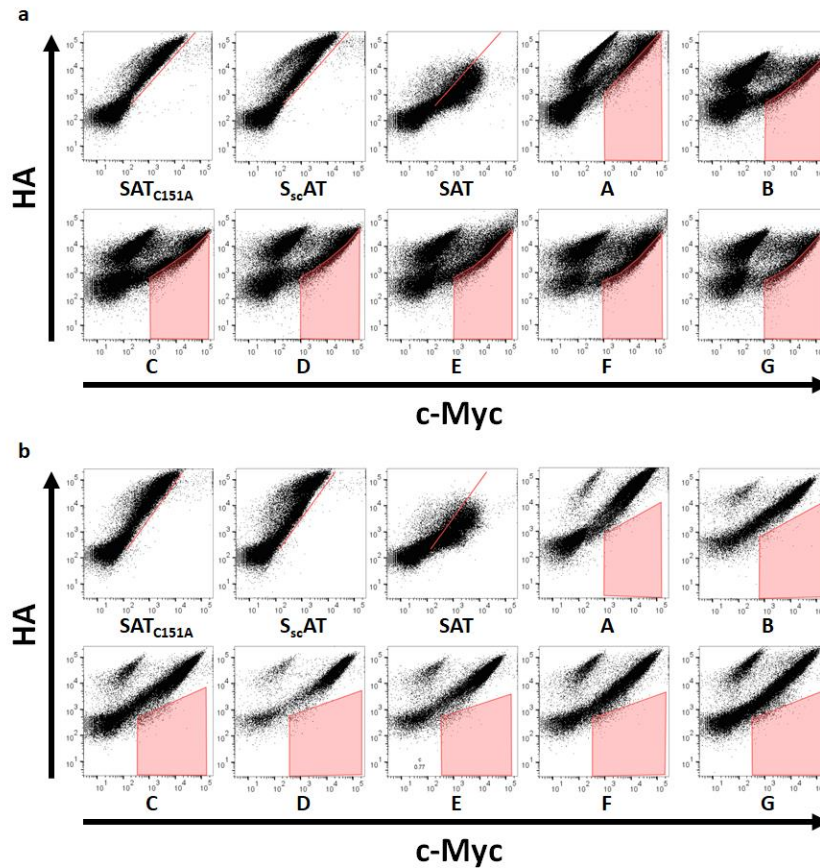


Figure 2-8. **Flow cytometric analysis of the first (a) and second (b) rounds of sorting of the first-generation TEVp library.** Two negative controls (SAT<sub>C151A</sub> and S<sub>sc</sub>AT), a single positive control (SAT), and each library were evaluated prior to each round of sorting. A consistent red reference diagonal is shown for the positive and negative controls. For both rounds of sorting, the collected population is shown in a red gate drawn uniquely for each library.

Using the unsorted and doubly sorted reads from the first-generation TEVp libraries (Table 2-4), amino acid enrichments were calculated at every mutated site as the ratio of frequencies from the doubly sorted versus unsorted populations (Figure 2-9). The wild-type amino acid was the most enriched residue at 29 of the 34 sites. Libraries C and G, which diversified adjacent regions (pink and light blue in Figure 2-1) exhibit no substantial enrichment other than wild-type. Of the five sites with preferred mutations (Table 2-5), three were in library A proximal to the cleaved substrate bond (Figure 2-10). A fourth site, F139, interacts with the substrate away from the cleavage site. The fifth site, L204, is relatively distant from the substrate and, thus, the preferred mutants (M, I, F, V, and C) likely aid activity via structural modulation. Seven additional sites exhibit amino acids that are effectively enriched, albeit less than the wild-type residue, in the more active TEVp population (pink spheres in Figure 2-10). Collectively, the sitewise amino acid enrichments provide guidance for a second-generation library with simultaneous mutations across the active site.

Library	Unsorted Reads	Doubly Sorted Reads
A	116,468	174,788
B	6,515	265
C	2,514	292
D	8,989	312
E	81,993	142,743
F	93,319	77,240
G	504,264	373,372

Table 2-4. **Deep sequencing reads for unsorted and doubly sorted populations from first-generation TEVp libraries A-G.**

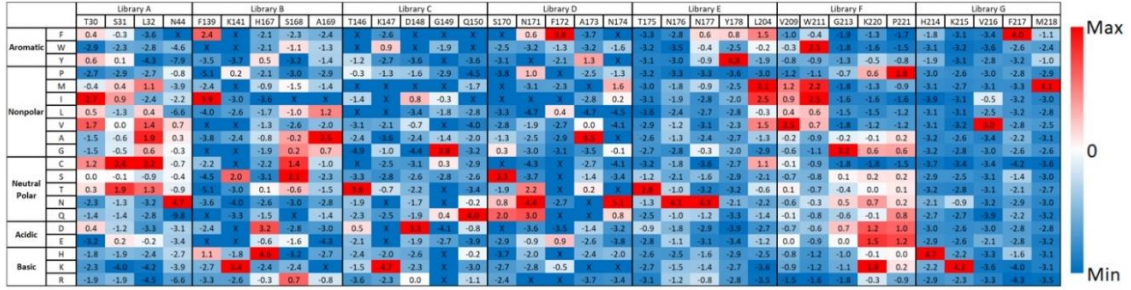


Figure 2-9. Heat map of sitewise enrichment for first-generation TEVp libraries. Enrichment (as the log<sub>2</sub> of the ratio of frequencies from doubly sorted versus unsorted populations) was calculated for every residue at every site in the seven TEVp libraries. 'X' denotes an amino acid that was observed in the unsorted reads but not observed in the doubly sorted reads. Each site is scaled independently.

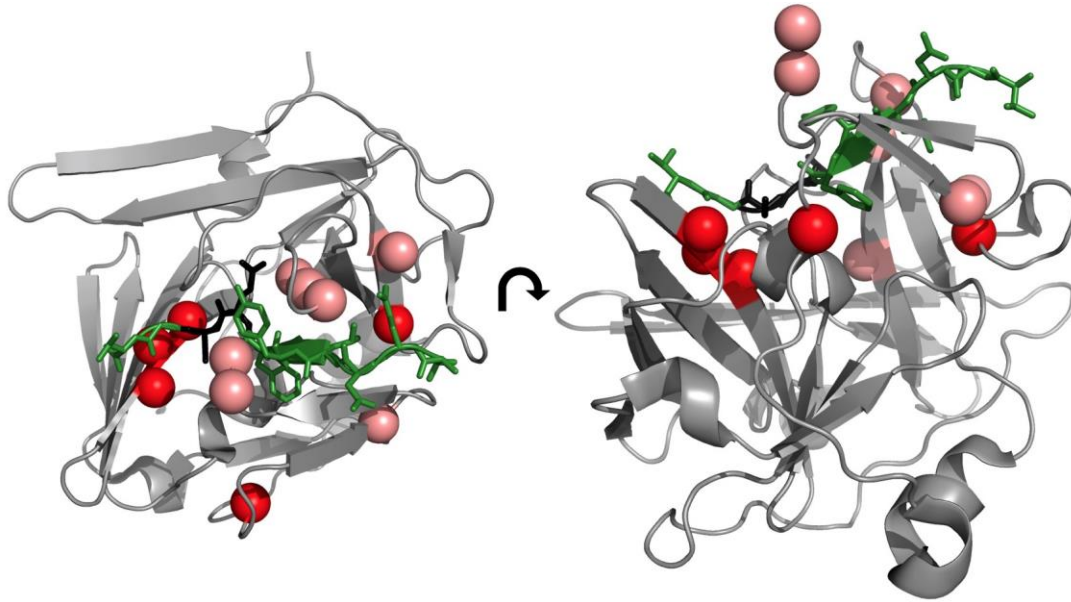


Figure 2-10. Candidate sites for second-generation TEVp library. TEVp is shown complexed with substrate (green), with sites containing at least one beneficial mutation highlighted in red, and sites containing at least one mutation that is substantially positively enriched, but not beneficial, highlighted in pink. Mutations were considered substantially positively enriched if they possessed an enrichment score of at least 1.5 and were within 2.0 of the wild-type score, or if they possessed a score of at least 1.0 if the wild-type score was below 2.0; this rubric identified seven sites: H167, N171, W211, K141, S170, K220, and P221.

T30	S31	L32	F139	L204
I	C	C	I	M
V	T	A		I
C	I	V		F
		T		V
		M		C

Table 2-5. List of beneficial mutations from the first-generation TEVp libraries. A mutation was deemed beneficial if its enrichment was greater than the wild-type enrichment plus ten percent of that site's enrichment range, which was calculated by subtracting the smallest enrichment value from the largest enrichment value at that site.

#### 2.4.3.2 *Second-generation TEVp library*

Beneficial mutations across first-generation libraries (Table 2-5) were, in general, incorporated into a single second-generation library (Table 2-6). An exception to this rule existed if the entire wild-type sequence for an individual first-generation sublibrary was clearly the most enriched clone, in which case the wild-type sequence was exclusively used in the second-generation library regardless of any compelling individual mutations at a particular site in the sequence. For example, while I139 is 2.9-fold more enriched than wild-type F139, the wild-type sequence for library B (F139, K141, H167, S168, A169) had the highest observed clonal enrichment (Figure 2-11), 5.7-fold more than the second-place sequence (I139, S141, D167, R168, L169). Thus, the second-generation library maintained the wild-type sequence in the library B region. It should be acknowledged that the single mutant F139I may not have been tested in the first generation given the 17% sampling of library B (Table 2-3). At two sites (W211 and K220), potentially compelling mutations that were either equally enriched or only slightly less enriched than wild-type (W211I, W211M, and K220E) were also incorporated into the second-generation library. Further, initial deep sequencing of library G, which yielded only 301 reads for the unsorted population, identified multiple mutations with enrichment greater than wild-type: H214 (K, N, W), V216 (E, Q) M218 (P, G); or comparable: K215G and K215E (Figure 2-12). These mutations were included in the second generation design, although later deeper sequencing revealed reduced enrichment in the first generation.

<b>T30</b>	<b>S31</b>	<b>L32</b>	<b>L204</b>	<b>W211</b>	<b>H214</b>	<b>K215</b>	<b>V216</b>	<b>M218</b>	<b>K220</b>
I	T	V	M	I	K	G	E	P	E
V	I	M	I	M	N	E	Q	G	
C	C	C	F		Q	R	L		
A		T	V		W				
		A	C						

Table 2-6. **Second-generation TEVp library design.** List of residue positions (bold) mutated in second-generation TEVp library, with the available amino acid mutations at each site shown below. Wild-type was incorporated at every site as well.

Library A		Library B		Library C		Library D	
T30, S31, L32, N44	Enrichment	F139, K141, H167, S168, A169	Enrichment	T146, K147, D148, G149, Q150	Enrichment	S170, N171, F172, A173, N174	Enrichment
V,T,C,N	6.5	F,K,H,S,A	8.9	T,K,D,G,Q	6.9	S,N,F,A,N	7.8
I,T,M,G	6.4	I,S,D,R,L	6.4	D,K,D,G,Q	5.2	Q,Q,F,A,N	6.8
D,T,T,N	6.1	F,A,H,S,A	2.6	D,K,R,C,Q	4.1	N,P,F,T,I	6.3
F,G,A,V	6.1	F,R,H,S,A	2.5	A,T,I,Q,H	3.7	Q,F,E,Y,M	5.0
L,R,E,G	6.0	L,Q,G,F,F	2.4	S,V,I,I,N	3.7	S,N,F,V,N	3.8
I,T,L,N	5.9	C,A,D,M,W	2.3	D,K,R,W,S	3.1	N,V,W,S,Q	1.8
T,T,T,N	5.9	F,L,H,S,A	1.9	A,F,D,W,K	3.1	T,N,Y,A,M	1.5
L,C,T,N	5.7	F,H,H,S,A	1.9	P,W,V,A,R	1.8	L,K,D,F,P	0.7
V,I,T,N	5.6	F,T,H,S,A	1.6	E,K,D,G,Q	1.0	G,T,L,V,G	0.2
C,T,C,N	5.6	C,L,D,M,W	1.1	T,H,S,G,N	0.6	R,E,K,A,W	0.1

Library E		Library F		Library G	
T175, N176, N177, Y178, L204	Enrichment	V209, W211, G213, K220, P221	Enrichment	H214, K215, V216, F217, M218	Enrichment
T,N,N,Y,M	3.6	V,W,G,R,P	3.1	N,V,E,V,G	8.9
T,N,N,Y,I	3.2	T,M,G,K,P	3.1	Y,R,I,G,F	8.6
T,N,N,Y,F	2.2	L,M,G,K,P	2.5	H,W,I,S,V	8.5
T,N,N,Y,V	1.7	H,N,H,N,E	2.2	S,A,I,W,V	8.3
T,N,N,Y,C	1.1	V,M,G,R,P	2.1	H,S,V,A,Y	8.1
S,T,R,Y,L	0.8	M,I,G,R,P	2.0	A,R,D,R,L	8.0
T,T,G,Y,L	0.8	N,P,Q,W,E	1.9	W,K,V,F,M	7.7
T,A,M,Y,L	0.6	L,L,G,K,P	1.9	E,W,V,W,W	7.2
T,H,R,R,R	0.4	M,L,G,K,Q	1.8	R,I,V,W,W	7.2
T,T,W,F,M	0.3	A,P,T,V,S	1.8	H,K,V,F,G	7.2

Figure 2-11. **Top 10 most enriched clones for each of the 7 TEVp first-generation libraries.**

	H214	K215	V216	F217	M218
F	0.0	ND	ND	0.6	-0.1
W	0.4	-0.1	-0.6	-0.2	-0.3
Y	-0.1	-0.5	-0.7	-0.2	-0.3
P	-0.6	ND	ND	-0.5	1.1
M	-0.6	-0.6	-0.7	-0.2	0.3
I	-0.4	-0.4	-0.1	-0.5	0.2
L	-0.3	-0.3	-0.3	-0.5	-0.2
V	-0.5	-0.1	0.2	0.2	0.2
A	-0.2	-0.4	-0.1	-0.2	-0.7
G	0.1	0.5	-0.5	-0.1	0.4
C	-0.5	-0.3	ND	-0.6	-0.4
S	-0.3	0.2	-0.6	-0.1	-0.2
T	ND	-0.6	-0.2	0.2	-0.3
N	0.6	0.0	-0.7	-0.8	ND
Q	ND	-0.4	0.4	-0.7	-0.1
D	ND	-0.4	-0.2	-0.5	-0.3
E	0.1	0.3	0.8	-0.5	-0.5
H	0.2	-0.4	ND	-0.3	ND
K	0.7	0.5	-0.4	ND	-0.2
R	-0.2	0.0	0.2	-0.4	-0.5

Figure 2-12. **Initial heat map for library G using a limited number of reads.** ND indicates a mutation that was not found in the unsorted population, and thus an enrichment was not determined.

The second-generation library was constructed with diversity at 10 sites including the aforementioned compelling mutations, wild-type, and a few amino acids resulting from the genetic code with constrained degenerate codon design (Table 2-6), which yields a theoretical diversity of 1 million variants. The library insert, created by fusion PCR of three fragments (Figure 2-3), was homologously recombined with pCT vector upon electroporation into yeast, which yielded 20 million transformants.

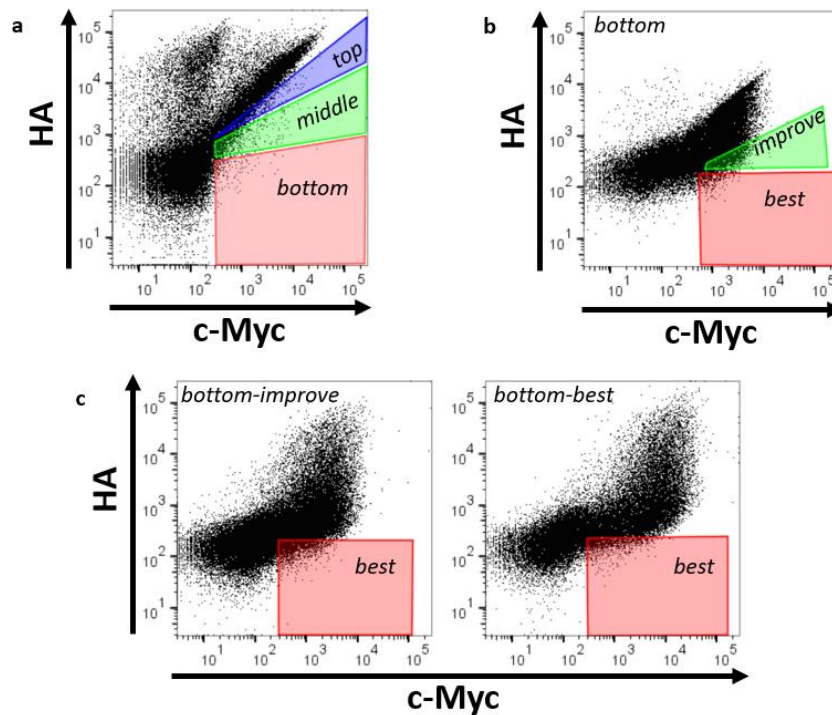


Figure 2-13. Flow cytometric analysis of first (a), second (b), and third (c) rounds of sorting of second-generation TEVp library. For the first sort, cells collected in a *bottom* red gate were propagated for a second sort. *Top* (blue) and *middle* (green) gates from the first sort collected mutants with sub-optimal activity; plasmid DNA was later extracted from these populations to create and test the activity of a limited number of *top* and *middle* gate mutants to further verify the yeast display construct's ability to differentiate mutants based on activity. For the second sort, cells were collected in the *improve* (green) or *best* (red) gates. These *bottom-improve* and *bottom-best* cells were then sorted once more, and another *best* gate (red) was used to collect cells for each.

The most active TEVp variants were enriched via three consecutive rounds of sorting (Figure 2-13, Table 2-7). The plasmid DNA of the *bottom-improve-best* (BIB) and *bottom-best-best* (BBB) triply-sorted populations was extracted and deep sequenced (Table 2-8). 30 mutants from the unsorted population were also Sanger sequenced to



estimate amino acid frequencies at each mutated location; these values agreed well with the theoretical amino acid frequencies (Figure 2-14), which were then used in sitewise enrichment score calculations (Figure 2-15). The top 100 most enriched clones in both BIB and BBB populations were separately analyzed for sitewise mutational preferences (Figure 2-15). Together, this sitewise analysis elucidated clear preferences for wild-type at sites 214, 216, 218, and 220, and clear mutational preferences of T30I and L32V. Two non-wild-type options were preferred at sites 31 (T and C) and 211 (I and M). While the mutational preferences at sites 204 and 215 were less clearly differentiated, the combined evidence suggests a preference for M or wild-type L at site 204 and non-wild-type E or R at site 215.

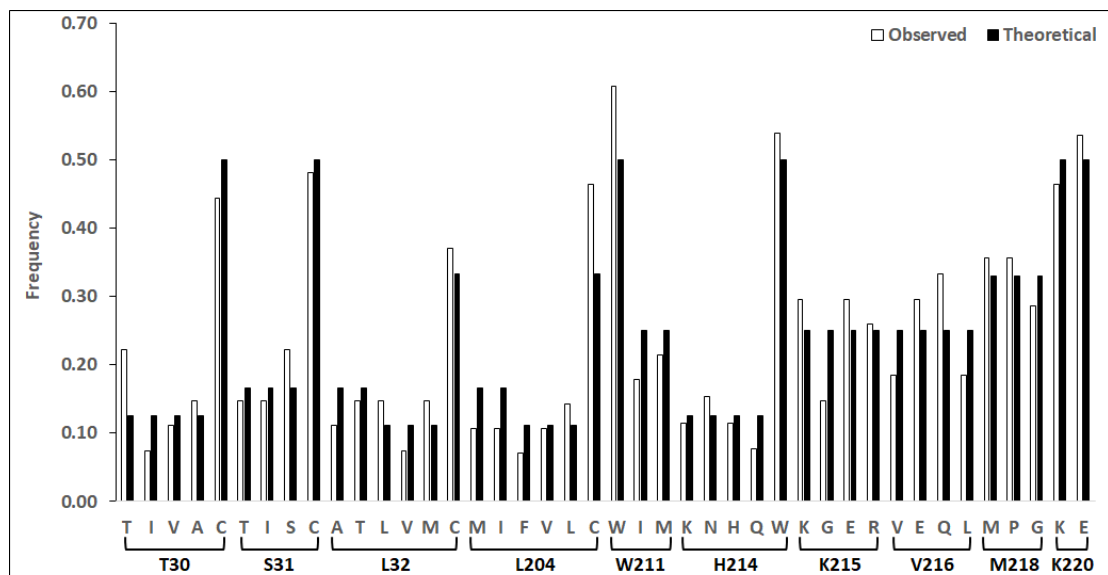


Figure 2-14. **Second-generation naive library analysis.** Theoretical amino acid frequencies compared to observed frequencies (via limited Sanger sequencing of 30 samples) for each position (T30, S31, L32, L204, W211, H214, K215, V216, M218, K220) of the second-generation TEVp library. Median difference between design frequency and observed frequency was 4%.

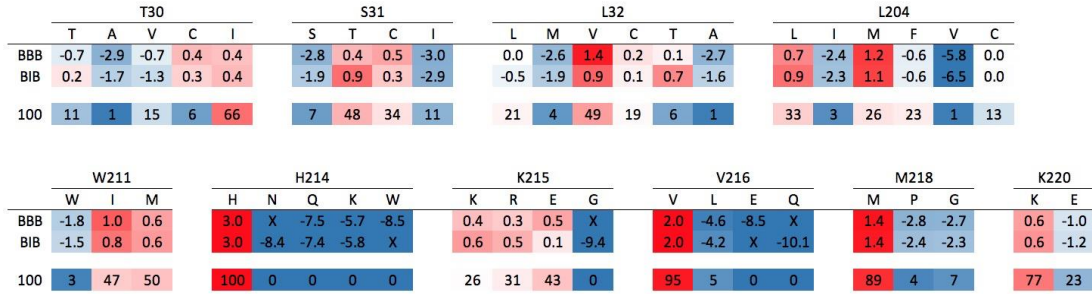


Figure 2-15. Heat map of sitewise enrichment for second-generation TEVp library. BBB and BIB log<sub>2</sub> enrichment values for every residue at every site in the second-generation TEVp library. An additional analysis was performed on the 100 most enriched clones for just the BBB population. These top 100 clones were assigned a value equal to their enrichment versus the 100<sup>th</sup>-ranked clone. For every residue at every site in the library, a weighted summation using these values was performed for the top 100 clones. At a given site, summation values for a particular amino acid were normalized so that the addition of all normalized values for every amino acid at a given site added up to 100. Optimal mutations are highlighted in dark red, and the most disfavored mutations are shown in dark blue.

	Cells Evaluated (millions)	c-Myc+ %	Cells Collected (% of c-Myc+)
First Sort	76	18	217,203 (1.6)
Second Sort	4.1	52	<i>Improve</i> – 154,719 (7.4) <i>Best</i> – 21,351 (1.0)
Third Sort – <i>Bottom Improve</i>	1.0	43	48,148 (11)
Third Sort – <i>Bottom Best</i>	1.1	52	33,138 (5.8)

Table 2-7. FACS data from the second-generation TEVp library.

	Reads
BIB	274,288
BBB	140,054

Table 2-8. Deep sequencing reads for the BIB and BBB populations from the second-generation TEVp library.

		30	31	32	151	204	211	214	215	216	218	220	kcat (1/s)	Km (uM)	Cat Eff (1/s/uM)
Wild-type	TEVp1	T	S	L	C	L	W	H	K	V	M	K	0.39±0.04	24.3±4.2	0.016±0.001
Sitewise Frequency	TEVp2	I	T	V	•	M	I	•	R	•	•	•	0.26±0.01	5.5±0.5	0.047±0.003
	TEVp3	I	T	V	•	M	M	•	E	•	•	•	0.77±0.04	22.1±4.7	0.035±0.006
	TEVp4	I	T	V	•	M	M	•	R	•	•	•	0.19±0.01	4.7±0.5	0.040±0.004
	TEVp5	I	T	V	•	•	I	•	E	•	•	•	0.39±0.01	10.9±0.8	0.036±0.002
	TEVp6	I	T	V	•	•	I	•	R	•	•	•	0.075±0.004	2.9±1.5	0.029±0.013
	TEVp7	I	T	V	•	•	M	•	E	•	•	•	0.20±0.003	10.7±0.3	0.019±0.0003
	TEVp8	I	T	V	•	•	M	•	R	•	•	•	0.17±0.004	4.3±0.04	0.038±0.0007
	TEVp9	I	C	V	•	M	I	•	E	•	•	•	0.16±0.01	5.4±0.6	0.029±0.002
	TEVp10	I	C	V	•	M	I	•	R	•	•	•	0.22±0.01	4.8±0.1	0.046±0.003
	TEVp11	I	C	V	•	M	M	•	E	•	•	•	0.88±0.08	19.7±2.7	0.045±0.002
	TEVp12	I	C	V	•	M	M	•	R	•	•	•	0.079±0.006	2.3±0.4	0.035±0.004
	TEVp13	I	C	V	•	•	I	•	E	•	•	•	0.18±0.01	7.0±1.2	0.026±0.004
	TEVp14	I	C	V	•	•	M	•	E	•	•	•	0.081±0.007	7.4±1.6	0.011±0.002
	TEVp15	I	C	V	•	•	M	•	R	•	•	•	0.076±0.002	2.6±0.4	0.030±0.005
	Sitewise Frequency (Failed Production)	TEVp16	I	T	V	•	M	I	•	E	•	•	•		
	TEVp17	I	C	V	•	•	I	•	R	•	•	•			
Epistatic	TEVp18	I	T	C	•	M	I	•	E	•	•	•	0.058±0.009	4.8±3.2	0.014±0.006
	TEVp19	I	T	C	•	M	I	•	R	•	•	•	0.037±0.012	3.3±0.8	0.011±0.001
	TEVp20	I	T	C	•	•	I	•	E	•	•	•	0.027±0.008	9.3±8.0	0.007±0.009
	TEVp21	I	T	C	•	•	I	•	R	•	•	•	0.083±0.001	2.8±0.3	0.029±0.002
F204 (TEVp20) and E220 (TEVp21)	TEVp22	I	•	•	•	F	M	•	•	•	G	•	0.29±0.04	88.5±13.5	0.0033±0.0003
	TEVp23	V	•	•	•	M	M	•	R	•	•	E	0.71±0.02	45.8±1.8	0.016±0.0001
Active Site Cysteine Replacements (C151A)	TEVp24	I	C	V	A	F	M	•	R	•	•	•			
	TEVp25	C	T	V	A	•	M	•	E	•	•	•			
	TEVp26	I	C	V	A	•	M	•	E	•	•	•			
	TEVp27	I	T	C	A	•	M	•	E	•	•	•			
Inactive Mutants from Sort (Top Gate)	TEVp28	V	T	C	•	C	M	W	R	•	•	E			
	TEVp29	V	C	C	•	C	•	N	•	•	G	•			
	TEVp30	V	I	M	•	•	•	W	E	E	•	•			
Inactive Mutants from Sort (Middle Gate)	TEVp31	C	C	M	•	M	•	•	•	•	G	E			
	TEVp32	C	C	T	•	•	•	W	•	•	•	•			
	TEVp33	V	C	C	•	M	•	Q	•	L	•	•			
	TEVp34	•	C	C	•	I	•	W	G	•	P	E			
	TEVp35	I	T	T	•	C	•	W	•	E	G	E			
	TEVp36	C	C	•	•	•	M	•	•	•	•	•			
	TEVp37	C	•	•	•	C	I	Q	R	•	G	E			
	TEVp38	V	T	T	•	C	M	W	•	•	•	•			

Table 2-9. List of TEVp mutants produced in *E. coli*.

#### 2.4.4 TEVp mutant production and characterization from sitewise analysis

This sitewise analysis motivated the production of 16 TEVp mutants (TEVp2-17 in Table 2-9, where wild-type is denoted as TEVp1), with sites 30 (I) and 32 (V) fixed as their dominant mutations, and two options at sites 31 (T or C), 204 (M or L), 211 (I or M), and 215 (E or R). TEVp variants were expressed from a pET vector in NEB5α *E. coli* for 5 hrs at 25 °C. Cell lysate was analyzed by SDS-PAGE, along with carbonic anhydrase and bovine serum albumin reference proteins, to quantify TEVp concentration via densitometry (Figure 2-16). All TEVp variants except TEVp16-17 were effectively produced. Enzymatic activity of variants was measured with a peptide substrate with a Dnp-quenched 2-Abz fluorophore (2-Abz-ENLYFQSGTK-Dnp;  $\lambda_{ex} = 320$  nm,  $\lambda_{em} = 420$  nm) (Figure 2-17, Table 2-9).

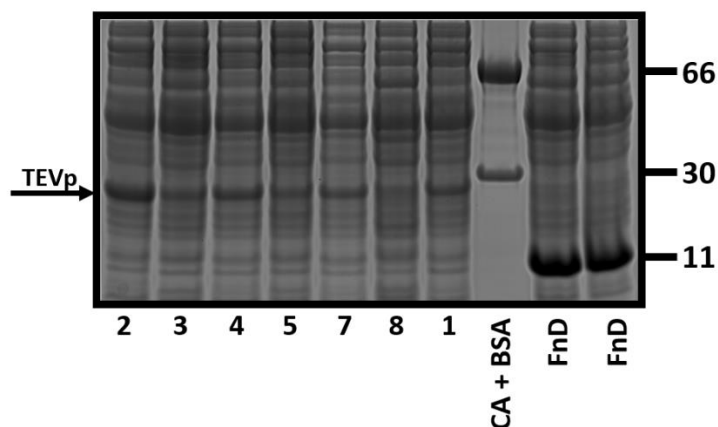


Figure 2-16. **SDS-PAGE analysis to quantify TEVp production yield.** Example of an SDS-PAGE of *E. coli* lysate from productions of TEVp variants 1-5 and 7-8, along with a well with both carbonic anhydrase (CA) and bovine serum albumin (BSA) reference protein and two wells of TEVp-deficient *E. coli* lysate from producing fibronectin D (FnD). The lack of any band in the 28 kDa region for the FnD samples confirms that the 28 kDa band in the TEVp 1-5 and 7-8 lanes is specific to TEVp.

Twelve of the 14 novel mutants showed statistically significantly improved catalytic efficiency compared to wild-type TEVp1 (Figure 2-17a-b). The three most active clones were hexamutants: TEVp2, TEVp10, TEVp11, which demonstrated up to 3-fold improvement in  $k_{cat}/K_M$  over wild-type (respective fold change of  $3.0 \pm 0.3$ ,  $2.9 \pm 0.2$ ,  $2.8 \pm 0.2$ ); TEVp2 and TEVp10 owe this improvement to  $4.4 \pm 0.7$ -fold and  $5.0 \pm 0.8$ -fold improvement in  $K_M$  (offset by a respective  $1.5 \pm 0.2$ -fold and  $1.7 \pm 0.2$ -fold hindrance in  $k_{cat}$ ), and TEVp11 is aided by both a  $2.3 \pm 0.3$ -fold improvement in  $k_{cat}$  and a slight decrease of  $1.2 \pm 0.2$ -fold in the  $K_M$ . All 14 mutants exhibited reduced  $K_M$  (with 12 of 14 being statistically significant) (Figure 2-17c). Conversely, only two mutants robustly improved  $k_{cat}$  (Figure 2-17d). Notably, these were the only two clones with L204M, W211M, and K215E.

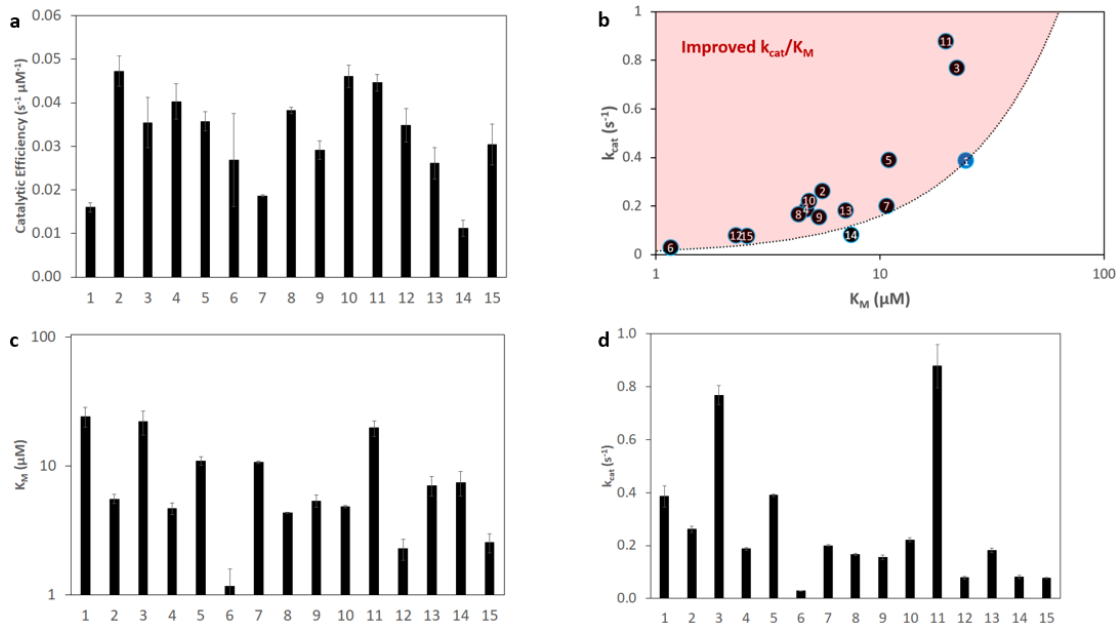


Figure 2-17. **Comparison of enzymatic parameters of TEVp 2-15 to wild-type enzyme (TEVp1).** a) Catalytic efficiency ( $s^{-1} \mu M^{-1}$ ) for TEVp clones 1-15. b) Graph of  $k_{cat}$  vs.  $K_M$  for TEVp clones 1-15, with TEVp clones (labeled in the marker with the mutant number) falling on the dotted line maintaining, and those above the line improving from, wild-type catalytic efficiency. c)  $K_M$  ( $\mu M$ ) and d)  $k_{cat}$  ( $s^{-1}$ ) for TEVp clones 1-15. Error bars are 95% confidence intervals.

To elucidate the impact of each mutation in the context of the most efficient mutant, TEVp2, we examined the enzymatic activity of all six single-mutation reversions (Figure 2-18, Table 2-10). All six reversions reduced catalytic efficiency thereby supporting the benefit of combining all six mutations. Five of the six reversions had a dramatic negative impact (fold change ranging from  $3.2 \pm 0.4$  to  $9.0 \pm 0.4$ ) on the turnover number albeit with moderate benefit to the Michaelis-Menten constant. The exception is the V32L reversion, which increased  $k_{cat}$   $1.3 \pm 0.1$ -fold and hindered  $K_M$   $1.8 \pm 0.2$ -fold. These reversion results are intriguing in that TEVp2 achieves a  $3.0 \pm 0.3$ -fold increase in catalytic efficiency relative to wild-type via a  $4.4 \pm 0.7$ -fold reduction in  $K_M$  but a  $1.5 \pm 0.2$ -fold reduction in  $k_{cat}$ . Yet the reversion mutants predominantly hinder  $k_{cat}$  and aid  $K_M$ . Thus, the mutations act collectively to aid  $K_M$  while limiting detriment to  $k_{cat}$ .

	kcat (1/s)	Km (uM)	Cat Eff (1/s/uM)
TEVp2	0.26±0.01	5.5±0.5	0.047±0.003
TEVp2 I30T	0.06±0.01	3.5±0.9	0.018±0.004
TEVp2 T31S	0.04±0.003	1.8±1.1	0.026±0.011
TEVp2 V32L	0.33±0.01	9.7±0.6	0.034±0.003
TEVp2 M204L (TEVp6)	0.03±0.0004	1.2±0.4	0.027±0.011
TEVp2 I211W	0.05±0.004	3.3±0.5	0.015±0.001
TEVp2 R215K	0.08±0.01	4.5±1.5	0.019±0.003

Table 2-10. Summary of the effect on enzymatic parameters of TEVp2 single-mutant reversion to wild-type.

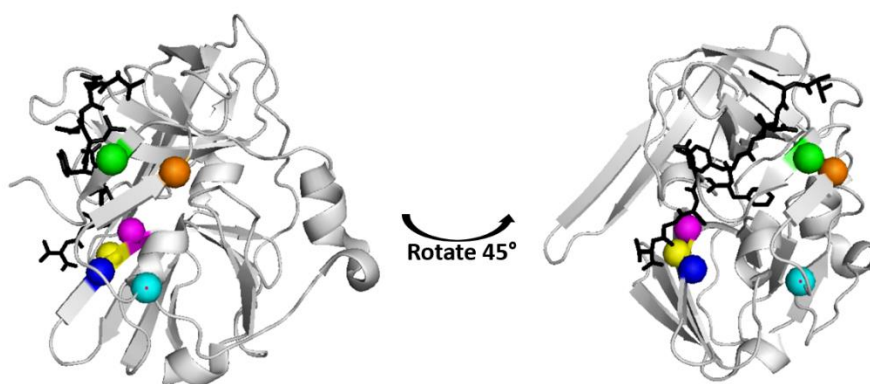


Figure 2-18. Illustration of sites mutated in optimal variant TEVp2. PyMOL image of inactivated C151A TEVp complexed with its substrate ENLYFQ↓S (black), with the  $\alpha$ -carbon of the mutated amino acid residues in TEVp2 highlighted: T30I (blue), S31T (yellow), L32V (magenta), L204M (cyan), W211I (orange), K215R (green).

#### 2.4.5 TEVp mutant production and characterization from epistatic analysis

An epistatic analysis was performed for both BIB and BBB populations to evaluate the mutational preferences at various positions in context with particular mutations at other specific sites (Figure 2-19). While the epistatic analysis largely bolstered the rationale for producing the aforementioned 16 TEVp variants, it also suggested a possible preference for L32C, as opposed to L32V, when S31 is mutated to threonine. As there also appeared to be a mild preference for W211I over W211M with the S31T and L32C mutations, four more TEVp variants (TEVp18-21 in Table 2-9), with four sites fixed (I30, T31, C32, I211) and two options at sites 204 (M or L) and 215 (E or

R), were produced and characterized. Only one (TEVp21) of the four C32-containing TEVp mutants (TEVp18-21) yielded a meaningful improvement over wild-type (Figure 2-20). In the context of T30I, S31T, L204M, W211I, K215E, the L32C mutation enables production whereas the L32V did not; yet the catalytic efficiency is nominally lower than wild-type. When this context is modified at K215 to K215R,  $k_{cat}/K_M$  is nominally lower for the L32C variant whereas the L32V exhibited the optimal activity. Yet, when the context is modified further to retain wild-type at L204 (*i.e.*, T30I, S31T, W211I, K215R), the L32V and L32C variants are equivalently effective with ~8-fold improved  $K_M$  and hindered  $k_{cat}$  to yield  $k_{cat}/K_M$  1.8-fold superior to wild-type.

		S31				L32				L204				W211		K215							
		T	C			V	C			M	L			I	M	E	K		R				
T30	I	0.7	1.4	0.9	0.7	1.5	1.5	1.3	0.7	1.1	1.2	1.4	1.2	1.4	1.3	0.8	0.9	1.2	0.4	0.8	0.8	0.3	1.2
S31	T					0.1	0.9	1.2	1.2	1.2	1.7	1.3	1.8	1.3	1.7	0.8	1.1	0.9	1.1	1.1	1.6	0.4	1.1
	C					2.3	1.6	0.4	0.2	1.8	1.5	1.2	1.2	1.5	1.1	1.1	1.1	1.1	0.4	0.8	0.9	0.9	0.9
L32	V									2.5	2.3	2.0	1.6	2.3	1.8	2.0	1.6	1.9	1.3	1.8	1.5	1.7	1.1
	C									1.1	1.1	1.2	1.3	1.2	1.0	0.6	0.5	0.7	0.1	0.6	0.6	0.5	0.7
L204	M													2.2	2.1	1.8	1.8	1.4	1.3	1.6	1.7	1.7	1.6
	L													1.2	1.1	0.8	1.3	0.8	0.4	1.7	1.9	0.5	1.1
W211	I																	1.7	1.2	1.0	1.0	1.3	1.5
	M																	1.2	1.0	0.6	0.9	1.1	1.2

Figure 2-19. **Epistatic analysis.** Log<sub>2</sub> enrichment values for double mutants in the BBB (left) and BIB (right) populations for select compelling residues at sites of interest in the second-generation TEVp library.

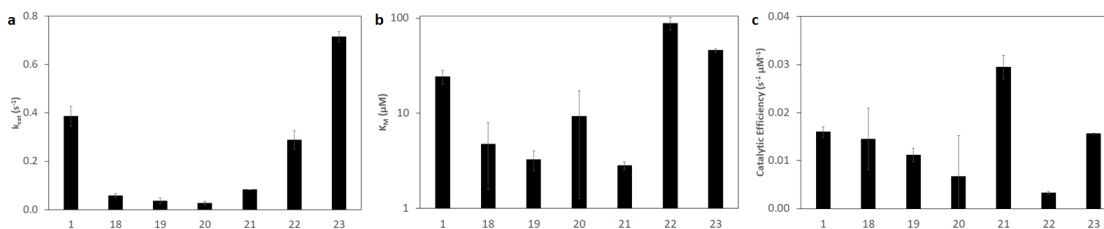


Figure 2-20. **Comparison of enzymatic parameters of TEVp 18-23 to wild-type enzyme (TEVp1).** a)  $k_{cat}$  ( $s^{-1}$ ), b)  $K_M$  ( $\mu M$ ), and c) catalytic efficiency ( $s^{-1} \mu M^{-1}$ ) for TEVp clones 1 and 18-23. Error bars are 95% confidence intervals.

In evaluating TEVp variants 2-15 and 18-21, a mostly consistent pattern emerged regarding the enzymatic effects of one mutation over another at a particular position when evaluated in the context of an otherwise fixed TEVp sequence (Figure 2-21). For

the V32-containing variants TEVp2-15, the turnover number  $k_{\text{cat}}$  was generally aided at the expense of  $K_M$  for S31T (vs. S31C), L204M (vs. wild-type leucine), and K215E (vs. K215R), while any correlation for C32-containing variants TEVp18-21, applicable only for L204M vs. wild-type leucine and K215E vs. K215R, was less clearly determinable. W211I (vs. W211M) caused  $k_{\text{cat}}$  and  $K_M$  to generally move in the same direction, either both increasing or both decreasing depending on the context. For L32V (vs. L32C), two of the three cases demonstrated an improved  $k_{\text{cat}}$  with minimal hindrance to  $K_M$ , and in the third cases there was minimal impact on both parameters.

Two additional mutations, F204 and E220, were relatively frequent, albeit not dominant, in the pool of functional variants (Figure 2-15). Thus, the F204- and E220-containing clones (TEVp22 and TEVp23, respectively, in Table 2-9) that exhibited the highest enrichment were also selected for production and characterization. TEVp22 dropped the catalytic efficiency by  $4.9 \pm 0.5$ -fold from wild-type and TEVp23 had a catalytic efficiency equivalent to wild-type with a trade-off in improved  $k_{\text{cat}}$  ( $1.9 \pm 0.2$ -fold) offset by an increase in  $K_M$  ( $1.9 \pm 0.3$ -fold) (Figure 2-20).



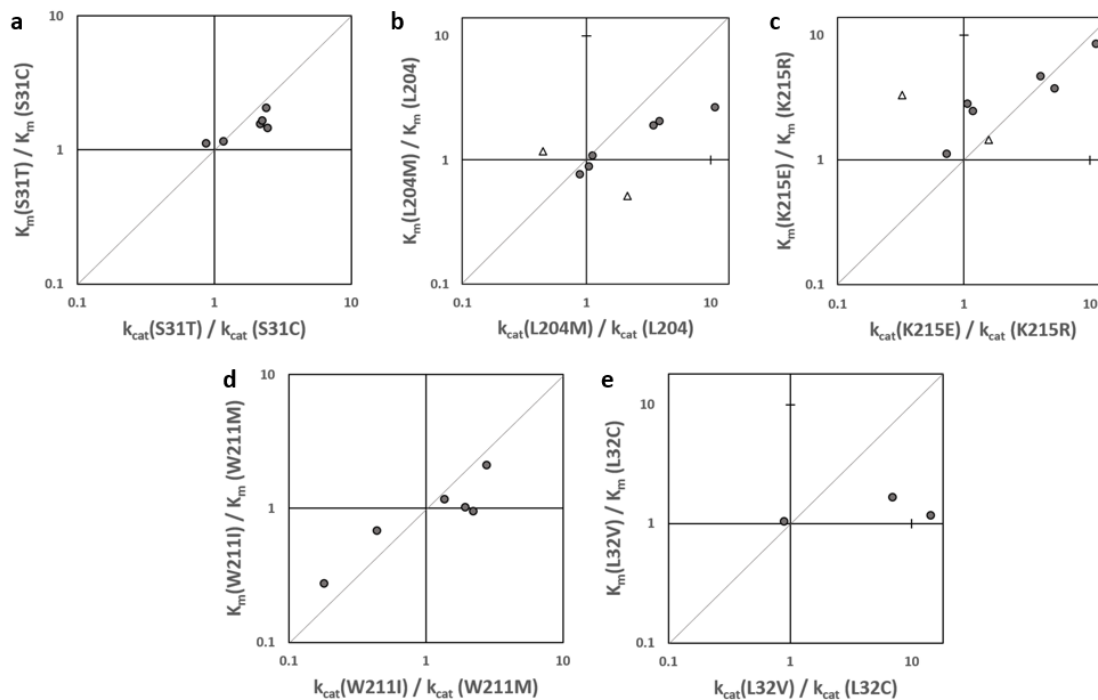


Figure 2-21. **Summary of the effect on enzymatic parameters of singular mutations when evaluated in the context of a various TEVp clones.** V32-containing variants TEVp2-15 and C32-containing variants TEVp18-21 are shown in grey circles and white triangles, respectively.

#### 2.4.6 Investigation of potential auxiliary cysteine protease activity at sites 30-32

Due to the strong enrichment of cysteine at sites 30-32 and the relative proximity of these sites to the catalytic triad cysteine (C151) and the substrate (specifically the cleavage point, Figure 2-22), it was hypothesized that cysteine was enriched at sites 30-32 because it possibly provided supplementary cysteine proteolytic activity. Of the deep sequenced clones that contained cysteine in these sites, the top four most enriched variants (TEVp24-27 in Table 2-9) were selected for production with C151A mutation; TEVp26 coincidentally also represented the C151A variant of TEVp14, which conveniently enabled a direct comparison to determine if supplemental proximal cysteines can preserve, or prevent the absence of, cysteine proteolytic activity when the wild-type active site is inactivated. Evaluating reaction velocity at a substrate concentration of 50  $\mu\text{M}$  and an enzyme concentration of 0.1  $\mu\text{M}$ , it is evident that

TEVp24-27 registered no activity (Figure 2-23). This provides substantial evidence to disprove the hypothesis that supplemental cysteine protease activity can be engineered into TEVp at sites 30-32.

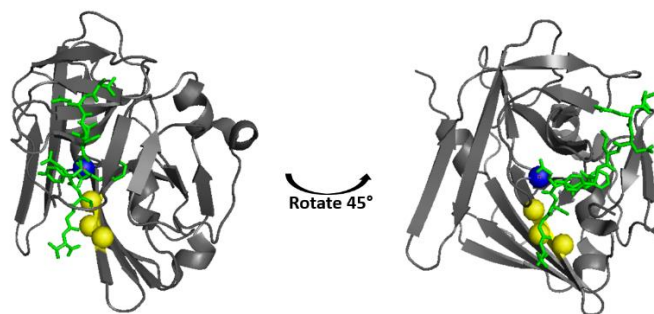


Figure 2-22. **Illustration of proximity of active site cysteine (C151) to sites 30-32.** PyMOL image of inactivated C151A TEVp complexed with its substrate ENLYFQ↓S (green), with the  $\alpha$ -carbon of the amino acid residues at positions 30-32 shown in yellow and at position 151 shown in blue.

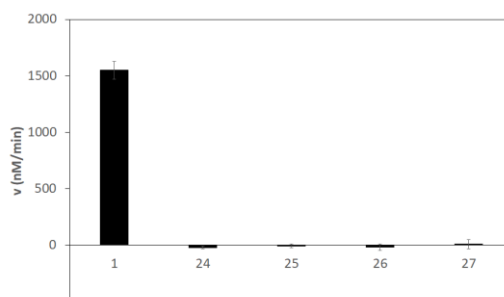


Figure 2-23. **Comparison of enzymatic activity of TEVp 24-27 to wild-type enzyme (TEVp1).** Substrate and enzyme concentrations were 50  $\mu$ M and 0.1  $\mu$ M, respectively, and error bars are 95% confidence intervals.

#### 2.4.7 Enzyme-substrate co-display system successfully stratifies protease mutants by activity

To further demonstrate the yeast surface display construct's ability to differentiate mutants based on activity, plasmid DNA was extracted from cells in the *top* and *middle* gates from the second-generation library's first sort (Figure 2-13). Three random clones from *top* and eight from *middle* (TEVp28-30 and TEVp31-38, respectively, in Table 2-9) were isolated for evaluation. There was no activity or minimal activity for the three clones from the *top* gate (TEVp28-30) and six of the eight TEVp variants from the *middle* gate (TEVp clones 31-32, 34-35, 37-38); of the remaining two *middle* gate TEVp

variants, one (TEVp36) showed activity comparable to wild-type (TEVp1) and the other (TEVp33) showed above-negligible activity, though significantly below wild-type (Figure 2-24). Compared to the improved activity of TEVp mutants 2-15, this provides further evidence that the TEVp/substrate fusion display construct effectively stratifies yeast cells on FACS based on their encoded TEVp from low to high enzymatic activity when selecting for high to low HA:c-Myc ratio, respectively.

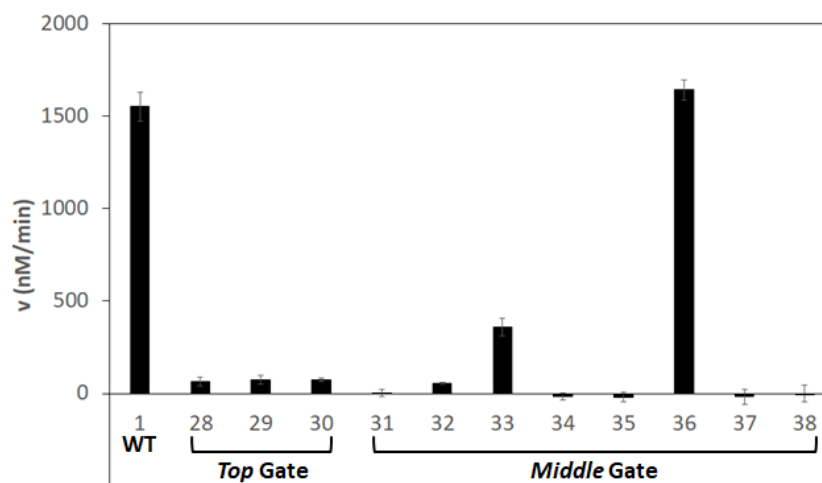


Figure 2-24. **Comparison of enzymatic activity of *top* and *middle* gate mutants to wild-type enzyme (TEVp1).** Three *top* gate mutants (TEVp28-30) and eight *middle* gate mutants (TEVp31-38) were analyzed at substrate and enzyme concentrations of 50  $\mu$ M and 0.1  $\mu$ M, respectively. Error bars shown are 95% confidence intervals.

#### 2.4.8 Extension of the enzyme-substrate co-display system to sortase enzyme

The generalizability of this enzyme/substrate fusion protein display construct in differentiating enzymatic activity was also examined. The genes for the TEVp enzyme and substrate from the SAT construct were swapped out for a heptamutant version of sortase (Srt7) and its substrate (LPET↓GG, with the final G not part of the canonical substrate), respectively. SrtA can function as a protease cutting the substrate LPET↓G if water is used as a nucleophile instead of an oligoglycine molecule, resulting in only a 41% decrease in catalytic efficiency<sup>103</sup>. Specifically, we examined an N-terminal

truncated (residues 60-206) SrtA heptamutant (five mutations – P94R/D160N/D165A/K190E/K196T – for enhanced activity and two more – E105K/E108A – for calcium-independence) with 11-fold improvement in catalytic efficiency compared to wild-type<sup>28,104-106</sup>. A negative control was also tested that replaced the canonical substrate with a scrambled version (TEGLGP). Yeast transformed with one of these plasmids were induced for 4 hrs and labeled with the same antibodies as before for flow cytometry analysis (Figure 2-25). As with the analogous SAT versions, 4 hrs of induction elicits standard surface display. There is also a clear shift in the displayed population of SAS compared to the negative control S<sub>sc</sub>AS (Table 2-11). Display levels of c-Myc appear equivalent between the two, but the SAS sample has a demonstrably weaker surface display of the HA tag, indicative of enzymatic cleavage of the tethered substrate. Further, a theoretical gate for collecting improved Srt7 variants can still be drawn below this population in the SAS sample. These results in total suggest that our system is possibly generalizable and that different proteolytic enzymes could be examined for the discovery of improved mutants using this yeast display framework.

	SAS	S <sub>sc</sub> AS
HA	6,148	34,445
c-Myc	1,947	2,033

Table 2-11. **Summary statistics of flow cytometry-assisted Srt7 activity profiling.** Median fluorescence signal originating from HA and c-Myc tags in the c-Myc+ population of SAS and S<sub>sc</sub>AS samples after 4 hours of SG-CAA induction.

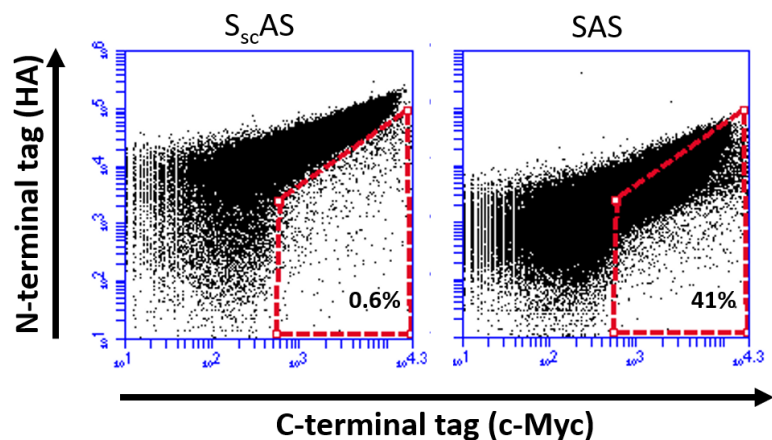


Figure 2-25. **Protease activity profiling of Srt7 via flow cytometry.** Flow cytometry plots of SAS (Srt7 with substrate LPETGG) and S<sub>sc</sub>AS (incorporating scrambled substrate TEGLGP in place of LPETGG) after 4 hrs induction. Cells were labeled with fluorescent antibodies to detect HA and c-Myc tags.

## 2.5 Discussion

We demonstrated in this study that co-display of a protease and its corresponding substrate on the same Aga2 yeast surface protein can be harnessed to differentiate cells based on the catalytic activity of the protease variant expressed. This extends the utility of yeast display for enzyme engineering beyond previous co-display constructs employed for bond-forming enzymes<sup>27-28</sup> via library-scale evolution of the concept initially published by Cochran and team<sup>27</sup>.

Seven first-generation saturation mutagenesis TEVp libraries, which were rationally designed based on structure to target the area around the active site, were sorted using the TEVp/substrate dual-display system to identify mutations that increased catalytic efficiency. A retrospective analysis was performed to evaluate sitewise mutational tolerance (Figure 2-9) in relation to expectations predicated on a variety of common metrics (relative solvent accessibility<sup>107</sup>, distance to the substrate cut site<sup>30,89</sup>, FoldX-predicted changes in stability<sup>56</sup>, substitution matrices such as BLOSUM62<sup>61</sup>, and consensus design<sup>39</sup>). The relative lack of predictive power for these metrics highlighted the need for thorough sampling of sequence space via the yeast display activity screen to

identify improved mutations. An analysis of the number of mutations at a particular site with positive enrichment versus that site's relative solvent accessibility (Figure 2-26) or distance to the glutamine residue in the TEVp substrate's cut site (Figure 2-27) both demonstrated no meaningful correlation. Assessment of the relationship between increased functional enrichment and computationally predicted stability (Figure 2-28) revealed weak nominal benefit of stability.

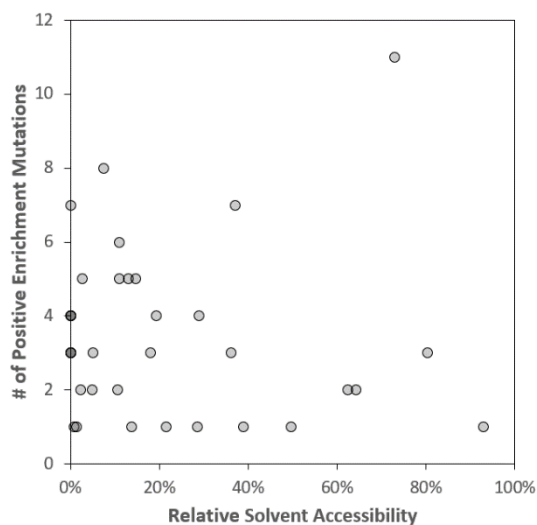


Figure 2-26. **Relative solvent accessibility is a poor predictor of mutational tolerance in the sites around TEVp's active site.** The number of mutations exhibiting positive enrichment in the first-generation TEVp library were obtained from the data in Figure 2-9. Relative solvent accessibility at a particular site was computed as the solvent-accessible surface area of the wild-type residue, obtained using the `get_area` command in PyMOL, divided by the empirical maximum possible solvent-accessible surface area of the same amino acid<sup>108</sup>.

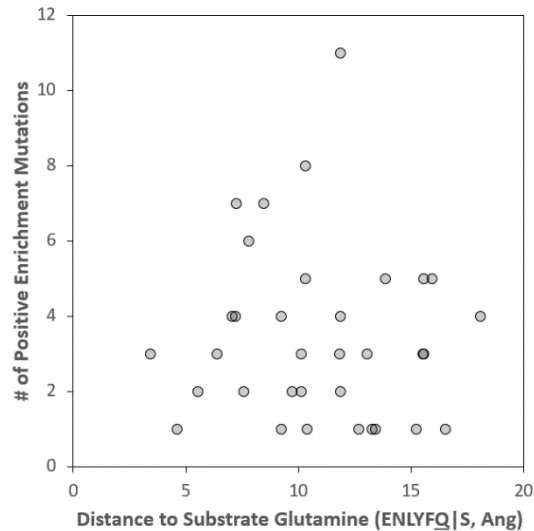


Figure 2-27. **Distance to the substrate cut site is a poor predictor of mutational tolerance in the sites around TEVp's active site.** The number of mutations exhibiting positive enrichment in the first-generation TEVp library were obtained from the data in Figure 2-9. Distance between residues was calculated in PyMOL.

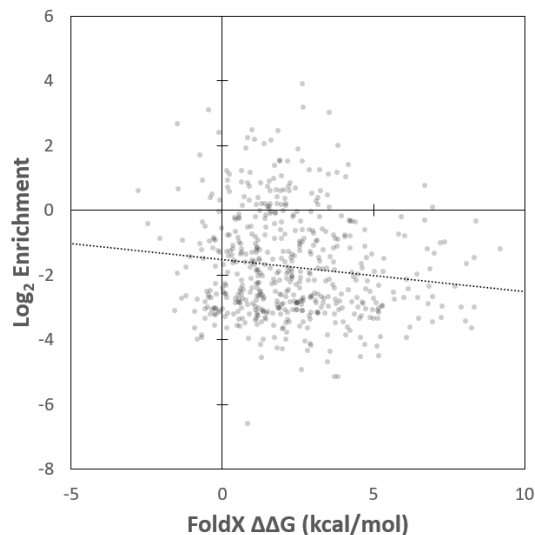


Figure 2-28. **FoldX-predicted stability changes from point mutations is a poor predictor of mutational tolerance in the sites around TEVp's active site.** Enrichment values in the first-generation TEVp library were obtained from the data in Figure 2-9.  $\Delta\Delta G$  values were calculated using the PDB file 1LVB<sup>102</sup> and the fourth version of FoldX<sup>56</sup>. 85% of data points possessed both  $\Delta\Delta G$  values below 10 kcal/mol and  $\log_2$  enrichment values above -8; the remaining 15% were excluded for analysis.

The overall relationship between increased functional enrichment and chemical homology appeared similarly weak (Figure 2-29). To evaluate the performance of homologous vs. non-homologous mutations in predicting beneficial mutations, mutations

with a score of 0 or greater in the BLOSUM62 matrix<sup>61</sup> were considered homologous. 9 of 182 (5.0%) homologous mutations, and 8 of 464 (1.7%) non-homologous mutations, were beneficial. Interestingly, while the general correlation between BLOSUM62 score and enrichment appears weak, homologous mutations with a BLOSUM62 score of 0 or greater were three times as likely to be beneficial compared to non-homologous mutations. It is also worth noting that all 6 mutations in TEVp2 are quite homologous: T30I, S31T, L32V, L204M, W211I, and K215R. Additionally, in evaluating all 17 beneficial mutations (Table 2-5), 53% of beneficial mutations were chemically homologous and 47% were non-homologous. Some examples of chemically homologous mutational preference include: S31 strongly preferring mutation to the neutral polar residue threonine, L32 preferring mutation to the nonpolar amino acids valine and methionine, and L204 strongly preferring mutation to the nonpolar residues methionine, isoleucine, phenylalanine, and valine. Yet, notable exceptions were also observed, such as T30 strongly preferring the nonpolar residue isoleucine, and significant positive enrichment for cysteine and threonine at site L32. Our results lend credence to the assertion that homologous mutations are more likely to be beneficial than non-homologous mutations<sup>109</sup>, but library designs that neglects non-homologous mutations will likely miss beneficial mutations that are potentially roughly equal in number to that generated by homologous mutations.



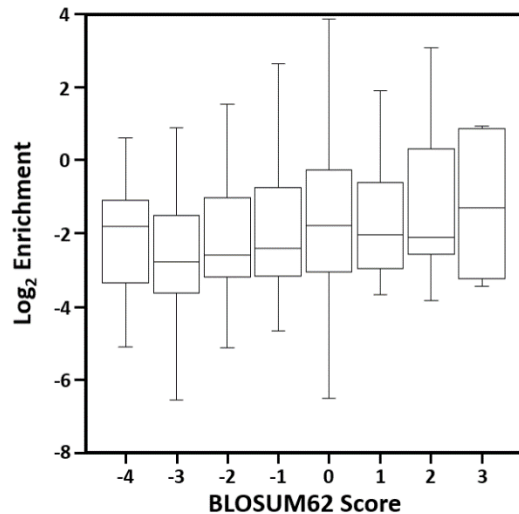


Figure 2-29. **BLOSUM62 score is a poor predictor of mutational tolerance in the sites around TEVp's active site.** Enrichment values in the first-generation TEVp library were obtained from the data in Figure 2-9. BLOSUM62 scores were obtained from literature<sup>61</sup>. Box and whisker plot shown with the bottom and top of the box indicating 1<sup>st</sup> and 3<sup>rd</sup> quartiles, respectively, with the line drawn in the box indicating the median enrichment. Whiskers extend to minimum and maximum data points that are still within the median plus or minus 1.5 times the interquartile range.

Consensus design<sup>39</sup> using the peptidase C4 family in Pfam<sup>110</sup> (PF00863) would not have efficiently identified the six sites that benefit from mutation that our two-generation strategy found; while 5 of 6 exhibit high mutational tolerance (63-96% non-wild-type, Table 2-12), this is also true of 14 of the other 28 sites in the proximal region. Moreover, W211 would not have been identified as it is 96% conserved in homologs. Beyond site choice, the particular beneficial amino acids would not have been identified by consensus design as only S31T (23%), L32M (40%) and L204F (11%) were present at more than 5% in homologs.

T30		S31		L32		L204		W211		K215	
A: 9	0.4%	A: 1	0.0%	A: 10	0.4%	A: 0	0.0%	A: 0	0.0%	A: 39	1.5%
C: 5	0.2%	C: 35	1.4%	C: 61	2.4%	C: 0	0.0%	C: 1	0.0%	C: 1	0.0%
D: 93	3.8%	D: 118	4.7%	D: 0	0.0%	D: 3	0.1%	D: 0	0.0%	D: 29	1.1%
E: 524	21.4%	E: 658	26.2%	E: 0	0.0%	E: 0	0.0%	E: 0	0.0%	E: 138	5.4%
F: 3	0.1%	F: 0	0.0%	F: 1	0.0%	F: 277	10.5%	F: 0	0.0%	F: 0	0.0%
G: 16	0.7%	G: 4	0.2%	G: 2	0.1%	G: 0	0.0%	G: 0	0.0%	G: 2	0.1%
H: 0	0.0%	H: 17	0.7%	H: 0	0.0%	H: 32	1.2%	H: 0	0.0%	H: 31	1.2%
I: 26	1.1%	I: 1	0.0%	I: 242	9.6%	I: 0	0.0%	I: 1	0.0%	I: 30	1.2%
K: 8	0.3%	K: 19	0.8%	K: 0	0.0%	K: 0	0.0%	K: 0	0.0%	K: 668	25.9%
L: 8	0.3%	L: 0	0.0%	L: 939	37.3%	L: 99	3.8%	L: 24	0.9%	L: 1	0.0%
M: 13	0.5%	M: 0	0.0%	M: 999	39.7%	M: 0	0.0%	M: 0	0.0%	M: 4	0.2%
N: 677	27.7%	N: 119	4.7%	N: 0	0.0%	N: 0	0.0%	N: 0	0.0%	N: 1,361	52.8%
P: 1	0.0%	P: 0	0.0%	P: 0	0.0%	P: 0	0.0%	P: 0	0.0%	P: 0	0.0%
Q: 53	2.2%	Q: 40	1.6%	Q: 4	0.2%	Q: 0	0.0%	Q: 0	0.0%	Q: 174	6.7%
R: 152	6.2%	R: 10	0.4%	R: 0	0.0%	R: 0	0.0%	R: 1	0.0%	R: 30	1.2%
S: 707	28.9%	S: 902	35.9%	S: 36	1.4%	S: 0	0.0%	S: 3	0.1%	S: 25	1.0%
T: 149	6.1%	T: 579	23.1%	T: 90	3.6%	T: 0	0.0%	T: 0	0.0%	T: 31	1.2%
V: 1	0.0%	V: 4	0.2%	V: 132	5.2%	V: 0	0.0%	V: 4	0.2%	V: 5	0.2%
W: 0	0.0%	W: 0	0.0%	W: 0	0.0%	W: 397	15.1%	W: 2,521	96.1%	W: 1	0.0%
Y: 1	0.0%	Y: 4	0.2%	Y: 0	0.0%	Y: 1,822	69.3%	Y: 68	2.6%	Y: 8	0.3%
X: 1	0.0%	X: 0	0.0%	X: 0	0.0%	X: 0	0.0%	X: 1	0.0%	X: 0	0.0%

N44	F139	K141	H167	S168	A169	T146	K147	D148	G149	Q150	S170	N171	F172
A: 0	0.0%	A: 0	0.0%	A: 0	0.0%	A: 1	0.0%	A: 89	3.4%	A: 6	0.2%	A: 16	0.6%
C: 0	0.0%	C: 0	0.0%	C: 67	2.5%	C: 0	0.0%	C: 3	0.1%	C: 0	0.0%	C: 0	0.0%
D: 0	0.0%	D: 0	0.0%	D: 0	0.0%	D: 0	0.0%	D: 4	0.2%	D: 0	0.0%	D: 311	11.8%
E: 0	0.0%	E: 0	0.0%	E: 0	0.0%	E: 0	0.0%	E: 0	0.0%	E: 6	0.2%	E: 159	6.0%
F: 0	0.0%	F: 2,255	85.8%	F: 0	0.0%	F: 0	0.0%	F: 0	0.0%	F: 3	0.1%	F: 13	0.5%
G: 0	0.0%	G: 0	0.0%	G: 12	0.5%	G: 0	0.0%	G: 25	0.9%	G: 41	1.6%	G: 19	0.7%
H: 0	0.0%	H: 0	0.0%	H: 2,635	99.5%	H: 0	0.0%	H: 0	0.0%	H: 4	0.2%	H: 23	0.9%
I: 0	0.0%	I: 4	0.2%	I: 371	14.1%	I: 0	0.0%	I: 7	0.3%	I: 12	0.5%	I: 0	0.0%
K: 0	0.0%	K: 1,843	70.1%	K: 0	0.0%	K: 0	0.0%	K: 1	0.0%	K: 1,191	45.2%	K: 8	0.3%
L: 0	0.0%	L: 110	4.2%	L: 0	0.0%	L: 0	0.0%	L: 2,358	89.2%	L: 0	0.0%	L: 2	0.1%
M: 1	0.0%	M: 90	3.4%	M: 6	0.2%	M: 0	0.0%	M: 27	1.0%	M: 0	0.0%	M: 0	0.0%
N: 2,285	95.0%	N: 0	0.0%	N: 0	0.0%	N: 31	1.2%	N: 18	0.7%	N: 0	0.0%	N: 344	13.1%
P: 378	14.2%	P: 0	0.0%	P: 0	0.0%	P: 0	0.0%	P: 2	0.1%	P: 3	0.1%	P: 153	5.8%
Q: 0	0.0%	Q: 0	0.0%	Q: 101	3.8%	Q: 0	0.0%	Q: 0	0.0%	Q: 375	14.2%	Q: 2	0.1%
R: 0	0.0%	R: 78	3.0%	R: 67	2.5%	R: 7	0.3%	R: 2	0.1%	R: 0	0.0%	R: 7	0.3%
S: 0	0.0%	S: 0	0.0%	S: 64	2.4%	S: 1	0.0%	S: 2,108	79.6%	S: 16	0.6%	S: 96	3.6%
T: 3	0.1%	T: 0	0.0%	T: 21	0.8%	T: 1	0.0%	T: 8	0.3%	T: 0	0.0%	T: 41	1.6%
V: 2	0.1%	V: 3	0.1%	V: 31	1.2%	V: 0	0.0%	V: 72	2.7%	V: 90	3.4%	V: 0	0.0%
W: 0	0.0%	W: 0	0.0%	W: 0	0.0%	W: 0	0.0%	W: 1	0.0%	W: 0	0.0%	W: 0	0.0%
Y: 0	0.0%	Y: 89	3.4%	Y: 1	0.0%	Y: 0	0.0%	Y: 0	0.0%	Y: 0	0.0%	Y: 12	0.5%
X: 0	0.0%	X: 0	0.0%	X: 0	0.0%	X: 3	0.1%	X: 0	0.0%	X: 0	0.0%	X: 14	0.5%

A173	N174	T175	N176	N177	Y178	V209	G213	K220	P221	H214	V216	F217	M218
A: 123	4.7%	A: 4	0.2%	A: 21	0.8%	A: 13	0.5%	A: 21	0.8%	A: 0	0.0%	A: 3	0.1%
C: 1	0.0%	C: 1	0.0%	C: 0	0.0%	C: 8	0.3%	C: 0	0.0%	C: 4	0.2%	C: 57	2.2%
D: 318	12.1%	D: 105	4.0%	D: 195	7.4%	D: 9	0.3%	D: 1	0.0%	D: 0	0.0%	D: 11	0.4%
E: 45	1.7%	E: 3	0.1%	E: 254	9.6%	E: 103	3.9%	E: 22	0.8%	E: 60	2.3%	E: 0	0.0%
F: 3	0.1%	F: 5	0.2%	F: 0	0.0%	F: 0	0.0%	F: 0	0.0%	F: 846	31.9%	F: 0	0.0%
G: 9	0.3%	G: 262	10.0%	G: 57	2.2%	G: 101	3.8%	G: 2	0.1%	G: 0	0.0%	G: 449	17.3%
H: 23	0.9%	H: 229	8.7%	H: 0	0.0%	H: 47	1.8%	H: 1	0.0%	H: 3	0.1%	H: 1	0.0%
I: 11	0.4%	I: 4	0.2%	I: 10	0.4%	I: 695	26.2%	I: 1	0.0%	I: 13	0.5%	I: 1,555	59.2%
K: 155	5.9%	K: 51	1.9%	K: 67	2.5%	K: 403	15.2%	K: 0	0.0%	K: 14	0.5%	K: 0	0.0%
L: 7	0.3%	L: 9	0.3%	L: 27	1.0%	L: 58	2.2%	L: 0	0.0%	L: 4	0.2%	L: 91	3.5%
M: 5	0.2%	M: 4	0.2%	M: 7	0.3%	M: 19	0.7%	M: 0	0.0%	M: 33	1.2%	M: 31	1.2%
N: 140	5.3%	N: 1,075	40.9%	N: 228	8.5%	N: 10	0.4%	N: 2,502	94.5%	N: 0	0.0%	N: 0	0.0%
P: 1	0.0%	P: 1	0.0%	P: 0	0.0%	P: 0	0.0%	P: 0	0.0%	P: 459	17.6%	P: 5	0.2%
Q: 780	29.8%	Q: 251	9.0%	Q: 0	0.0%	Q: 394	14.9%	Q: 0	0.0%	Q: 27	1.0%	Q: 1,295	52.5%
R: 71	2.7%	R: 21	0.8%	R: 122	4.6%	R: 110	4.2%	R: 0	0.0%	R: 0	0.0%	R: 2	0.1%
S: 171	6.5%	S: 495	18.8%	S: 903	34.1%	S: 31	1.2%	S: 36	1.4%	S: 17	0.6%	S: 1,322	50.8%
T: 440	16.8%	T: 96	3.3%	T: 756	28.5%	T: 522	20.0%	T: 1	0.0%	T: 95	3.6%	T: 2	0.1%
V: 314	12.0%	V: 13	0.5%	V: 3	0.1%	V: 60	2.3%	V: 2	0.1%	V: 90	3.7%	V: 868	33.1%
W: 0	0.0%	W: 0	0.0%	W: 0	0.0%	W: 1	0.0%	W: 0	0.0%	W: 0	0.0%	W: 0	0.0%
Y: 4	0.2%	Y: 7	0.3%	Y: 0	0.0%	Y: 5	0.2%	Y: 0	0.0%	Y: 1,479	55.8%	Y: 0	0.0%
X: 0	0.0%	X: 1	0.0%	X: 1	0.0%	X: 0	0.0%	X: 0	0.0%	X: 0	0.0%	X: 0	0.0%

Table 2-12. Results of homologous protein sequence analysis. The six sites that benefited from mutation across our two-generation study are shown in green, with all other library sites in blue.

Of the 34 sites that were subjected to saturation mutagenesis (646 possible single mutations) in TEVp, 17 (2.6%) were deemed beneficial (Table 2-5), which is in the middle of previously published frequencies: 0.1-14.5%<sup>111-114</sup>. This beneficial mutation rate, as well as our success in ultimately engineering improved enzymatic activity in TEVp, further bolsters the literature findings that active site mutations can provide a clear benefit<sup>30,89</sup>. However, there are likely as many failures as reported successes with this approach<sup>115</sup>, and distal mutations have also consistently demonstrated the ability to

improve enzymatic activity or stability<sup>48,85-88</sup>, so a balanced approach that incorporates different regions for mutagenesis is still useful, if not recommended.

When the beneficial mutations from the first-generation libraries were pooled to create a single larger and more-constrained second-generation library, these mutations did not necessarily remain beneficial. A graphical analysis of enrichment in both the first- and second-generation TEVp libraries illustrates that roughly half of the residues that were positively enriched in the first-generation were negatively enriched in the second-generation TEVp library (Figure 2-30). For example, L204I was strongly enriched in the first-generation library, but was extremely disfavored in the second-generation library.

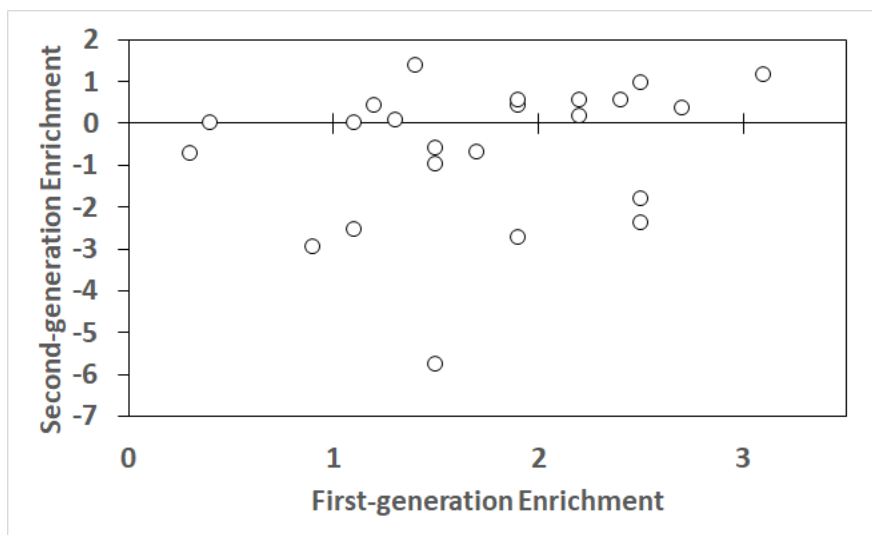


Figure 2-30. **Relationship between first-generation enrichment and second-generation enrichment.** Two groups of amino acids were omitted: 1)) wild-type residues with negative first-generation enrichment (S31, L204) and 2) amino acids included in the second-generation due to codon degeneracy (T30A).

Utilizing the sitewise enrichment data from our second-generation library, a handful of TEVp multi-mutants were produced in *E. coli* and subsequently analyzed to determine the impact on catalytic activity when these beneficial individual mutations were merged in different combinations. Using our strategy, the vast majority of TEVp multi-mutants produced represented an improvement over wild-type in terms of catalytic

efficiency (Figure 2-17a-b). However, this improvement was generally, though not always, manifested principally via a reduction in  $K_M$  as opposed to an improved  $k_{cat}$  (Figures 2-17c and 2-17d). There were no instances where both parameters were improved in a single multi-mutant. Rather, there usually was a tradeoff in slightly diminished  $k_{cat}$  with significantly lower  $K_M$ , although some mutants (TEVp3 and TEVp11, specifically) preserved  $K_M$  while improving turnover number. It is unclear whether this display framework is technologically more suited towards finding mutations that improve  $K_M$  instead of  $k_{cat}$  in general, or if our findings are more specific to the enzyme used and sites identified for mutation. Further, by comparing multi-mutants that differed only at one site, we were able to examine the effect of specific mutations in various contexts to examine if there was a particularly reliable effect (Figure 2-21). This analysis demonstrated that, in general, a consistent narrative emerged for each site (e.g., K215R generally leads to a relatively equivalent drop in both turnover number and  $K_M$  as compared to K215E); this insight can be used for future enzyme engineering efforts if particular functionality (binding vs. catalysis) needs to be emphasized. Finally, these improvements in catalytic efficiency could be used as the starting point for another round of directed evolution, where areas adjacent the recently-discovered hot spots (e.g., sites 30-32) are targeted for mutagenesis.

We considered the potential for a yeast cell's displayed substrate being cleaved by a TEVp variant on a neighboring cell, as opposed to the TEVp variant that is tethered to the same Aga2p as the substrate. The sampleable space of a tethered TEVp variant was modeled as a hemisphere (Figure 2-31), and both the substrate and TEVp variant were assumed restricted to a radius of nearly 15 nm from the yeast cell surface based on the

(G<sub>4</sub>S)<sub>2</sub>-ASASPAAPAPASPAAPAPSA-(G<sub>4</sub>S)<sub>2</sub> linker that connects both TEVp and its substrate to Aga2p. This suggests that a given TEVp variant occupies a volume of approximately 10<sup>-17</sup> mL around Aga2p, resulting in an effective localized concentration of roughly 250 μM for each tethered TEVp variant. The overall concentration of TEVp in solution was calculated based on the induction OD<sub>600</sub> of 1.0, which corresponds to 10<sup>7</sup> cells/mL. Using a high estimate of 10<sup>5</sup> displayed constructs per yeast cell<sup>97</sup>, the molar concentration of all TEVp in solution would be roughly 1 nM. Thus, the effective concentration of the tethered TEVp is over 10<sup>5</sup>-fold more than that of any neighboring TEVp variant, theoretically making cross-sampling a negligible concern and suggesting that cells exhibiting a decreased HA/c-Myc ratio are the result of increased catalytic efficiency of the encoded TEVp variant of that particular cell. Further, our ability to leverage this construct and our screening assays to enrich TEVp variants with demonstrably enhanced catalytic efficiency empirically disproves any major concerns about cross-sampling.

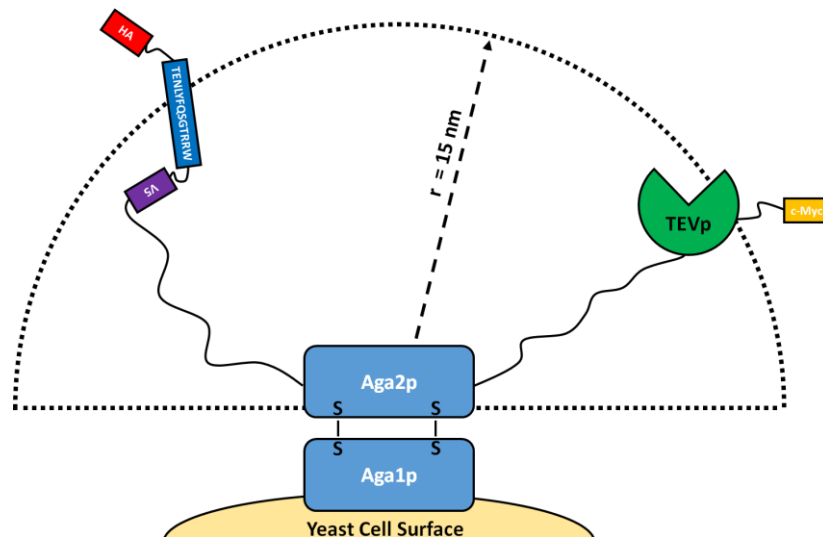


Figure 2-31. **Schematic of yeast surface display construct.** Theoretical half-sphere volume is denoted by dotted line. Radius of 15 nm is determined by length of linker.

Our display framework's ability to distinguish based on proteolytic activity was shown to function for two proteases, as both TEVp (Figure 2-7) and SrtA (Figure 2-25) effectively separated wild-type and inactive protease/substrate pairings on FACS plots. For both enzymes, 4-hr induction of display provided adequate signal differential for a collection gate for improved mutants in the low HA/high c-Myc area of the FACS plot. Notably, this region was less well-defined for SrtA as the displayed cells were more diffusely distributed on FACS plots compared to TEVp. This suggests that while our display construct appears generalizable to other proteases, the optimal experimental conditions and effectiveness may be protein-dependent. Also, there is a tradeoff with additional induction time, as increased display increases epitope signal, which creates more opportunity for activity differentiation but at the cost of more exposure time of surface-bound substrates to enzymatic cleavage, which reduces the fraction of intact substrate tethered with any active protease and thus reduces the space for a collection gate. An induction time of 4 hrs yielded an effective compromise for the proteases we tested, but the overall delicate position potentially motivates future alteration and optimization of the display construct to widen the dynamic range, specifically when using proteases with higher catalytic activity. These screens could be made more stringent by lengthening linkers or altering their composition to reduce substrate engagement by enzyme. Alternatively, a reversible enzyme inhibitor could also be utilized during induction to allow for increased display without enzymatic cutting, followed by a window of time at the experimenter's choosing for proteolysis after washing the inhibitor out. Nevertheless, the current experimental design enabled effective differentiation of enzyme activity.

Our novel yeast dual-display construct effectively differentiates cells quantitatively via FACS based on the catalytic activity of encoded proteases. Rationally-designed TEVp libraries were screened for mutants with enhanced catalytic efficiency in a directed evolution scheme, and the success of this approach was not noticeably hindered by deliberate under-sampling. Due to the targeted nature of our libraries, this approach likely had a higher probability of finding beneficial mutations than an error-prone PCR approach given the immensity and sparsity of sequence space. Further, other approaches such as multiple sequence alignment and consensus design, if the relevant sequence data existed, would possibly miss some of the hot spots our structure-guided libraries uncovered. Our approach represents a novel and more comprehensive way to find more active protease mutants. Additionally, while this display construct was mostly employed to study TEVp in this study, its success with SrtA demonstrates a potential tolerability for the use of other enzymes, which can be harnessed during efforts to improve other proteases. Finally, while this new comprehensive methodology does represent another addition to the protein engineering toolbox, there is specific and immediate benefit from multiple enhanced TEVp variants. Any biotechnological use of TEVp, such as its original intent as a specific tag remover, will be aided by an approximate 3-fold improvement in catalytic efficiency.

## **Chapter 3 – Applying ligand-enzyme fusion proteins to synthetic urinary biomarkers for noninvasive detection of abnormal disease-related receptor expression**

---

### **3.1 Abstract**

Blood-based biomarkers indicative of diseases like cancer are an appealing means for disease detection, but issues such as degradation and low concentration limit efficacy. To circumvent this, a novel concept was introduced: synthetic biomarkers. In brief, nanoparticles, which are connected to reporters by a substrate linkage, are first administered intravenously. When these nanoparticles survey diseased tissue, reporters are generated in a manner that reflects the concentration of a disease-specific enzyme. These reporters can be filtered into the urine for noninvasive disease detection. This approach is, however, limited by its reliance on upregulation of disease-specific proteases. Many diseases are characterized by abnormal expression of cell-surface receptors, but current diagnostic techniques are plagued by invasiveness, unreliability, and cost. In this chapter, we introduce a novel extension of the synthetic reporter approach to noninvasively detect abnormal receptor expression by harnessing ligand-enzyme fusion proteins, which impart exogenous enzymatic activity to tissue with aberrant receptor expression. A mathematical model for epidermal growth factor receptor (EGFR) tumor xenografts in mice demonstrated feasibility of this approach with tobacco etch virus protease (TEVp)-based fusions, suggesting detection of tumors as small as 0.28 mm when using typical substrate concentrations for synthetic reporter *in vivo* studies. Using a flexible linker, a variety of fusions were produced using different enzymes, ligands, and orientations. In each fusion combination tested, binding and catalytic activity



was well preserved, indicating a modular fusion framework. Demonstrating feasibility with anti-EGFR TEVp-based fusions in an *in vitro* cellular assay was not consistently successful. However, the following limitations were identified for improvement: high substrate lability, and insufficient fusion-specific product generation due to inadequate catalytic activity – which would motivate protease engineering – or suboptimal fusion linker design that resulted in ineffective projection of receptor-bound fusion’s enzyme component to engage soluble substrate.

### **3.2 Introduction**

The global incidence, prevalence and mortality rates of noncommunicable diseases such as cancer have been on the rise for decades, and the overall burden is increasingly and disproportionately being borne by the developing world, with cancer the second leading cause of mortality in developing countries for over a decade<sup>68,116</sup>. Early detection of cancer is linked with improved patient outcomes<sup>117</sup>. In the United States, colorectal cancer mortality rates have substantially declined mainly due to advancements in colonoscopy that have led to earlier diagnosis of colorectal polyps<sup>116</sup>. Similarly, mammography screening has led to a significant reduction in breast cancer-specific mortality in the industrialized world<sup>118</sup>, and evidence suggests that lung cancer screening via low-dose computed tomography has led to reduced mortality from lung cancer, the largest individual cause of death from cancer worldwide<sup>119</sup>. Nevertheless, a substantial percentage of cancer patients in the developed world are first identified with advanced stage disease<sup>116</sup>, and many require an invasive biopsy that, due to intratumor heterogeneity and sampling variation (Figure 3-1), can incorrectly diagnose or stage the cancer, and often requires a repeat biopsy<sup>69</sup>. Overall, this illustrates the need to improve

early diagnosis and screening methods to catch cancerous lesions when therapy can be curative, more cost-effective, and result in lower morbidity<sup>116</sup>. Furthermore, imaging modalities, colonoscopy technology, histopathology laboratories, and other essential medical infrastructure are often cost-prohibitive for remote or resource-limited settings<sup>68</sup>. Also factoring in the lack of requisite trained medical personnel, it is abundantly evident that early diagnosis and screening in developing countries is quite difficult. Thus, there is a pressing need for improved cancer screening tools and diagnostics for global health applications that are cheap, noninvasive, quantitative, highly sensitive and specific, image-free, user-friendly and point-of-care. This technology could also be revolutionary in the United States and the rest of the industrialized world, as there is a compelling case for a first-line imaging alternative that, while it sacrifices information on tumor location, provides for dramatic improvements in convenience and cost<sup>68</sup>. Using inexpensive point-of-care assays or devices to detect naturally-occurring molecular biomarkers from circulation is an appealing solution to the aforementioned dilemma. However, very few endogenous blood-based cancer biomarkers can reliably identify disease in an early state, either because of low concentration<sup>65</sup>, rapid degradation<sup>66</sup>, or an inability to detect the marker in such a complex biological fluid<sup>67</sup>. The dilution effect biomarkers experience upon entrance into the bloodstream from tumor cells is particularly dramatic; computational calculations indicate a solid tumor will not release sufficient biomarker amounts for disease detection until it has reached a spherical diameter of greater than 2.5 cm, which would take more than 10 years to develop<sup>120</sup>.

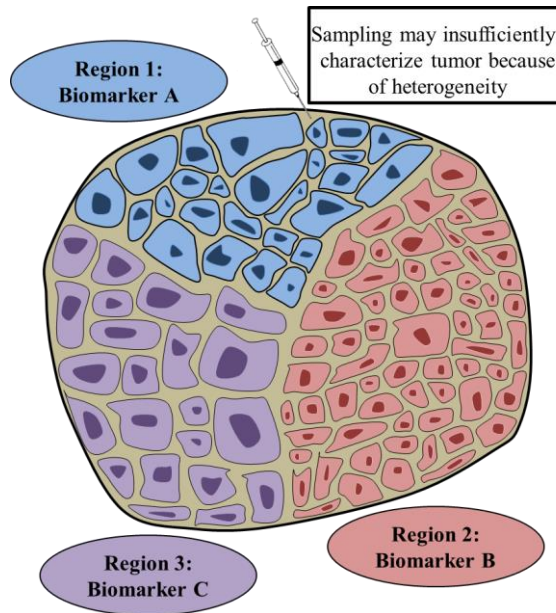


Figure 3-1. **Biopsy of a heterogeneous tumor.** Intratumor heterogeneity leads to biopsy sampling variability, which can result in insufficient characterization of the total molecular profile of the tumor, necessitating a repeat biopsy.

The concept of synthetic biomarkers was introduced to circumvent these particular issues, and has shown promise in noninvasively detecting noncommunicable diseases such as thrombosis, liver fibrosis, and colorectal cancer<sup>68-70</sup>. Instead of relying on low-level and unstable endogenous blood-based biomarkers, patients receive an intravenous dose of an engineered nanoscale agent that can interrogate potentially diseased tissue and remotely report on the existence of natural molecules or molecular processes by generating reporter molecules in circulation that can then be filtered into the urine for disease detection (Figure 3-2). The nanoscale agent consists of reporter peptides conjugated to a biocompatible nanoparticle via a protease-sensitive cleavage site. The nanoparticle bestows the conjugated peptides with long circulation times to facilitate prolonged surveying of the diseased tissue and also restricts the conjugated peptides from being filtered into the urine<sup>69</sup>. Additionally, in the case of a variety of cancers, transport of the nanoscale entity

to the tumor is greatly aided by angiogenic tumor vasculature and the enhanced permeability and retention effect<sup>121</sup>. Tumor proteases cut the susceptible peptide substrate, releasing reporter peptides into the circulation<sup>68-70</sup> to be selectively filtered and concentrated into the urine<sup>70</sup>. Localized tumor-specific upregulation of matrix metalloproteinases found in many solid cancers has previously been exploited by tailoring cleavage sites to a specific matrix metalloproteinase<sup>68-70</sup>. Many reporter types in unmodified urine have been investigated, including mass-encoded reporters analyzed via mass spectrometry<sup>68</sup> and ligand-encoded reporters detected via enzyme-linked immunosorbent assay (ELISA) or paper test strips<sup>68,70</sup> based on the lateral flow assay technology used in home pregnancy tests and in diagnostic tests for liver damage<sup>122</sup> and HIV<sup>123</sup>. Further, the use of mass spectrometry or various specific capture antibodies on a paper test strip allows for multiplexing and simultaneous quantification<sup>68</sup>. Unfortunately, while the injection of nanoparticles and the subsequent collection and analysis of a readily accessible and compositionally simple bodily fluid like urine is amenable to point-of-care disease detection, a mass spectrometer will likely not be available for global health cases; however, the aforementioned sandwich immunoassays would be an acceptable accompanying point-of-care assay for both the industrialized world and resource-limited settings<sup>68</sup>.

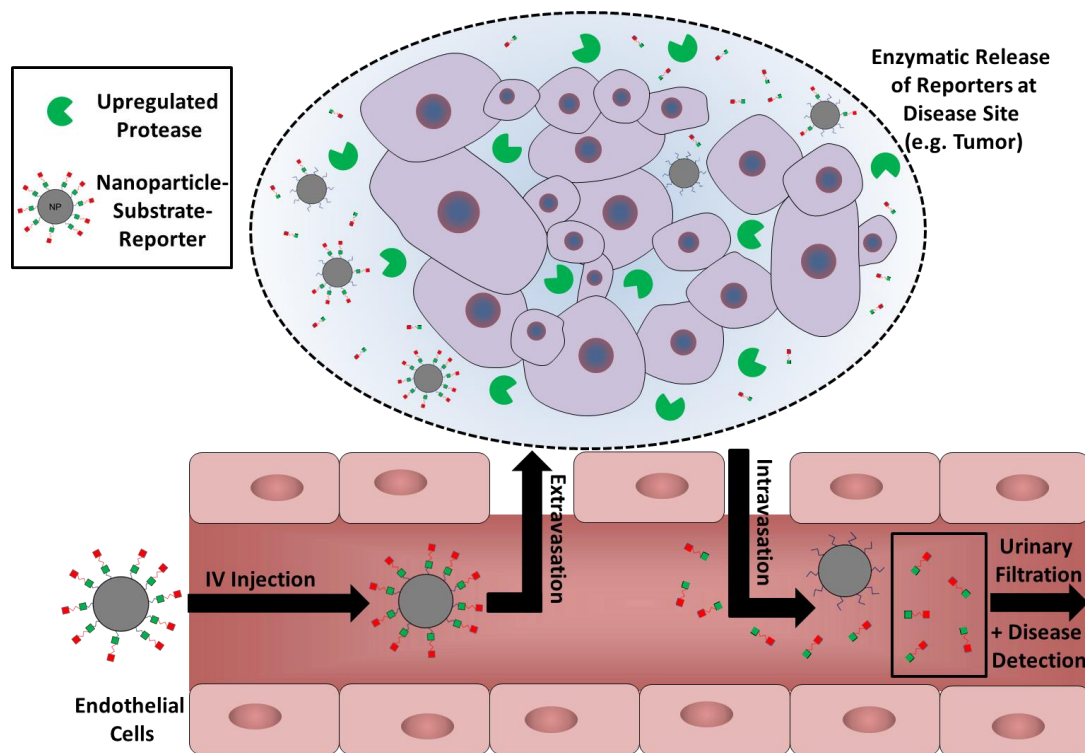


Figure 3-2. **Illustration of how synthetic reporters can non-invasively detect disease.** Nanoparticle-substrate-reporter composites are administered intravenously, circulate in the blood, and sample the disease environment to probe for aberrant enzymatic activity. Reporter peptides are conjugated to nanoparticles via linkages sensitive to disease-specific upregulated proteases, facilitating release of reporters that can then be renally filtered to enable non-invasive disease detection using a patient's urine.

Yet, this technology is dependent on a local dysregulated enzymatic profile. It would be helpful to diagnose and characterize noncommunicable diseases like cancer based on abnormal expression of cell-surface markers such as growth factor receptors. Growth factor receptors, including vascular endothelial growth factor receptor (VEGFR), insulin-like growth factor 1 receptor (IGF1R), mesenchymal epithelial transition (MET) receptor, and those found in the ErbB family of receptor tyrosine kinases, are often overexpressed in a multitude of cancers<sup>124</sup>. The ErbB family, also referred to as the epidermal growth factor receptor (EGFR) family, consists of four members: EGFR/ErbB1/HER1, ErbB2/HER2, ErbB3/HER3, and ErbB4/HER4<sup>125</sup>. Each ErbB receptor has three regions: an extracellular ligand-binding region, a transmembrane domain, and an intracellular tyrosine kinase region<sup>126-127</sup>. Growth factor binding induces homo- and hetero-dimerization

of the ErbB receptors, leading to autophosphorylation of several tyrosine residues on the intracellular domain that can then associate with a multitude of proteins possessing phosphotyrosine-binding SH2 domains<sup>128</sup>. This results in activation of various downstream signaling pathways responsible for pleiotropic biological events, including cell proliferation, differentiation, survival, adhesion and migration<sup>129-130</sup> (Figure 3-3). Normally, ErbB signaling is tightly regulated, but releasing this control, and activating the receptors via overexpression or mutation<sup>127</sup>, has been linked to several cancers<sup>128</sup>. Overexpression or activating mutations of EGFR have been noted in non-small cell lung cancer<sup>131-132</sup>, glioblastoma<sup>133</sup>, breast carcinoma<sup>134</sup>, as well as esophageal<sup>135</sup> and head and neck<sup>136-137</sup> squamous cell carcinoma, among others. Similarly, overexpression of HER2 has been observed in gastric and gastroesophageal junction adenocarcinoma<sup>138</sup>, as well as breast and ovarian cancer<sup>139-140</sup>. Overexpression or activating mutations of these ErbB receptors confers unique survival advantages to malignant cells by enhancing cellular proliferation, enabling metastatic spread and suppressing apoptosis, which together creates a more aggressive form of cancer resistant to standard therapies, decreases disease-free and overall survival rates and leads to an increased risk of disease recurrence<sup>141-142</sup>. Consequently, the ErbB family of receptors is an appealing biological target for cancer therapy. Several therapeutics, either antibodies against the extracellular domain such as trastuzumab (HER2) or small-molecule tyrosine kinase inhibitors such as gefitinib (EGFR), have shown promise in blocking the ErbB receptor's pathway, retarding tumor growth and improving prognosis<sup>127</sup>. Thus, characterizing a patient's ErbB receptor expression levels and mutational status is not only crucial for initial diagnosis and

monitoring of disease recurrence, but also for patient stratification, treatment guidance, and monitoring of treatment response over time.

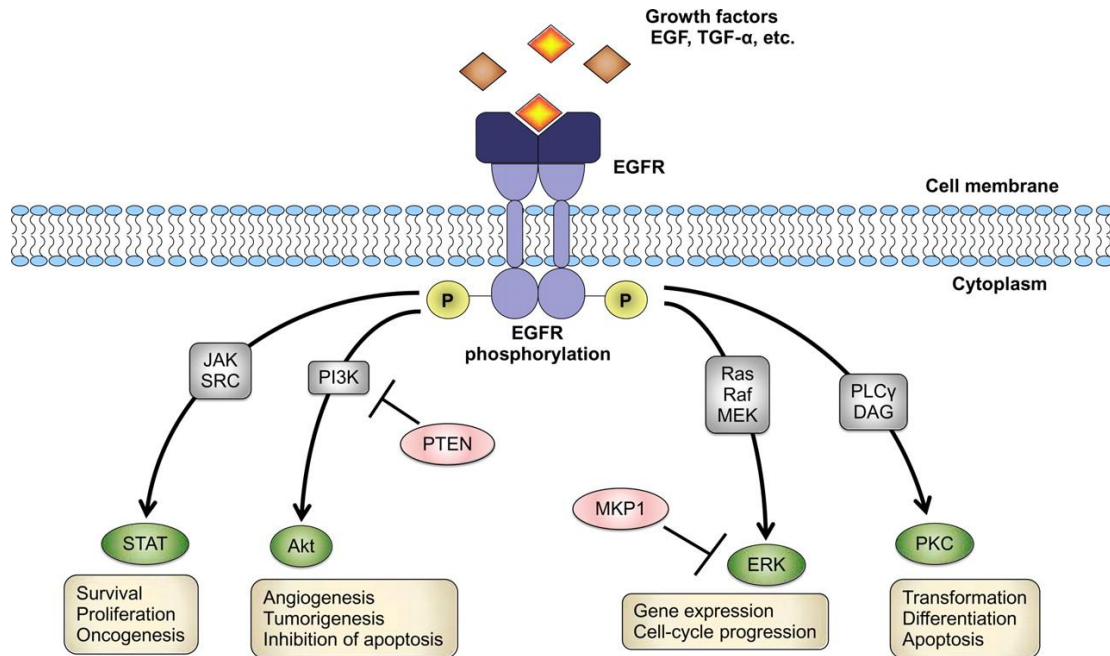


Figure 3-3. **Signaling pathways of EGFR.** EGF, transforming growth factor alpha (TGF- $\alpha$ ), and other ligands can activate the EGFR signaling pathway. Some of the important signaling pathways regulated by EGFR are shown, with other pathways and cross-talk left out for clarity. Figure reproduced from [137].

The current gold standard techniques for identifying abnormal ErbB expression are immunohistochemistry at the protein level and fluorescence in situ hybridization (FISH) at the gene level<sup>143-147</sup>, yet both of these approaches possess significant limitations. Both techniques require a tissue sample from either an excisional biopsy or core needle biopsy<sup>148</sup>, with multiple biopsies sometimes necessary to account for tissue heterogeneity<sup>143-144</sup>. These procedures are inherently invasive and thus carry particular risks, some of them quite significant; for instance, when performing a transthoracic needle biopsy of a pulmonary nodule, there is a 15% risk of pneumothorax, which could necessitate a chest tube, a longer length of stay, and even lead to respiratory failure requiring mechanical ventilation<sup>149</sup>. Additionally, there are major concerns regarding the

reproducibility and reliability of both immunohistochemistry and FISH. There is considerable intra- and inter-laboratory inconsistency in results due to differing methodology, instrumentation and experience in tissue fixation, antigen retrieval and staining<sup>143-147</sup>; further, in using immunohistochemistry or FISH to determine ErbB status, there is even significant discordance between samples taken via different biopsy procedures<sup>148</sup>. In addition, staining interpretation can fluctuate dramatically between pathologists, and the threshold setting for reporting positive results is varying; a large number of false negatives may be observed if the threshold is too high, and many patients must then undergo a repeat biopsy if dictated by their clinical characteristics<sup>143-145</sup>. Finally, resource-limited settings typically do not possess the expensive histopathology facilities or skilled personnel needed to run such procedures and assays, and thus this form of detection is ill-equipped to impact the global health cancer crisis<sup>68</sup>. These shortcomings underscore the need for inexpensive, noninvasive, sensitive and specific point-of-care detection technologies to quantitatively characterize ErbB receptor status, making EGFR an ideal model target for our initial studies.

Here we propose an innovation that potentially, and significantly, broadens applicability of the synthetic biomarker approach (Figure 3-4). Pairing excessive receptor proteins with an exogenous enzyme, accomplished by administering a dose of a fusion protein that combines exogenous enzymatic activity with binding capabilities (in this case anti-EGFR binding), provides an artificially increased enzyme concentration in the local disease environment. If sufficiently small, unbound fusions will be rapidly cleared by the renal system *in vivo*, leaving only those fusions bound to EGFR. Administration of a substrate, such as reporter peptides conjugated to the nanoparticle surface by the exogenous



enzyme's substrate<sup>68</sup>, will lead to unique product release into circulation upon substrate sampling of the disease site, and then urinary analysis of these filtered products can reveal underlying receptor expression levels.

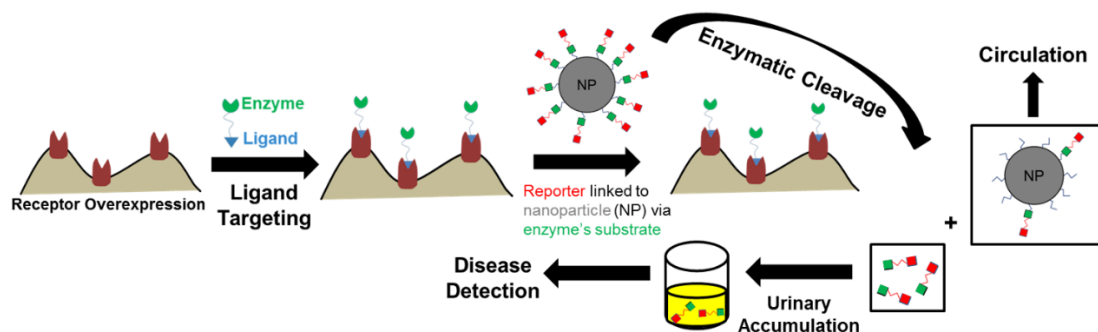


Figure 3-4. **Synthetic reporters to detect or quantify receptor expression.** The proposed system involves the binding of ligand-enzyme fusions, followed by delivery of reporters linked to nanoparticles via a cleavable linkage, and subsequent enzymatic release of reporters that are filtered into urine for analysis.

The proposed synthetic reporter strategy requires ligands and enzymes that meet rigorous criteria. For the ligand, small molecular size is desired to facilitate rapid urinary clearance of unbound molecules – to ensure low background – and enhanced extravasation and solid tumor penetration – to maximize delivery to tumor and drive sensitivity<sup>98,150-151</sup>. Antibodies perform poorly in this category due to restrictions imposed by large size (immunoglobulin G is 150 kDa, for example)<sup>150</sup>. High specificity and affinity, preferably in the low-nM dissociation constant ( $K_D$ ) region, are needed for effective ligand delivery and reduced clinical side effects<sup>151</sup>. The Hackel Lab has engineered multiple small-molecule protein scaffolds for low-nM affinity EGFR binding, including affibody (7 kDa)<sup>150</sup>, gene 2 protein (Gp2, 6 kDa)<sup>151</sup>, and fibronectin (Fn, 11 kDa)<sup>98</sup>. The tenth type III domain of human Fn has proven effective as a scaffold for molecular recognition, with three solvent-exposed loops (denoted BC, DE, and FG) that resemble an immunoglobulin complementarity-determining region and are amenable to sequence modification to create

novel binding capabilities<sup>152</sup> (Figure 3-5). Multiple anti-EGFR Fn binders have been engineered to bind various epitopes on the receptor (Table 3-1). Fn clone C (FnC) possesses a low-nM  $K_D$ , whereas FnA and FnD has a sub-nM  $K_D$ , making these variants ideal ligands for this proposed technology. Additionally, a non-binding Fn clone (FnNB), which replaces the wild-type RGD (arginine, glycine, aspartate) binding motif with an RDG scrambled sequence to abolish binding<sup>153</sup>, can serve as an experimental negative control. Additionally, many of these anti-EGFR Fn mutants have already been successfully used in a fusion protein context with a flexible GGGSGGGKGGG linker<sup>98</sup>. Other anti-EGFR clones of interest include the affibody clone EA68<sup>150</sup> and the 45-amino acid Gp2<sub>E2.2.3</sub><sup>151</sup> (hereafter referenced as simply Gp2), which exhibit binding affinities to A431 cells of  $5.3 \pm 1.7$  nM and  $18 \pm 8$  nM, respectively.



Figure 3-5. **The wild-type tenth type III domain of fibronectin (PDB: 1TTF).** The three loops engineered for EGFR binding are shown: DE (red), BC (purple), FG (blue).

Name	BC loop	DE loop	FG loop	Epitope <sup>a</sup>	K <sub>D</sub> (nM) <sup>b</sup>
A	FDYAVTY	GWIST	DNSHWPFIRST	L14H, Q16R, Y45F, H69(Q,R,Y)	0.26 ± 0.13
C	YFRDPRYVDY	WYLPE	GDDQNAGL	I341V, E376K	1.4 ± 0.2
D	LHHRSDVRS	GSRSL	WGSYCCSN	K430E, S506R	0.25 ± 0.05
NB	DAPAVTVRY	GSKST	GRDGSPASSK	--	--

Table 3-1. **The sequences, epitopes and affinities of EGFR-binding Fn mutants.** <sup>a</sup>Epitope mapping identified EGFR mutations that reduced affinity for the clone of interest<sup>98</sup>. <sup>b</sup>The affinity per clone was determined by titration of Fn binding to A431 cells at pH 7.4 (on ice to prevent internalization).

For the enzyme, four major criteria were considered: a) small size to enhance tissue permeability and urinary clearance, as with the binder; b) high specificity to minimize any off-target promiscuous enzymatic activity; c) high substrate stability to maintain low background product formation from endogenous enzymes; and d) high catalytic activity (high  $k_{cat}$ ) and substrate affinity (low  $K_M$ ) at physiological conditions (pH, temperature, ionic strength, oxidation state) to ensure fusions will generate sufficient reporter in the urine for sensitive detection. Two enzymes were selected: the aforementioned TEVp1 (28 kDa) and Srt7 (17 kDa, referenced in Chapter 2 as a heptamutant version of sortase A with five mutations for enhanced activity – providing an 11-fold improvement in catalytic efficiency – and two more for calcium-independence<sup>28,104-106</sup>), which are both remarkably substrate-stringent proteases with respective literature  $k_{cat}$  values of  $0.30 \pm 0.02$  and  $1.85 \pm 0.1 \text{ s}^{-1}$ , and  $K_M$  values of  $65 \pm 8$  and  $320 \pm 50 \mu\text{M}$ <sup>26,106</sup>. Further, both proteases possess an optimal temperature of approximately 30 °C, with only a minor dropoff (~20-25%) in activity from this maximum at 37 °C, though TEVp possesses only ~5-10% of this maximal activity at 4 °C<sup>154-155</sup>.

Here we evaluate the performance of multiple elements of the strategy for synthetic reporters via ligand-enzyme fusion activity. We demonstrate the ability to produce various recombinant fusion proteins in *E. coli*, including those with TEVp1 and Srt7 as the enzyme

component, and FnA, FnC, FnD, FnNB, EA68 and Gp2 as the ligand component. In many instances, both the ligand-enzyme and enzyme-ligand orientation, were tested. The vast majority of these fusions demonstrated significant binding activity, which in some cases was nearly identical as their parental individual ligand component. Two fusions, FnD-TEVp1 and FnNB-TEVp1, were also investigated for enzymatic activity and were shown to exhibit similar values for the turnover number, Michaelis-Menten constant and catalytic efficiency as compared to TEVp1. These fusions were then evaluated on EGFR-expressing A431 cells to assess the ability of cell surface-bound fusion to generate a reporter fluorescence signal significantly above background (e.g., from non-binding fusions). Inconsistencies in the results from this cellular assay suggested that fusions must be improved, either in terms of enhancements in enzymatic activity or alterations of the linking entity connecting ligand and enzyme to better project enzyme from the cell surface to engage soluble substrate molecules.

### **3.3 Methods**

#### 3.3.1 Creation of fusion protein-encoding DNA constructs

Plasmid pET constructs containing various ligand-enzyme and enzyme-ligand fusion proteins were created via several steps. To start, a pET Fn-linker-Fn-His<sub>6</sub> (or simply Fn-Fn) construct was used that featured wild-type Fn proteins separated by a GGGSGGGKGGG linker; the N-terminal Fn was flanked by NheI and BamHI cut sites and the C-terminal Fn was flanked by KpnI and SacI cut sites. Using analogous protocols as described in Chapter 2, the N-terminal Fn in the pET Fn-Fn vector was removed by digesting the plasmid with NheI and BamHI restriction enzymes, followed by recovery of cut vector using gel electrophoresis. FnA (primers 1-2), FnC (3-4), FnD (5-6), FnNB (7-

8), EA68 (9-10), Gp2 (11-12), TEVp1 (13-14), and Srt7 (15-16) genes were all PCR amplified (Supplemental Table 3-1) to add flanking NheI and BamHI cut sites; NheI/BamHI restriction enzyme digestion of these PCR amplified genes, and subsequent recovery via gel electrophoresis, created a gene with the appropriate NheI and BamHI sticky ends to permit ligation of these eight entities into the cut vector to create FnA-Fn, FnC-Fn, FnD-Fn, FnNB-Fn, EA68-Fn, Gp2-Fn, TEVp1-Fn, and Srt7-Fn. For the first six of these genetic constructs (i.e., featuring N-terminal ligand), the C-terminal Fn was excised via a digestion using KpnI and SacI; TEVp1 or Srt7 genes, PCR amplified to add KpnI and SacI cut sites using primers 17-18 and 19-20, respectively, were then digested using KpnI and SacI enzymes to facilitate ligation with the cut vector to create six versions of ligand-TEVp1 and four versions ligand-Srt7 (FnC-Srt7 and EA68-Srt7 were not created). The C-terminal Fn was similarly removed from pET TEVp1-Fn and pET Srt7-Fn with an analogous KpnI/SacI restriction digest; PCR amplification of the genes for four ligands (FnA, FnD, FnNB, Gp2) was performed to add KpnI and SacI cut sites using primers 21-22, 23-24, 25-26, and 27-28, respectively; subsequent KpnI/SacI restriction digest facilitated ligation of the four ligand options into constructs possessing N-terminal TEVp1 or Srt7. pET vectors were thus created for 15 fusion proteins: FnA-TEVp1, FnA-Srt7, FnC-TEVp1, FnD-TEVp1, FnD-Srt7, FnNB-TEVp1, FnNB-Srt7, EA68-TEVp1, Gp2-TEVp1, Gp2-Srt7, TEVp1-FnA, Srt7-FnA, TEVp1-FnD, Srt7-FnD, TEVp1-FnNB, Srt7-FnNB, TEVp1-Gp2, and Srt7-Gp2.

### 3.3.2 Fusion protein production

T7 Express competent *E. coli* were transformed with one of the 15 fusion protein-encoding pET plasmids, grown in 150 mL lysogeny broth cultures at 37 °C, 250 rpm and

then induced with 0.5 mM IPTG at an OD<sub>600</sub> of 0.7-1.0 for 5 hrs at 30 °C, 250 rpm. Cell pelleting, lysis, processing, and analysis by SDS-PAGE and densitometry was performed as before in Chapter 2, except 12 µL instead of 6 µL of sample was added to each well on the gel. Productions were performed in quadruplicate.

### 3.3.3 Cell culturing

A431 cells were grown at 37 °C in a humidified atmosphere with 5% CO<sub>2</sub>, cultured in Dulbecco's modified Eagle medium with 4.5 g/L glucose, sodium pyruvate, and glutamine supplemented with 10% (v/v) fetal bovine serum. When cells reached ~80% confluence, 0.25% trypsin and 1 mM EDTA was used to detach cells for passage or analysis, and when appropriate cells were counted with a Countess II Automated Cell Counter.

### 3.3.4 Affinity titration of fusion proteins

50,000 cells were labeled with differing concentrations of each fusion protein, rotating for 6 hrs on ice (i.e., at 4 °C) to ensure >95% equilibrium binding was achieved. Cells were pelleted at 400g for 3 min, washed with cold PBSA, and incubated for 30 min on ice in 50 µL PBSA with 20 µg/mL anti-His<sub>6</sub> tag antibody conjugated to FITC (ab1206, Abcam); cells were washed in PBSA to remove unbound antibody conjugate and resuspended in PBSA for flow cytometry analysis on an Accuri C6 instrument. The sum of squared errors was minimized to obtain the value of K<sub>D</sub> assuming a 1:1 binding interaction. The number of replicates for specific fusions were: three (FnA-TEVp1, Gp2-Srt7, and Srt7-Gp2), five (FnC-TEVp1), and two (all other fusions).

### 3.3.5 Fusion protein catalytic analysis

The peptide substrate 2-Abz-ENLYFQSGTK-Dnp was again used to examine fusion protein catalytic parameters. The peptide contains the canonical TEVp1 substrate sequence (ENLYFQSG), where cleavage occurs between Q and S, and is flanked by the 2-aminobenzoyl fluorophore and its quencher 2,4-dinitrophenyl. Thus, intact peptide exhibits minimal fluorescence but cleaved peptide is highly fluorescent<sup>156-158</sup>. Fusions (FnD-TEVp1 or FnNB-TEVp1) was diluted to a concentration of 0.4  $\mu\text{M}$  using TEVp-free lysate of *E. coli*-produced FnD. 40  $\mu\text{L}$  of this solution was added to five microplate wells. 40  $\mu\text{L}$  solution containing the peptide substrate in PBS at various concentrations (0, 5, 20, 50, and 150  $\mu\text{M}$ ) was then added to the five wells to give a final fusion protein concentration of 0.2  $\mu\text{M}$  and final substrate concentrations of 0, 2.5, 10, 25, and 75  $\mu\text{M}$ . Fluorescence ( $\lambda_{\text{ex}} = 320 \text{ nm}$ ,  $\lambda_{\text{em}} = 420 \text{ nm}$ ) measurements were taken every 30 s for 150 s total on a Synergy H1 microplate reader. Background fluorescence from a TEVp-free negative control, consisting of FnD lysate, was subtracted, and slopes of background-subtracted fluorescence versus time were calculated at each substrate concentration. A 2-Abz standard curve allowed for the conversion of slopes to initial reaction velocities. These reaction velocities at various substrate concentrations were fit to a Michaelis-Menten model to calculate  $k_{\text{cat}}$  and  $K_{\text{M}}$ . Triplicate measurements were conducted.

### 3.3.6 *In vitro* cell assay

Four samples, each tested in triplicate, were identified for the *in vitro* cellular assay to validate the efficacy of the synthetic reporter approach applied to surface-bound fusion proteins generating signal correlative to cell-surface receptor density: fusion (FnD-TEVp1) with substrate, and three negative controls that either eliminated binding (i.e., using FnNB-TEVp1 instead of FnD-TEVp1), catalytic activity (i.e., using FnD instead of

FnD-TEVp1), or substrate (PBSA added instead). FnD-TEVp1, FnNB-TEVp1, or FnD ligand were first diluted in cold PBSA to yield a concentration of 60 nM. 100  $\mu$ L of FnD-TEVp1, FnNB-TEVp1 or FnD at this concentration was used to resuspend 150,000 pelleted A431 cells. This represented a ligand:receptor ratio of approximately 10:1 assuming two million EGFR molecules per A431 cell<sup>159-161</sup>. Samples were placed on ice on a rotator for 1 hr to approach equilibrium binding. Samples were pelleted at 400g for 3 min, washed twice with 1 mL cold PBSA, and resuspended in 50  $\mu$ L of 100  $\mu$ M peptide substrate. This volume of solution was added to 96-well plates, which were covered in aluminum foil and rotated in the cold room when not being read by the fluorimeter. Fluorescence measurements were made every hour for the first five hours followed by measurements at 11.5 hr, 17.5 hr, and 22.5 hr.

### **3.4 Results**

#### 3.4.1 Mathematical modeling of ligand-enzyme-reporter system

Our proposed extension of the synthetic reporter approach harnesses ligand-enzyme fusions to create an artificial exogenous enzymatic presence in the disease site that is reflective of the underlying receptor burden and can be exploited by substrate to generate reporters for noninvasive urinary disease detection. To first ensure feasibility of this approach, we modeled the intended *in vivo* application. Our mathematical model made several reasonable assumptions to predict the rate of reporter generation from an EGFR-bearing tumor in mice. The model assumes two million EGFR molecules per A431 cell, which has been experimentally measured in cell culture<sup>159-161</sup> and murine models<sup>162</sup>. The model also assumes a 20% tumor void fraction and that 2% of EGFR are labeled by ligand-enzyme fusion, which is consistent with published results for



molecularly targeted ligands in tumors<sup>163-164</sup>. Michaelis-Menten parameters for TEVp1 from literature were used<sup>26</sup>. A total mouse blood volume of 1.5 mL<sup>165</sup> was used to estimate the dilution factor once reporters move from the tumor environment to circulation. Finally, the blood:urine concentrating factor, estimated as approximately 70, was determined by dividing the total mouse blood volume of 1.5 mL by the mouse urine volume after 1 hour, equal to the daily urine volume of 0.5 mL<sup>166</sup> divided by 24. Considering the standard ELISA detection limit ( $\sim 0.1$  nM)<sup>167</sup>, our mathematical model predicts robust reporter generation after one hour for a variety of substrate concentrations (40 nM to 100  $\mu$ M) and a range of tumor burdens (up to 5 mm in diameter). For example, a tumor diameter of 5 mm and a substrate concentration of 1  $\mu$ M, which is in line with the concentrations used in previous *in vivo* synthetic reporter experiments<sup>68,70</sup>, results in fusion-bound TEVp generating 52 pM of reporter per second, and a urinary reporter concentration after one hour of approximately 600 nM, between three and four orders of magnitude greater than the limit of detection (Figure 3-6, Supplemental Figure 3-1). Additionally, the minimum tumor diameter required for detection predictably decreases with increased substrate concentration, from approximately 0.81 mm for 40 nM to 0.079 mm for 100  $\mu$ M (Figure 3-7). Holding the substrate concentration at 1  $\mu$ M, the minimum tumor diameter needed for detection decreases from 0.58 mm to 0.28 mm to 0.17 mm for  $k_{cat}$  values of 0.03 s<sup>-1</sup>, 0.3 s<sup>-1</sup>, and 3 s<sup>-1</sup>, respectively, and these same minimum tumor diameter values held for varying  $K_M$  from 650  $\mu$ M to 65  $\mu$ M to 6.5  $\mu$ M, respectively. Finally, operating at a substrate concentration of 1  $\mu$ M and predicting the effect of  $K_D$  on EGFR labeling from Schmidt and Wittrup<sup>164</sup>, the minimum tumor diameter needed for detection only modestly increases from 0.28 mm to 0.29 mm to 0.37 mm for  $K_D$  values of

1 nM, 10 nM, and 100 nM, although an additional 10-fold increase in  $K_D$  to 1  $\mu$ M results in a substantially larger minimum tumor diameter. The results of the *in vivo* mathematical model persuasively suggest that, in context with the literature, the system is indeed practical.

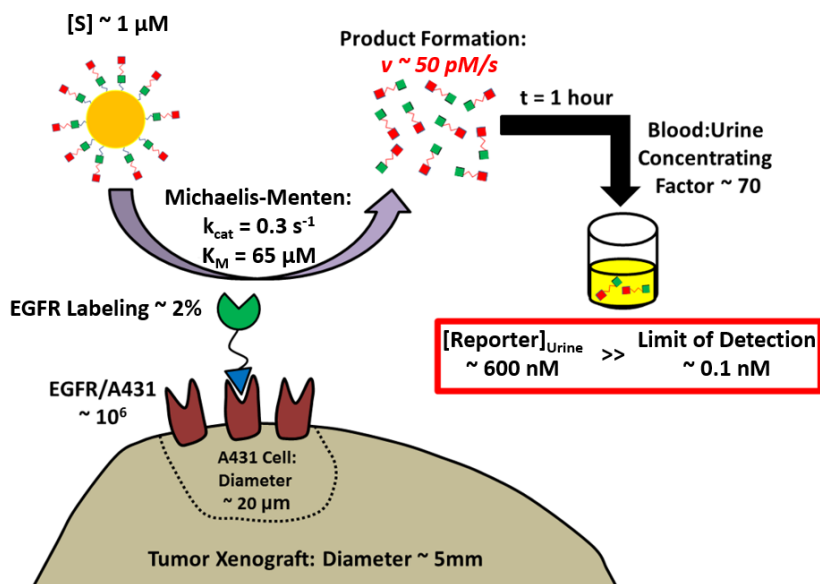


Figure 3-6. **Schematic illustrating feasibility of our extension of the *in vivo* synthetic reporter approach that leverages ligand-enzyme fusions to remotely detect aberrant receptor expression.** Our model evaluated an EGFR-bearing tumor, shown here with diameter of 5 mm, and assumed TEVp catalytic properties in accordance with literature<sup>26</sup>. For a substrate concentration of 1  $\mu$ M, the average product formation rate over one hour in the tumor environment was approximately 50 pM/s and the predicted urinary reporter peptide concentration after one hour was between three and four orders of magnitude above the ELISA limit of detection<sup>167</sup>.

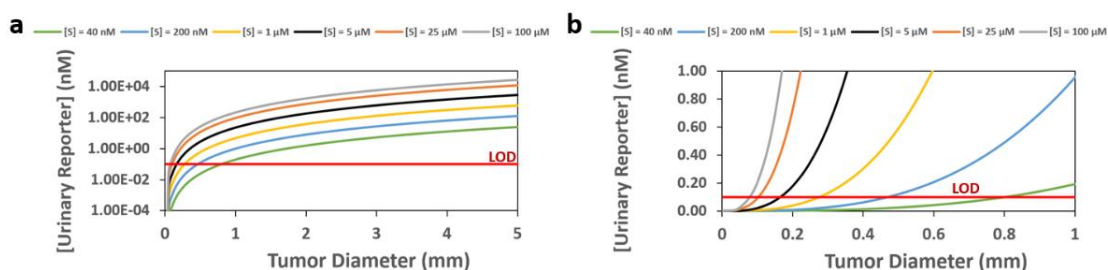


Figure 3-7. **Urinary reporter concentration after one hour as a function of EGFR-expressing tumor burden at various substrate concentrations.** A) Substrate concentrations of 0.04, 0.2, 1, 5, 25, and 100  $\mu$ M were tested for a tumor burden of 0-5 mm in diameter. Y-axis is presented on a log scale for better data visualization. ELISA limit of detection (LOD:  $\sim 0.1 \text{ nM}$ ) is shown with a horizontal red line. B) Equivalent data shown, but zoomed in on the 0-0.5 mm tumor diameter and 0-1 nM urinary reporter range in order to better show performance with respect to the LOD. Increased substrate concentration permits detection of smaller tumors, approximately 0.81, 0.47, 0.28, 0.16, 0.10 and 0.079 mm in diameter for substrate concentrations of 0.04, 0.2, 1, 5, 25 and 100  $\mu$ M, respectively.

Before progressing to any *in vivo* studies, however, we sought to demonstrate practicality in an *in vitro* cellular assay (Figure 3-8). This assay examines the performance of fusion proteins, compared to relevant negative controls such as ligand-only or non-binding fusions, in first binding to EGFR proteins on A431 cells on ice *in vitro*, and then subsequently generating product from the cell surface upon substrate exposure. Mathematical modeling of this assay, even at reduced temperature, suggested robust product generation from the A431 cell surface. Assuming two million EGFR molecules per A431 cell<sup>159-161</sup>, 95% binding of fusion proteins to EGFR, and adequate washing out of unbound fusions, 150,000 cells in 50  $\mu$ L should yield an approximate bound fusion, and thus bound TEVp1, concentration of 10 nM. Applying Michaelis-Menten kinetics to a substrate concentration of 100  $\mu$ M, and factoring in a roughly 10-fold reduction<sup>154</sup> in the TEVp turnover number at 4 °C compared to room temperature, roughly 1  $\mu$ M of unquenched Abz should be generated every hour. Using the standard curve for 2-Abz (Figure 2-4), this roughly correlates to an hourly increase in arbitrary fluorescence units of 10,000. Therefore, while this system incorporates a wide array of components, our models suggest robust signal generation in both the *in vitro* and *in vivo* settings to make this extension of the synthetic reporter technology feasible.

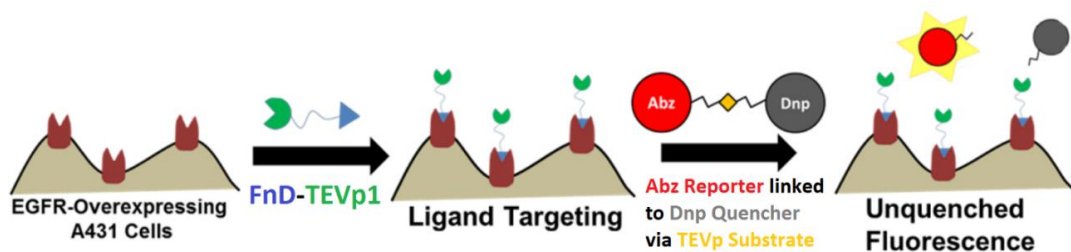


Figure 3-8. **Schematic illustrating the sequence of the synthetic reporter *in vitro* assay using ligand-enzyme fusion proteins.** Fusions (e.g., FnD-TEVp1) are first bound to A431 cells, and substrate peptide is added after unbound fusions are washed away. The increase in fluorescence signal should be correlative with the number of fusion proteins, and thus the number of accessible EGFR molecules.

### 3.4.2 Fusion protein production

Ligand-enzyme fusion proteins were expressed from pET vector in NEB5 $\alpha$  *E. coli* in 150 mL volume for 5 hrs at 30 °C. Cell lysate was analyzed by SDS-PAGE, along with carbonic anhydrase (CA) and bovine serum albumin (BSA) reference proteins, to quantify fusion protein concentration via densitometry (Figure 3-9). Noticeably, the Srt7-based fusions consistently generated greater yields than the TEVp1-based fusions, particularly when FnA or Gp2 is used as the ligand (Figure 3-10). There did not appear to be a consistent trend in yields when comparing different ligands aside from FnNB-containing fusions, which tended to offer higher fusion yields, and there was not a significant difference when comparing different orientations, irrespective of the ligand or enzyme under consideration. Regardless, for each fusion protein produced, yields were sufficient to permit affinity and catalytic analysis.

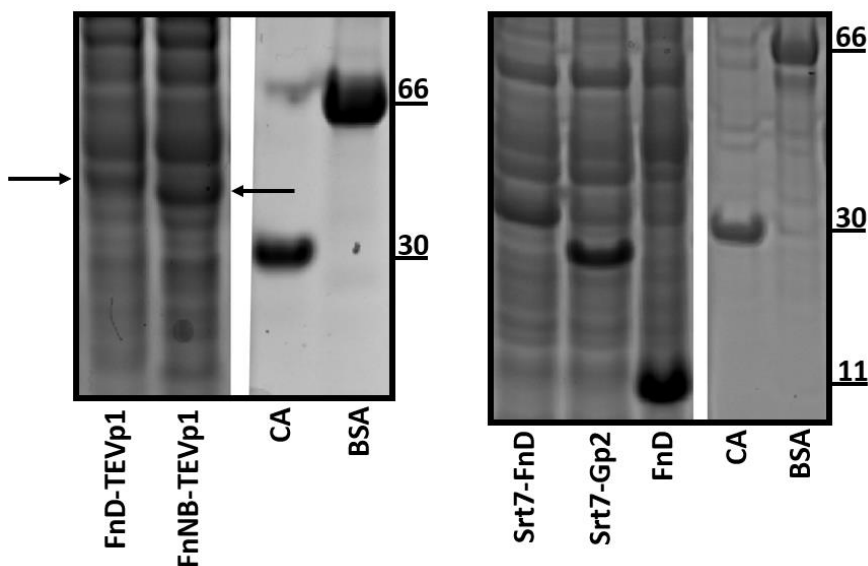


Figure 3-9. **SDS-PAGE analysis to quantify fusion production yield.** Example of SDS-PAGE gels of *E. coli* lysate from productions of four fusions [FnD-TEVp1 (40 kDa), FnNB-TEVp1 (39 kDa), Srt7-FnD (30 kDa), and Srt7-Gp2 (25 kDa)] along with references CA, BSA and FnD. Arrows are shown to illustrate the band corresponding to FnD-TEVp1 and FnNB-TEVp1.

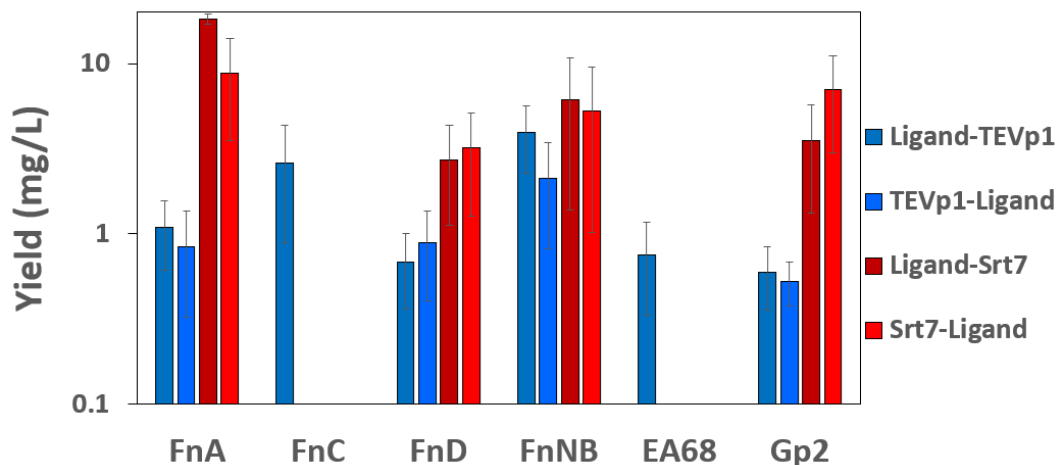


Figure 3-10. **Production yield of various fusions.** Ligand-enzyme fusions were expressed in NEB5 $\alpha$  *E. coli*, and yield was quantified by SDS-PAGE and densitometry with calibrant protein standards. For ligands FnA, FnD, FnNB and Gp2, fusions were made for both TEVp1 and Srt7, in both orientations for each. For FnC and EA68, only the ligand-TEVp1 version was made. Error bars shown are 95% confidence intervals.

### 3.4.3 Affinity titration of fusion proteins

Fusion proteins were titrated to assess binding affinity to EGFR-expressing A431 cells (Figure 3-11). It appears that only the FnA- and FnD-containing TEVp1 fusions exhibited a meaningful decrease in affinity, specifically one to two orders of magnitude compared to published  $K_D$  values for the corresponding ligand. Nevertheless, low nanomolar affinities are preserved, making FnA- and FnD-based fusions still suitable for further efforts in the synthetic reporter approach. FnC, EA68, and Gp2 binding was not substantially impacted by TEVp1-based fusions. All four Srt7-based fusions tested exhibited roughly unaffected  $K_D$  values. FnNB-based fusions were evaluated at the highest concentration tested (200 nM) and consistently demonstrated negligible binding, as expected. All tested fusions exhibited acceptable binding performance; FnD-TEVp1 was chosen for further evaluation because of high affinity and good yield in the context of TEVp, which was chosen for initial enzyme focus.

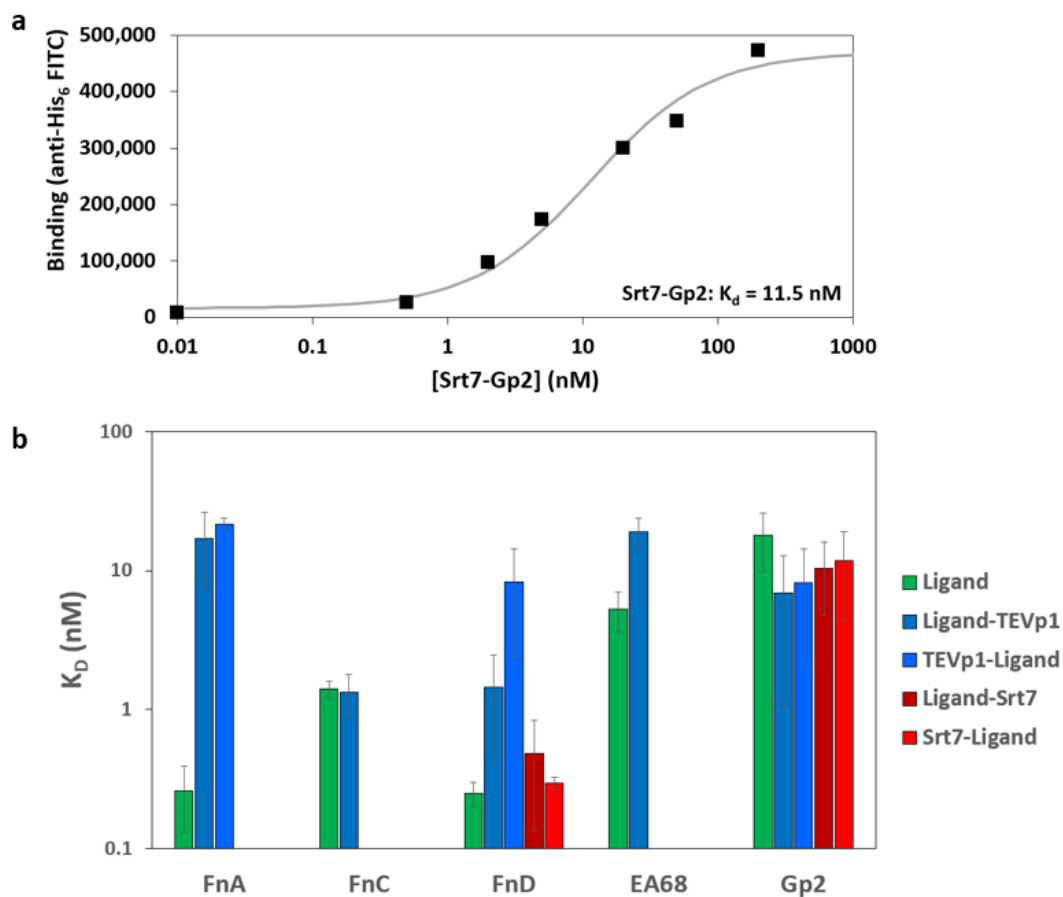


Figure 3-11. **Affinity estimations for various fusions compared to literature  $K_D$  values for ligand only.**  
 A) Example of Srt7-Gp2 dissociation constant ( $K_D$ ) being determined by titration, specifically using seven fusion concentrations (0.01, 0.5, 2, 5, 20, 50, and 200 nM). EGFR-overexpressing A431 cells were incubated in a particular fusion concentration while on ice, followed by labeling of EGFR-bound fusions with FITC-conjugated anti-His<sub>6</sub> antibody. Analysis by flow cytometry provided fluorescence that was correlative with bound fusion amount. B)  $K_D$  values shown for each fusion evaluated, including for various ligands, enzymes and orientation of components. Error bars shown are 95% confidence intervals.  $K_D$  values for ligand-only samples were obtained from literature<sup>41-43</sup>.

### 3.4.4 Fusion protein catalytic analysis

Two fusions, FnD-TEVp1 and non-binding control FnNB-TEVp1, were examined for catalytic activity via a quenched fluorophore peptide assay (Figure 3-12, Table 3-2). The  $k_{cat}$  values of both fusions were similar to wild-type ( $0.39$  s<sup>-1</sup> from Table 2-9 in Chapter 2) and the  $K_M$  values were only increased by approximately 2-fold (from wild-type value of  $24.3$   $\mu$ M from Table 2-9 in Chapter 2), yielding a decrease in catalytic efficiency of roughly two for both fusions. This result, combined with the widespread

preservation of all or most of the binding affinity of the ligand portion of the fusion, motivated progressing to *in vitro* cellular assays to examine if the synthetic reporter approach could be achieved with our particular fusions.

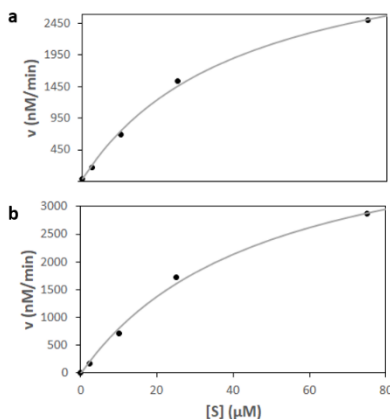


Figure 3-12. **Michaelis-Menten fit to kinetic data.** Example trials are shown for two TEVp1-based fusion proteins: a) FnNB-TEVp1 and b) FnD-TEVp1. The cleavage of peptide substrate 2-Abz-ENLYFQSGTK-Dnp, at various concentrations, by fusion protein (0.2 μM final concentration) was examined via fluorimetry. Background-subtracted fluorescence versus time graphs allowed for determination of reaction velocities and thus turnover number and Michaelis-Menten constant upon data fitting to a Michaelis-Menten kinetic model.

	$k_{cat}$ (1/s)	$K_m$ (uM)	Cat Eff (1/s/uM)
FnD-TEVp1	$0.38 \pm 0.03$	$46.9 \pm 5.1$	$0.0083 \pm 0.0005$
FnNB-TEVp1	$0.34 \pm 0.02$	$42.7 \pm 3.3$	$0.0079 \pm 0.0006$

Table 3-2. **Catalytic performance of fusion proteins FnD-TEVp1 and FnNB-TEVp1.** Errors shown are 95% confidence intervals.

### 3.4.5 *In vitro* cellular assay

The results of the *in vitro* cellular assay proved to be inconsistent across several iterations. In three trials, the use of FnD-TEVp1 to bestow an increased TEVp1 enzymatic activity to the surface of A431 cells was successful, as fluorescence from unquenching substrate outpaced that of the relevant negative controls in a statistically significant manner (Figure 3-13a). However, in four other trials this differentiation failed

to materialize (Figure 3-13b). Of the substrate-containing samples, the fluorescence range for the FnD-TEVp1-containing samples was ~45,000-80,000, while the two relevant negative controls ranged in fluorescence from ~45,000-60,000. Even in Figure 3-13a, two major limitations were observed. First, the increase in fluorescence was far below what was expected based on the experimentally established binding affinity and enzymatic kinetics. Assuming EGFR-bound fusion internalization was indeed negligible at 4 °C<sup>168-169</sup>, the increase in fluorescence compared to negative controls during the first five hours in Figure 3-13a was projected to be approximately 50,000 units, but experimentally it was short of 10,000 units. Potential explanations for this discrepancy are legion, but include a failure to adequately minimize fusion internalization with incubation on ice, a larger than anticipated decrease in TEVp turnover number due to the decreased temperature, and suboptimal fusion linker design that hinders bound fusion's enzyme component in accessing soluble substrate. Clearly, an increased activity TEVp variant would be beneficial in generating higher signal. The evolved variants from Chapter 2, which were engineered after the studies in the current chapter, offer a compelling solution. There is also potentially a need for linker length/composition optimization for this system so that bound fusions can more effectively project the TEVp1 enzyme from the surface to engage soluble substrate molecules. The second limitation revealed by the cellular reporter assay is the appreciable increase in fluorescence over time for both substrate-containing negative controls. The lack of cellular binding of FnNB should render minimal residual TEVp1 after washing for FnNB-TEV1p. More strikingly, the TEVp1-free FnD sample has no enzyme in the system at any point, and previous experiments involving peptide substrate mixed with FnD in *E. coli* lysate generated



negligible signal (Figure 3-14). The possibility that the FnNB-TEVp1-containing sample generated high background fluorescence due to inadequate washing out of the TEVp1-containing fusion seems unlikely as the rise in fluorescence generally mirrored that of the samples containing only FnD. Therefore, it appears that this particular TEVp1 substrate is far more labile than ideal in the environment of A431 cells over this particular timeline (i.e., approaching 24 hours). While engineering substrate stability (without sacrificing enzymatic activity) in future experiments to lower background non-TEVp-related fluorescence is an intriguing avenue to explore, and likely a necessary one for any *in vivo* experiments using TEVp-containing fusions to proceed successfully, the proof-of-concept in this *in vitro* cellular assay should still be achievable with sufficient bound fusion-specific signal. This perspective reiterates the importance of the aforementioned first point regarding improving enzymatic activity and/or optimizing linkers.

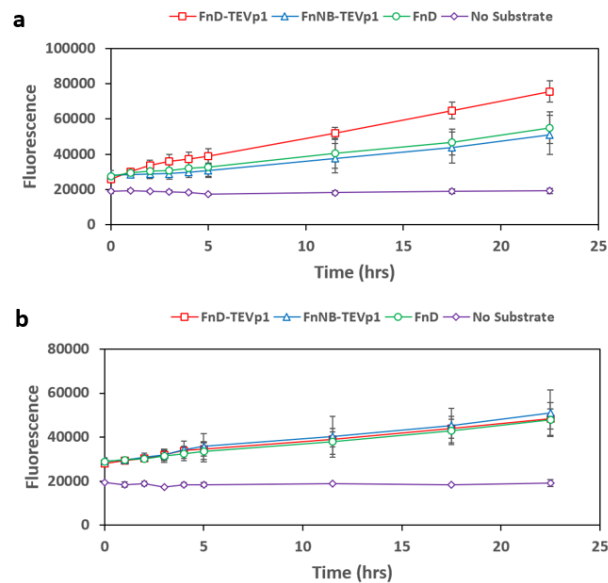


Figure 3-13. **Results of *in vitro* cellular assay.** a) Example of the FnD-TEVp1 sample successfully outpacing the relevant negative controls (i.e., no enzyme or non-binding fusion). b) Example of the FnD-TEVp1 failing to differentiate from the negative controls. Error bars shown are 95% confidence intervals. The inconsistency in the result, illustrated via numerous attempts at the cellular assay, indicated a need to increase enzymatic activity, optimize linkers, or decrease substrate lability to maximize signal/noise separation and validate the theoretical predictions and functionality of this approach.

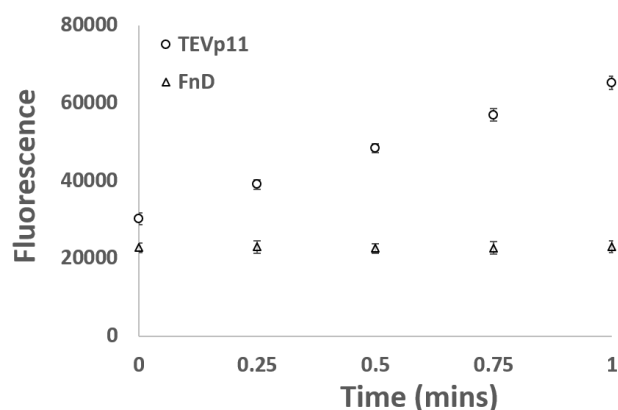


Figure 3-14. **FnD in *E. coli* lysate generates minimal background fluorescence signal compared to TEVp mutants in lysate.** Fluorescence from cutting a 2-Abz-ENLYFQSGTK-Dnp peptide was evaluated for protein and substrate concentrations of 0.1  $\mu$ M and 50  $\mu$ M, respectively, in a volume of 80  $\mu$ L. TEVp11 is shown here as a representative example. All measurements were made in triplicate.

### 3.5 Discussion

Our novel approach to using synthetic biomarkers involves targeting aberrant cell-surface receptor expression by first creating an artificially increased local enzyme concentration via the use of ligand-enzyme fusions, and then exploiting this by administering the exogenous enzyme's substrate to generate a reporter signal in congruence with the underlying abnormal receptor expression. This differs from the established synthetic reporter approach in the literature<sup>68,70</sup> that targets a disease based on its aberrant local soluble enzymatic upregulation. Thus, to first demonstrate the feasibility of this approach, we modeled our proposed approach to ensure sufficient product generation. Compared to negative controls that substituted non-binding fusions or only the ligand component, our fusions were predicted to generate added hourly fluorescence from the surface of 150,000 A431 cells of approximately 10,000 units. This, however, was predicated on a number of assumptions such as near-saturation of receptors with fusions and a 10-fold decrease in the fusion turnover number while on ice<sup>154</sup>. While the

specific product generation from bound fusions was predicted to be robust and easily outpace negative controls, any deviation from these assumptions, such as decreased receptor expression, poor fusion binding, greater-than-expected temperature-related decline in fusion  $k_{cat}$ , or difficulties with the bound fusion accessing soluble substrate, could threaten an otherwise favorable signal-to-noise ratio.

We established the ability to produce a variety of fusion proteins that simultaneously combined enzymatic activity with the binding capabilities of ligands. The modularity of this fusion protein system was effectively demonstrated, as tolerance for an array of enzyme and ligand components, as well as switching the order of each component, was illustrated by observing only a minor or moderate impact on the binding affinity and/or catalytic efficiency of the fusion when such changes were made. This result was not necessarily surprising as fusion systems with the same GGGSGGGKGGG linker<sup>98</sup>, as well as other similar linkers<sup>170</sup>, have previously demonstrated the ability to preserve much, if not all, of the activity of both fusion components for many combinations, although certainly some combinations were sub-optimal in that they failed to preserve an appreciable percentage of the wild-type component's activity<sup>98</sup>. In designing fusion proteins, adequate, much less optimal, activity is not guaranteed for two fusion components and a given linker<sup>170</sup>. While reasonable fusion modularity with the GGGSGGGKGGG linker has been suggested by successfully testing various components, the use of an enzyme as a component had not yet been attempted. Further, whereas longer linker length (i.e., two- to three-fold increase) for fusions with bivalent binding may harm overall performance<sup>98</sup>, other studies examining fusion systems with an enzyme component strongly suggest the benefit of increased linker length (i.e., 20 amino

acids instead of 5-10) to achieve fusion catalytic activity approaching that of the wild-type component<sup>171</sup>. It would be a worthwhile endeavor to evaluate different linkers to further improve upon the soluble activity of our linker-enzyme fusions. That being said, our fusions generally performed sufficiently like their wild-type components in a soluble setting to motivate progression to the *in vitro* cellular assay.

Linker optimization in terms of length, composition or hydrophobicity is crucial for engendering the desired specific function<sup>170</sup>, in this case in the cell-bound context. While our linkers may have been adequate for soluble fusion activity, our unique synthetic reporter approach to utilizing fusions requires the enzyme component to be active when the linked binder places the fusion in a cell-bound, instead of soluble, state. Enzymatic activity of fusions when bound may be hindered by a variety of factors including steric hindrance from neighboring cell-surface proteins. In light of the results from the *in vitro* cellular assay, other linkers, such as the rigid linker A(EAAAK)<sub>4</sub>ALEA(EAAAK)<sub>4</sub>A or the flexible linker (GGGS)<sub>4</sub><sup>170</sup>, should be investigated to examine if fusion-based catalytic activity can be preserved in the soluble state and also produce the necessary signal generation when bound to the cell-surface. The different properties of these linkers, namely added length, may assist the projection of surface-bound fusion's enzyme component away from the cell surface to enable proteolysis of soluble substrate, which could remedy the experimental limitation we observed of weak signal from cell surface-bound fusions cutting soluble substrate. Another option is to use a ligand, such as FnA, that targets an epitope less proximal to the cellular membrane. In hindsight it should be noted that FnD binds a membrane proximal region of EGFR<sup>98</sup>. Insufficient cutting could also possibly be remedied by substituting

TEVp1 for an improved version of TEVp. The cell assay studies were conducted before the entirety of Chapter 2, which is why a TEVp variant like TEVp11 has not yet been swapped into the fusion in place of TEVp1, but this is certainly compelling for future work.

A potential criticism of the cellular assay, and our extension of the synthetic reporter approach to quantifying receptor expression in general, either in the *in vitro* or *in vivo* setting, is that there is no guarantee that substrate cutting is occurring with cell surface-bound fusions exclusively. Rather, some bound fusions will slowly dissociate over time, and these soluble dissociated fusions could be contributing to some fraction of the catalysis. The lack of consistent fluorescence differentiation between ligand-enzyme samples and negative controls in the cellular assay over nearly 24 hours suggests this is of minimal concern. Further, it importantly should be of no consequence if catalysis is derived from still-bound fusions or fusions that have since dissociated from receptor, as the effective total fusion concentration in solution, and thus the rate of product generation, is still correlated to the underlying receptor burden.

In this study, we aimed to evaluate a variety of fusion proteins that combined binding and catalytic activity. Our work in this chapter demonstrates that recombinant ligand-enzyme fusions that preserve a significant portion, if not all, of the activity of the individual components can be produced. Further, we sought to study the modularity of a ligand-enzyme fusion system with a GGGSGGGKGGG linker, and we have demonstrated this particular linker allows for a modular fusion protein system that appears to work well with a variety of ligands and enzymes in both orientations. While these results in a soluble setting were satisfactory, we chiefly sought to establish an

extension of the synthetic reporter technology to quantify abnormal receptor expression. Our results in the *in vitro* cellular assay demonstrate a clear need for further optimization of this system to validate the practicality of this extension. While our mathematical model suggested this extension's feasibility, poor signal generation from receptor-bound fusions, among other technical issues, prevented this empirically. To boost this signal, there are numerous potential remedies: replace the TEVp variant used (TEVp1) with an improved mutant from Chapter 2 (e.g., TEVp2, TEVp10, TEVp11) or from an additional round of directed evolution. If, however, the lack of surface signal generation is less a result of insufficient enzymatic activity – which our mathematical modeling suggests it is not – and more an issue of soluble substrate accessibility for the bound enzyme, utilizing a ligand that targets an epitope further from the cell membrane or a linker with increased length to further project the linked enzyme from the cell surface are more likely to resolve the issue.

Any *in vivo* translation of our synthetic reporter adaptation after successful *in vitro* validation will require a significantly more stable substrate, particularly as serum, with its collection of proteases, is likely to be even less forgiving for the substrate<sup>172</sup>. If TEVp is to be used in the *in vivo* setting, substrate engineering would be necessary to decrease lability. A sensible approach would be to use the yeast display construct from Chapter 2, with TEVp1<sub>C151A</sub> installed to inactivate the enzyme, and design a substrate library to first identify those mutants that are impervious to background serum proteolysis (i.e., those cells with a lack of a decrease in HA signal). Of these stable substrate mutants, only those that contained some residual activity with TEVp could be propagated further. To identify this, improved substrate sequences inserted into the yeast display construct,

now with TEVp1, could be analyzed via flow cytometry to find samples that still permit TEVp-facilitated lysis above negative control signal (i.e., TEVp1<sub>C151A</sub>). Once this has been achieved, directed evolution to improve TEVp activity against this newly improved substrate could then be performed to recover, if not improve from, wild-type catalytic activity, now against the stable substrate. These experiments would also aid the *in vitro* cellular assay, as substrate lability was a major limitation as well. Another option that would primarily address the aforementioned weak signal generation, but also potentially assist in substrate lability, is to entirely replace the TEVp1 enzyme with another enzyme with improved catalytic efficiency, assuming that catalytic activity is mostly preserved in the fusion context. Srt7, for example, does not possess a meaningful improvement in catalytic efficiency, as its turnover number and Michaelis-Menten constant are approximately six-fold and five-fold higher, respectively<sup>26,106</sup>; however, if operating at a saturating substrate concentration, this six-fold improvement in  $k_{\text{cat}}$  would theoretically boost the signal generation rate by roughly the same amount, which could prove sufficient in differentiating from negative control signal. Regarding substrate lability, the Srt7 substrate may be more or less labile, with the latter option potentially being of significant benefit.

An additional avenue of compelling future work is to study the impact of various variables, such as substrate concentration, fusion concentration, and EGFR burden (either cell surface density or cell number), on reporter generation from cell-bound fusions once the base case cellular assay has been worked out. Below enzyme-saturating conditions, an increase in substrate very well might improve the signal-to-noise ratio. Increased fusion concentration should lead to more bound fusion at equilibrium up to a certain saturating

point, when further increases yield no meaningful improvement in signal. As EGFR burden increases (say from MDA-MB-435S to MDA-MB-231 to A431 or MDA-MB-468 cells<sup>173</sup>), assuming saturating amounts of fusion, signal should invariably increase. The proposed *in vivo* framework involves a multitude of moving parts, including but not limited to fusion dose, substrate dose, exposure time, and extent of disease. Thus, a heightened understanding of how altering these parameters impacts reporter generation *in vitro* will inform us regarding basic levers of system performance, as well as assist in future optimization of the *in vivo* framework.



## **Chapter 4 – Concluding Remarks and Future Work**

In general, there is a poor understanding of sequence-function relationships in enzymes<sup>18</sup>, and this work includes a significant advancement in the mapping of that relationship for a specific region – the active site – of TEVp, a protease of significant biotechnological importance<sup>6,90,174</sup>. Some of our key findings regarding TEVp's sequence-function relationship include the identification of clearly beneficial mutations at hotspots T30, S31, and L32, which were proximal to the substrate cut site, as well as F139, which interacts with the substrate distant from the cut site, and L204, which is distal from the substrate. We also identified multiple sites with substantially positive non-wild type enrichment at H167, N171, W211, K141, S170, K220 and P221. Future TEVp engineering efforts should investigate those areas directly adjacent to the hotspots we have identified. When many of these sites were included in a single second-generation library with diversity constrained by the first-generation results, we were able to identify those mutational pairings that tend to provide optimal catalytic efficiency: T30I, S31T or S31C, L32V, L204 or L204M, W211I or W211M, and K215R or K215E. In producing and evaluating a series of multi-mutants in accordance with these findings, we examined how specific mutations engendered particular catalytic function. For example, in examining a variety of V32-containing multi-mutants, we observed an improvement in  $k_{cat}$  at the expense of  $K_M$  for S31T (vs. S31C), L204M (vs. wild-type leucine), and K215E (vs. K215R). The ability to identify which mutations are likely to aid catalytic efficiency, and specifically which mutations grant a specific function (e.g., increased  $k_{cat}$  or decreased  $K_M$ ), will be of benefit to future TEVp engineering efforts that are explicitly trying to engineer a specific function.

The vast majority of our multi-mutant(s) improved catalytic efficiency, with a max improvement of 2.9-fold for TEVp2 (T30I/S31T/L32V/L204M/W211I/K215R). It should be noted that improvements in our tested multi-mutants were generally achieved with a lower  $K_M$  as opposed to a higher  $k_{cat}$ ; for example, TEVp2 successfully decreased  $K_M$  from 24.3  $\mu\text{M}$  to 5.5  $\mu\text{M}$ , but its turnover number fell, not as dramatically, from 0.39  $\text{s}^{-1}$  to 0.26  $\text{s}^{-1}$ . It is not necessarily clear if this was a function of the specific protease tested, the specific active site library screened, or rather a property of the display construct as currently constructed. Interestingly, per the calculations shown in the Discussion section of Chapter 2, the effective local concentration of both protease variant and substrate on the cell surface is 250  $\mu\text{M}$ , whereas the literature  $K_M$  for TEVp1 is 65  $\mu\text{M}$ <sup>26</sup>. This suggests we are likely already operating at close to saturating conditions, as the substrate concentration is substantially greater than the  $K_M$  value. One would expect this system to mostly favor mutants that improve  $k_{cat}$  as decreases in  $K_M$  should not improve catalytic cutting substantially. It is quite possible that the tendency to find beneficial mutations that mostly lowered  $K_M$  was related to the specific region of sequence space tested. That is, active site libraries may be heavily weighted towards  $K_M$ -lowering mutations as opposed to  $k_{cat}$ -increasing mutations, even when the substrate concentration at the cell surface favors finding the latter. Future efforts that use this construct to evaluate other regions of TEVp sequence space – such as sites distant from the active site – or other proteases will provide perspective. For other enzymes with larger values of  $K_M$  such as sortase A ( $320 \pm 50 \mu\text{M}$ <sup>26,106</sup>), there are adjustments to the surface display construct that should favor identifying mutants with improved  $k_{cat}$  as opposed to  $K_M$ , namely alterations that further increase the local substrate concentration.

This could be achieved by adding additional substrate sequences to the construct or decreasing the linker length, though not enough that tethered enzyme and substrate cannot physically interact. This increase in substrate concentration could more favor enzyme saturation conditions, and in this case improvements in turnover number would be more likely to generate increased cutting. This is an intriguing area of future work that should be investigated as it would further allow customization of enzyme screens to find mutations with a specific type of improved fitness.

It should also be noted that most of the other TEVp-focused directed evolution studies have been chiefly concerned with imparting novel substrate selectivity<sup>26</sup> or solubility<sup>175</sup>, making our study unique for TEVp in that its focus was on catalytic efficiency. TEVp, like all cysteine proteases, is noteworthy for its substrate selectivity, not its catalytic activity, especially when compared to serine proteases like factor Xa and thrombin, which possess much higher catalytic activity – including a turnover number that is approximately two orders of magnitude larger – but more substrate promiscuity<sup>90,93,176-179</sup>. Thus, activity engineering in TEVp is potentially of great import, as a protease with dually high activity and high substrate specificity could be immensely useful for a variety of applications<sup>90,174</sup>, and more TEVp-focused directed evolution attempts to improve catalytic efficiency in the future would be of substantial utility.

Another major goal of this dissertation was to provide further insight on optimal combinatorial library design, specifically in the context of sites around the protease active site. In past enzyme engineering efforts, including those involving TEVp<sup>175,26</sup>, random mutagenesis has almost exclusively been used for library construction. Given the immensity, sparsity and ruggedness of sequence space, this is a rather inefficient method

for scanning sequence space for increased fitness<sup>18</sup>. This area of enzyme engineering would benefit from the implementation of smarter, smaller libraries, which are becoming more commonplace in other areas of protein engineering<sup>180</sup>, that target particular sites with specific constrained diversity to maximize the chance of identifying beneficial mutations. We elected to create seven rational first-generation TEVp mini libraries (4-5 amino acids per library, all within 6 Angstroms of the substrate) by saturation mutagenesis, with beneficial individual mutations from the first-generation incorporated into a single second-generation library to capture any synergistic effects between beneficial mutations in different first-generation libraries. Individual libraries were capped at 5 amino acids because the exponential nature of combinatorics suggests that library size above this number, if created via saturation mutagenesis, is at best poorly sampleable using standard FACS techniques. An alternative approach that was considered was to create a single first-generation library based on the 34 amino acids examined, but with constrained sitewise diversity to limit the overall first-generation diversity so that reasonable sampling coverage could be achieved. In this approach, it is not fully clear what metrics one should use to constrain diversity. Relative solvent accessibility<sup>107</sup>, distance to the substrate cut site<sup>30,89</sup>, FoldX-predicted stability changes<sup>56</sup>, BLOSUM62-based chemical homology scoring<sup>61</sup> and consensus design using a database of homologous enzymes<sup>39</sup> are all viable options from the literature, but their comprehensive utilization in active site protease libraries has not been adequately studied. Thus, we opted to perform saturation mutagenesis to obtain comprehensive enrichment data on every possible amino acid mutation at each site, which would allow us to retrospectively evaluate whether any of the aforementioned metrics would have properly

constrained diversity or if they would have let beneficial, but unexpected, mutations evade analysis.

Relative solvent accessibility and distance to substrate cut site exhibited no meaningful relationship with the number of positively enriched mutations. FoldX-predicted stability changes did not fare significantly better as there was a very weak relationship with enrichment. However, a reasonably convincing argument could be made that our results suggest a general rule of precluding any mutations with FoldX-predicted  $\Delta\Delta G$  values at the extreme positive end (i.e.,  $>10$  kcal/mol; all mutations (19 total) with  $\Delta\Delta G > 15$  kcal/mol, and 94% (31/33) of mutations with  $\Delta\Delta G > 10$  kcal/mol, had negative enrichment. Consensus design in this case was shown to be poor at predicting both sites to target and diversity to incorporate. Of the six sites that benefit from our two-generation approach, consensus design would have likely identified all but one (W211); however, the consensus design recommendations for mutation at these five sites were poorly reflective of empirical improvement, as the optimal mutations of T30I, S31C, L32V, L204, L204M, K215E, and K215R were found in homologous enzymes at rates of 1.1, 1.4, 5.2, 3.8, 0.0, 5.4, and 1.2%, respectively. Further, consensus design also suggested inclusion of half of the remaining 28 amino acids that eventually yielded no benefit. The case for using chemical homology to constrain active site library diversity was stronger than these other metrics, though still imperfect. Homologous mutations were approximately three times as likely to be beneficial compared to non-homologous mutations, but fully neglecting non-homologous mutations is bound to exclude several unexpected beneficial mutations. For example, in the first-generation TEVp library (Figure 2-9) at site T30, the chemically homologous serine had a  $\log_2$  enrichment score of

0, but the chemically non-homologous isoleucine and valine had enrichment scores of 2.7 and 1.7, respectively. Thus, implementation of constrained diversity around the TEVp active site based on any of these poorly-predictive metrics would have invariably led to missing several beneficial, but unexpected, mutations, and our saturation mutagenesis approach with relatively thorough sampling was rather essential in finding these.

It should be stressed that a mutational analysis in the active site region around one protease does not necessarily generalize to distal mutations in the same protease or active site mutations for a different protease. It is quite possible that relative solvent accessibility would be more predictive for distal non-active site mutations, or that active site mutational analysis for another protease would yield different findings. Future work evaluating distal mutations in TEVp or active site mutations in another enzyme (e.g., sortase A or other cysteine proteases like tobacco vein mottling protease<sup>6</sup> or human rhinovirus 3C protease<sup>6,181</sup>) using a similar saturation mutagenesis approach should be applied to evaluate the predictive power of these metrics in different proteases and various protease domains. There is evidence that efforts to engineer protease activity by mutating the active site region likely result in as many failures as successes, so a combined approach targeting the active site and distal mutations may be of benefit<sup>113</sup>.

In this dissertation, we developed a yeast surface substrate/protease co-display construct for efficient analysis of protease activity. During our development, the concept was validated by Cochran and team<sup>27</sup> for clonal analysis but not high-throughput combinatorial evaluation. In that prior study, protease cleavage of tethered substrate was followed by 3-azido-1-propanamine conjugation to the tethered cut substrate and a click reaction between the incorporated 3-azido-1-propanamine and biotin-

dibenzocyclooctyne. The amount of cell surface-bound biotin, labeled with PE-avidin conjugate, was thus reflective of enzymatic cutting, so in this case an increased catalytic activity resulted in increased fluorescence. We skipped the 3-azido-1-propanamine conjugation step and everything thereafter, and instead looked for signal reduction after a substrate-adjacent epitope tag was removed from the display construct. This display system has the notable advantage of experimental simplicity compared to the procedure put forward by Cochran and team. While we successfully demonstrated this construct's ability to differentiate wild-type TEVp activity and negative controls with no activity, this general approach has a potential disadvantage: a theoretical difficulty in differentiating between two extremely active enzymes. Yet, the success of our efforts in screening libraries and enriching beneficial mutations over several rounds of sorting illustrates that this theoretical concern may not be a practical one. Additionally, there are opportunities to improve this display system by making changes to enable modulation of screening stringency. An ideal protocol would allow for significant surface display of construct without substrate cutting, followed by a brief and tunable amount of time when protease can cut tethered substrate before analysis immediately thereafter. Using a soluble inhibitor in the induction medium could help achieve this. In this format, as constructs are synthesized and displayed, their protease is inhibited; inhibitor would be washed out when display is satisfactory, allowing the experimenter to modulate the cutting time in accordance with desired screening stringency. A highly stringent screen would wash out inhibitor for a brief period of time, followed by reintroduction of inhibitor to stop further cutting. This could theoretically be useful in differentiating mutants with very high catalytic activity.

While some display constructs have been tested with multiple enzymes to demonstrate generalizability<sup>26</sup>, Cochran and team examined their construct's ability to differentiate activity for one enzyme: sortase A. We tested our modified construct on two enzymes, sortase A and TEVp, demonstrating activity differentiation in both cases and providing preliminary evidence that our system is generalizable. Cochran and team had also not published validation of their display construct's ability to screen combinatorial libraries. From that perspective, our work represents a significant advancement in the application of this type of construct to effectively screen combinatorial libraries for improved catalytic activity. This is especially important because heretofore most enzyme engineering display platforms for library screening have been designed for either bond-forming enzymes<sup>28</sup> or proteases localized to an intracellular compartment such as the endoplasmic reticulum<sup>26</sup>. To our knowledge, this is the first yeast cell surface display construct that has been validated for both differentiation of proteolytic activity and effective screening of combinatorial protease libraries.

We also explored a possible application of improved protease activity in ligand-enzyme fusion proteins to extend the concept of synthetic reporters to noninvasively detect disease based on aberrant receptor expression. Our *in vivo* mathematical modeling suggest that fusions composed of current ligands and enzymes can feasibly and robustly generate synthetic reporters from the surface of diseased cells to enable this noninvasive quantification and detection, specifically estimating urinary reporter signal three to four orders of magnitude above the limit of detection. The ligand-enzyme fusions, however, failed to consistently achieve this result in the *in vitro* cellular assay. It is possible this was a result of insufficient catalytic activity, which motivates the use of one of our



improved TEVp mutants, or a to-be-discovered TEVp mutant from future directed evolution efforts with further improvements in catalytic efficiency. It is also possible that the fusion linker length was insufficient, leading to inadequate projection away from the cell surface of the enzymatic component of receptor-bound fusions, resulting in poor engagement of soluble substrate. If this was the culprit, increased enzymatic activity, as documented by our mathematical model, would still present significant benefit in improving reporter generation in the *in vivo* and *in vitro* models. This application of improved protease activity represents an exciting way, among many others, to benefit from our directed evolution efforts by enabling the extension of a promising diagnostic technology. Despite the roadblock encountered in completing this synthetic reporter extension, our enhanced understanding of the system attributes has identified concrete areas to focus on to remedy the issue, and the methodology and discoveries in our protease engineering efforts make improving the synthetic reporter system more possible.

In summary, the work presented in this dissertation represent a significant advancement of a yeast cell surface display platform that enables cellular differentiation based on encoded protease activity and can be exploited for the screening of combinatorial libraries. This platform was successfully leveraged to improve the catalytic efficiency of a host of TEVp multi-mutants, up to 2.9-fold and generally via decreases in the Michaelis-Menten constant. A wealth of information was also obtained regarding optimal library design in the region around the active site of proteases, specifically TEVp. While chemical homology, and FoldX-predicted stability changes to a lesser extent, may have a role in constraining diversity, non-saturation mutagenesis approaches run the risk of missing several beneficial, but unexpected, mutations. Our sequence-

function mapping of TEVp will also be of benefit to future engineering endeavors. The broader biotechnological community will benefit from improved activity TEVp variants, and we also explored an application of these variants with regards to diagnosing receptor overexpression using synthetic urinary reporters. We demonstrated *in vivo* feasibility of this system using our novel ligand-enzyme fusions in a mathematical model, and estimated the benefit of improved enzymatic activity in this system. This exciting connection between engineering protease activity and improving the performance of a novel and noninvasive molecular diagnostic technology should be furthered in the future to achieve the diagnostic aims of this extension of the synthetic reporter approach.

## References

---

1. Barrett AJ, Mcdonald JK. Nomenclature: protease, proteinase and peptidase. *Biochem J.* 1986;237(3):935.
2. Rawlings ND, Barrett AJ, Bateman A. MEROPS: the database of proteolytic enzymes, their substrates and inhibitors. *Nucleic Acids Res.* 2012;40:D343-50.
3. Jemli S, Ayadi-zouari D, Hlima HB, Bejar S. Biocatalysts: application and engineering for industrial purposes. *Crit Rev Biotechnol.* 2016;36(2):246-58.
4. Kirk O, Borchert TV, Fuglsang CC. Industrial enzyme applications. *Curr Opin Biotechnol.* 2002;13(4):345-51.
5. Neurath H, Walsh KA. Role of proteolytic enzymes in biological regulation (a review). *Proc Natl Acad Sci USA.* 1976;73(11):3825-32.
6. Mótyán JA, Tóth F, Tózsér J. Research applications of proteolytic enzymes in molecular biology. *Biomolecules.* 2013;3(4):923-42.
7. Li Q, Yi L, Marek P, Iverson BL. Commercial proteases: present and future. *FEBS Lett.* 2013;587(8):1155-63.
8. Craik CS, Page MJ, Madison EL. Proteases as therapeutics. *Biochem J.* 2011;435(1):1-16.
9. Chanalia P, Gandhi D, Jodha D, Singh J. Applications of microbial proteases in pharmaceutical industry: An overview. *Rev Med Microbiol.* 2011;22(4):96–101.
10. Gupta R, Beg QK, Lorenz P. Bacterial alkaline proteases: molecular approaches and industrial applications. *Appl Microbiol Biotechnol.* 2002;59(1):15-32.
11. Wehr MC, Laage R, Bolz U, et al. Monitoring regulated protein-protein interactions using split TEV. *Nat Methods.* 2006;3(12):985-93.
12. Hanes J, Plückthun A. In vitro selection and evolution of functional proteins by using ribosome display. *Proc Natl Acad Sci USA.* 1997;94(10):4937-42.
13. Mccafferty J, Griffiths AD, Winter G, Chiswell DJ. Phage antibodies: filamentous phage displaying antibody variable domains. *Nature.* 1990;348(6301):552-4.
14. Beerli RR, Bauer M, Buser RB, et al. Isolation of human monoclonal antibodies by mammalian cell display. *Proc Natl Acad Sci USA.* 2008;105(38):14336-41.
15. Boder ET, Wittrup KD. Yeast surface display for screening combinatorial polypeptide libraries. *Nat Biotechnol.* 1997;15(6):553-7.
16. Cherf GM, Cochran JR. Applications of Yeast Surface Display for Protein Engineering. *Methods Mol Biol.* 2015;1319:155-75.
17. Moore JC, Rodriguez-granillo A, Crespo A, et al. "Site and Mutation"-Specific Predictions Enable Minimal Directed Evolution Libraries. *ACS Synth Biol.* 2018;7(7):1730-1741.
18. Romero PA, Arnold FH. Exploring protein fitness landscapes by directed evolution. *Nat Rev Mol Cell Biol.* 2009;10(12):866-76.
19. Aharoni A, Gaidukov L, Khersonsky O, Mcgould S, Roodveldt C, Tawfik DS. The 'evolvability' of promiscuous protein functions. *Nat Genet.* 2005;37(1):73-6.
20. Packer MS, Liu DR. Methods for the directed evolution of proteins. *Nat Rev Genet.* 2015;16(7):379-94.

21. Lipovsek D, Antipov E, Armstrong KA, et al. Selection of horseradish peroxidase variants with enhanced enantioselectivity by yeast surface display. *Chem Biol.* 2007;14(10):1176-85.
22. Antipov E, Cho AE, Wittrup KD, Klivanov AM. Highly L and D enantioselective variants of horseradish peroxidase discovered by an ultrahigh-throughput selection method. *Proc Natl Acad Sci USA.* 2008;105(46):17694-9.
23. White KA, Zegelbone PM. Directed evolution of a probe ligase with activity in the secretory pathway and application to imaging intercellular protein-protein interactions. *Biochemistry.* 2013;52(21):3728-39.
24. Han SY, Zhang JH, Han ZL, Zheng SP, Lin Y. Combination of site-directed mutagenesis and yeast surface display enhances *Rhizomucor miehei* lipase esterification activity in organic solvent. *Biotechnol Lett.* 2011;33(12):2431-8.
25. Fushimi T, Miura N, Shintani H, Tsunoda H, Kuroda K, Ueda M. Mutant firefly luciferases with improved specific activity and dATP discrimination constructed by yeast cell surface engineering. *Appl Microbiol Biotechnol.* 2013;97(9):4003-11.
26. Yi L, Gebhard MC, Li Q, Taft JM, Georgiou G, Iverson BL. Engineering of TEV protease variants by yeast ER sequestration screening (YESS) of combinatorial libraries. *Proc Natl Acad Sci USA.* 2013;110(18):7229-34.
27. Lim S, Glasgow JE, Filsinger interrante M, Storm EM, Cochran JR. Dual display of proteins on the yeast cell surface simplifies quantification of binding interactions and enzymatic bioconjugation reactions. *Biotechnol J.* 2017;12(5)
28. Chen I, Dorr BM, Liu DR. A general strategy for the evolution of bond-forming enzymes using yeast display. *Proc Natl Acad Sci USA.* 2011;108(28):11399-404.
29. Smith JM. Natural selection and the concept of a protein space. *Nature.* 1970;225(5232):563-4.
30. Hermes JD, Blacklow SC, Knowles JR. Searching sequence space by definably random mutagenesis: improving the catalytic potency of an enzyme. *Proc Natl Acad Sci USA.* 1990;87(2):696-700.
31. Taverna DM, Goldstein RA. Why are proteins marginally stable?. *Proteins.* 2002;46(1):105-9.
32. Axe DD. Estimating the prevalence of protein sequences adopting functional enzyme folds. *J Mol Biol.* 2004;341(5):1295-315.
33. Keefe AD, Szostak JW. Functional proteins from a random-sequence library. *Nature.* 2001;410(6829):715-8.
34. Taverna DM, Goldstein RA. Why are proteins so robust to site mutations?. *J Mol Biol.* 2002;315(3):479-84.
35. Xia Y, Levitt M. Funnel-like organization in sequence space determines the distributions of protein stability and folding rate preferred by evolution. *Proteins.* 2004;55(1):107-14.
36. Govindarajan S, Goldstein RA. Evolution of model proteins on a foldability landscape. *Proteins.* 1997;29(4):461-6.
37. Bloom JD, Labthavikul ST, Otey CR, Arnold FH. Protein stability promotes evolvability. *Proc Natl Acad Sci USA.* 2006;103(15):5869-74.

38. Wagner A. Robustness and evolvability: a paradox resolved. *Proc Biol Sci.* 2008;275(1630):91-100.
39. Porebski BT, Buckle AM. Consensus protein design. *Protein Eng Des Sel.* 2016;29(7):245-51.
40. Patrick WM, Firth AE. Strategies and computational tools for improving randomized protein libraries. *Biomol Eng.* 2005;22(4):105-12.
41. Lutz S. Beyond directed evolution--semi-rational protein engineering and design. *Curr Opin Biotechnol.* 2010;21(6):734-43.
42. Wijma HJ, Floor RJ, Janssen DB. Structure- and sequence-analysis inspired engineering of proteins for enhanced thermostability. *Curr Opin Struct Biol.* 2013;23(4):588-94.
43. Wijma HJ, Floor RJ, Jekel PA, Baker D, Marrink SJ, Janssen DB. Computationally designed libraries for rapid enzyme stabilization. *Protein Eng Des Sel.* 2014;27(2):49-58.
44. Magliery TJ. Protein stability: computation, sequence statistics, and new experimental methods. *Curr Opin Struct Biol.* 2015;33:161-8.
45. Steiner K, Schwab H. Recent advances in rational approaches for enzyme engineering. *Comput Struct Biotechnol J.* 2012;2:e201209010.
46. Fasan R, Meharena Y, Snow CD, Poulos TL, Arnold FH. Evolutionary history of a specialized p450 propane monooxygenase. *J Mol Biol.* 2008;383(5):1069-80.
47. Spiller B, Gershenson A, Arnold FH, Stevens RC. A structural view of evolutionary divergence. *Proc Natl Acad Sci USA.* 1999;96(22):12305-10.
48. Shimotohno A, Oue S, Yano T, Kuramitsu S, Kagamiyama H. Demonstration of the importance and usefulness of manipulating non-active-site residues in protein design. *J Biochem.* 2001;129(6):943-8.
49. Reetz MT, Boccola M, Carballeira JD, Zha D, Vogel A. Expanding the range of substrate acceptance of enzymes: combinatorial active-site saturation test. *Angew Chem Int Ed Engl.* 2005;44(27):4192-6.
50. Matsumura I, Ellington AD. In vitro evolution of beta-glucuronidase into a beta-galactosidase proceeds through non-specific intermediates. *J Mol Biol.* 2001;305(2):331-9.
51. Park S, Morley KL, Horsman GP, Holmquist M, Hult K, Kazlauskas RJ. Focusing mutations into the *P. fluorescens* esterase binding site increases enantioselectivity more effectively than distant mutations. *Chem Biol.* 2005;12(1):45-54.
52. Paramesvaran J, Hibbert EG, Russell AJ, Dalby PA. Distributions of enzyme residues yielding mutants with improved substrate specificities from two different directed evolution strategies. *Protein Eng Des Sel.* 2009;22(7):401-11.
53. Steipe B, Schiller B, Plückthun A, Steinbacher S. Sequence statistics reliably predict stabilizing mutations in a protein domain. *J Mol Biol.* 1994;240(3):188-92.
54. Lehmann M, Pasamontes L, Lassen SF, Wyss M. The consensus concept for thermostability engineering of proteins. *Biochim Biophys Acta.* 2000;1543(2):408-415.

55. Das R, Baker D. Macromolecular modeling with rosetta. *Annu Rev Biochem.* 2008;77:363-82.
56. Schymkowitz J, Borg J, Stricher F, Nys R, Rousseau F, Serrano L. The FoldX web server: an online force field. *Nucleic Acids Res.* 2005;33:W382-8.
57. Wijma HJ, Floor RJ, Jekel PA, Baker D, Marrink SJ, Janssen DB. Computationally designed libraries for rapid enzyme stabilization. *Protein Eng Des Sel.* 2014;27(2):49-58.
58. Christensen NJ, Kepp KP. Accurate stabilities of laccase mutants predicted with a modified FoldX protocol. *J Chem Inf Model.* 2012;52(11):3028-42.
59. Cargill M, Altshuler D, Ireland J, et al. Characterization of single-nucleotide polymorphisms in coding regions of human genes. *Nat Genet.* 1999;22(3):231-8.
60. Ng PC, Henikoff S. Predicting deleterious amino acid substitutions. *Genome Res.* 2001;11(5):863-74.
61. Henikoff S, Henikoff JG. Amino acid substitution matrices from protein blocks. *Proc Natl Acad Sci USA.* 1992;89(22):10915-9.
62. Grantham R. Amino acid difference formula to help explain protein evolution. *Science.* 1974;185(4154):862-4.
63. Sneath PH. Relations between chemical structure and biological activity in peptides. *J Theor Biol.* 1966;12(2):157-95.
64. Yampolsky LY, Stoltzfus A. The exchangeability of amino acids in proteins. *Genetics.* 2005;170(4):1459-72.
65. D'souza AL, Tseng JR, Pauly KB, et al. A strategy for blood biomarker amplification and localization using ultrasound. *Proc Natl Acad Sci USA.* 2009;106(40):17152-7.
66. Haun JB, Castro CM, Wang R, et al. Micro-NMR for rapid molecular analysis of human tumor samples. *Sci Transl Med.* 2011;3(71):71ra16.
67. Surinova S, Schiess R, Hüttenhain R, Cerciello F, Wollscheid B, Aebersold R. On the development of plasma protein biomarkers. *J Proteome Res.* 2011;10(1):5-16.
68. Warren AD, Kwong GA, Wood DK, Lin KY, Bhatia SN. Point-of-care diagnostics for noncommunicable diseases using synthetic urinary biomarkers and paper microfluidics. *Proc Natl Acad Sci USA.* 2014;111(10):3671-6.
69. Kwong GA, Von maltzahn G, Murugappan G, et al. Mass-encoded synthetic biomarkers for multiplexed urinary monitoring of disease. *Nat Biotechnol.* 2013;31(1):63-70.
70. Lin KY, Kwong GA, Warren AD, Wood DK, Bhatia SN. Nanoparticles that sense thrombin activity as synthetic urinary biomarkers of thrombosis. *ACS Nano.* 2013;7(10):9001-9.
71. Pusztai L, Mazouni C, Anderson K, Wu Y, Symmans WF. Molecular classification of breast cancer: limitations and potential. *Oncologist.* 2006;11(8):868-77.
72. Sapino A, Goia M, Recupero D, Marchiò C. Current Challenges for HER2 Testing in Diagnostic Pathology: State of the Art and Controversial Issues. *Front Oncol.* 2013;3:129.
73. Gown AM. Current issues in ER and HER2 testing by IHC in breast cancer. *Mod Pathol.* 2008;21 Suppl 2:S8-S15.

74. Ooi A, Takehana T, Li X, et al. Protein overexpression and gene amplification of HER-2 and EGFR in colorectal cancers: an immunohistochemical and fluorescent in situ hybridization study. *Mod Pathol*. 2004;17(8):895-904.
75. Atkins D, Reiffen KA, Tegtmeier CL, Winther H, Bonato MS, Störkel S. Immunohistochemical detection of EGFR in paraffin-embedded tumor tissues: variation in staining intensity due to choice of fixative and storage time of tissue sections. *J Histochem Cytochem*. 2004;52(7):893-901.
76. Arnedos M, Nerurkar A, Osin P, A'hern R, Smith IE, Dowsett M. Discordance between core needle biopsy (CNB) and excisional biopsy (EB) for estrogen receptor (ER), progesterone receptor (PgR) and HER2 status in early breast cancer (EBC). *Ann Oncol*. 2009;20(12):1948-52.
77. Wiener RS, Schwartz LM, Woloshin S, Welch HG. Population-based risk for complications after transthoracic needle lung biopsy of a pulmonary nodule: an analysis of discharge records. *Ann Intern Med*. 2011;155(3):137-44.
78. Yoo J, Park C, Yi G, Lee D, Koo H. Active Targeting Strategies Using Biological Ligands for Nanoparticle Drug Delivery Systems. *Cancers (Basel)*. 2019;11(5).
79. Chen K, Conti PS. Target-specific delivery of peptide-based probes for PET imaging. *Adv Drug Deliv Rev*. 2010;62(11):1005-22.
80. Zhao N, Qin Y, Liu H, Cheng Z. Tumor-Targeting Peptides: Ligands for Molecular Imaging and Therapy. *Anticancer Agents Med Chem*. 2018;18(1):74-86.
81. Davids T, Schmidt M, Böttcher D, Bornscheuer UT. Strategies for the discovery and engineering of enzymes for biocatalysis. *Curr Opin Chem Biol*. 2013;17(2):215-20.
82. Jochens H, Bornscheuer UT. Natural diversity to guide focused directed evolution. *Chembiochem*. 2010;11(13):1861-6.
83. Ebert MC, Pelletier JN. Computational tools for enzyme improvement: why everyone can - and should - use them. *Curr Opin Chem Biol*. 2017;37:89-96.
84. Chaparro-riggers JF, Polizzi KM, Bommarius AS. Better library design: data-driven protein engineering. *Biotechnol J*. 2007;2(2):180-91.
85. Yamakura F, Sugio S, Hiraoka BY, Ohmori D, Yokota T. Pronounced conversion of the metal-specific activity of superoxide dismutase from *Porphyromonas gingivalis* by the mutation of a single amino acid (Gly155Thr) located apart from the active site. *Biochemistry*. 2003;42(36):10790-9.
86. Ragland DA, Nalivaika EA, Nalam MN, et al. Drug resistance conferred by mutations outside the active site through alterations in the dynamic and structural ensemble of HIV-1 protease. *J Am Chem Soc*. 2014;136(34):11956-63.
87. Agniswamy J, Louis JM, Roche J, Harrison RW, Weber IT. Structural Studies of a Rationally Selected Multi-Drug Resistant HIV-1 Protease Reveal Synergistic Effect of Distal Mutations on Flap Dynamics. *PLoS ONE*. 2016;11(12):e0168616.
88. Wilding M, Hong N, Spence M, Buckle AM, Jackson CJ. Protein engineering: the potential of remote mutations. *Biochem Soc Trans*. 2019;47(2):701-711.

89. Tracewell CA, Arnold FH. Directed enzyme evolution: climbing fitness peaks one amino acid at a time. *Curr Opin Chem Biol.* 2009;13(1):3-9.
90. Waugh DS. An overview of enzymatic reagents for the removal of affinity tags. *Protein Expr Purif.* 2011;80(2):283-93.
91. Carrington JC, Dougherty WG. Small nuclear inclusion protein encoded by a plant potyvirus genome is a protease. *J Virol.* 1987;61(8):2540-8.
92. Dougherty WG, Parks TD. Post-translational processing of the tobacco etch virus 49-kDa small nuclear inclusion polypeptide: identification of an internal cleavage site and delimitation of VPg and proteinase domains. *Virology.* 1991;183(2):449-56.
93. Kapust RB, Tözsér J, Fox JD, et al. Tobacco etch virus protease: mechanism of autolysis and rational design of stable mutants with wild-type catalytic proficiency. *Protein Eng.* 2001;14(12):993-1000.
94. Parks TD, Howard ED, Wolpert TJ, Arp DJ, Dougherty WG. Expression and purification of a recombinant tobacco etch virus NIa proteinase: biochemical analyses of the full-length and a naturally occurring truncated proteinase form. *Virology.* 1995;210(1):194-201.
95. Lucast LJ, Batey RT, Doudna JA. Large-scale purification of a stable form of recombinant tobacco etch virus protease. *BioTechniques.* 2001;30(3):544-6, 548, 550.
96. Bordat A, Houvenaghel MC, German-Retana S. Gibson assembly: an easy way to clone potyviral full-length infectious cDNA clones expressing an ectopic VPg. *Virol J.* 2015;12:89.
97. Swers JS, Kellogg BA, Wittrup KD. Shuffled antibody libraries created by in vivo homologous recombination and yeast surface display. *Nucleic Acids Res.* 2004;32(3):e36.
98. Hackel BJ, Neil JR, White FM, Wittrup KD. Epidermal growth factor receptor downregulation by small heterodimeric binding proteins. *Protein Eng Des Sel.* 2012;25(2):47-57.
99. Breibeck J, Skerra A. The polypeptide biophysics of proline/alanine-rich sequences (PAS): Recombinant biopolymers with PEG-like properties. *Biopolymers.* 2018;109(1).
100. Schlapschy M, Binder U, Börger C, et al. PASylation: a biological alternative to PEGylation for extending the plasma half-life of pharmaceutically active proteins. *Protein Eng Des Sel.* 2013;26(8):489-501.
101. Gorbalenya AE, Donchenko AP, Blinov VM, Koonin EV. Cysteine proteases of positive strand RNA viruses and chymotrypsin-like serine proteases. A distinct protein superfamily with a common structural fold. *FEBS Lett.* 1989;243(2):103-14.
102. Phan J, Zdanov A, Evdokimov AG, Tropea JE, Peters III HK, Kapust RB, Li M, Wlodawer A, Waugh DS. Structural basis for the substrate specificity of tobacco etch virus protease. *J Biol Chem.* 2002;277(52):50564-72.
103. Ton-that H, Mazmanian SK, Faull KF, Schneewind O. Anchoring of surface proteins to the cell wall of *Staphylococcus aureus*. Sortase catalyzed in vitro transpeptidation reaction using LPXTG peptide and NH(2)-Gly(3) substrates. *J Biol Chem.* 2000;275(13):9876-81.



104. Ilangovan U, Ton-that H, Iwahara J, Schneewind O, Clubb RT. Structure of sortase, the transpeptidase that anchors proteins to the cell wall of *Staphylococcus aureus*. *Proc Natl Acad Sci USA*. 2001;98(11):6056-61.
105. Hirakawa H, Ishikawa S, Nagamune T. Design of Ca<sup>2+</sup>-independent *Staphylococcus aureus* sortase A mutants. *Biotechnol Bioeng*. 2012;109(12):2955-61.
106. Hirakawa H, Ishikawa S, Nagamune T. Ca<sup>2+</sup>-independent sortase-A exhibits high selective protein ligation activity in the cytoplasm of *Escherichia coli*. *Biotechnol J*. 2015;10(9):1487-92.
107. Chen H, Zhou HX. Prediction of solvent accessibility and sites of deleterious mutations from protein sequence. *Nucleic Acids Res*. 2005;33(10):3193-9.
108. Tien MZ, Meyer AG, Sydykova DK, Spielman SJ, Wilke CO. Maximum allowed solvent accessibilities of residues in proteins. *PLoS ONE*. 2013;8(11):e80635.
109. Cochran JR, Kim YS, Lippow SM, Rao B, Wittrup KD. Improved mutants from directed evolution are biased to orthologous substitutions. *Protein Eng Des Sel*. 2006;19(6):245-53.
110. Bateman A, Coin L, Durbin R, et al. The Pfam protein families database. *Nucleic Acids Res*. 2004;32(Database issue):D138-41.
111. Bloom JD, Arnold FH. In the light of directed evolution: pathways of adaptive protein evolution. *Proc Natl Acad Sci USA*. 2009;106 Suppl 1:9995-10000.
112. Klesmith JR, Su L, Wu L, et al. Retargeting CD19 Chimeric Antigen Receptor T Cells via Engineered CD19-Fusion Proteins. *Mol Pharm*. 2019;16(8):3544-3558.
113. Shusta EV, Vanantwerp J, Wittrup KD. Biosynthetic polypeptide libraries. *Curr Opin Biotechnol*. 1999;10(2):117-22.
114. Klesmith JR, Bacik JP, Wrenbeck EE, Michalczyk R, Whitehead TA. Trade-offs between enzyme fitness and solubility illuminated by deep mutational scanning. *Proc Natl Acad Sci USA*. 2017;114(9):2265-2270.
115. Nannemann DP, Birmingham WR, Scism RA, Bachmann BO. Assessing directed evolution methods for the generation of biosynthetic enzymes with potential in drug biosynthesis. *Future Med Chem*. 2011;3(7):809-19.
116. Pierce MC, Richards-kortum R. Low-cost, portable optical imaging systems for cancer diagnosis. *Conf Proc IEEE Eng Med Biol Soc*. 2010;2010:1093-6.
117. Etzioni R, Urban N, Ramsey S, et al. The case for early detection. *Nat Rev Cancer*. 2003;3(4):243-52.
118. Tabár L, Vitak B, Chen TH, et al. Swedish two-county trial: impact of mammographic screening on breast cancer mortality during 3 decades. *Radiology*. 2011;260(3):658-63.
119. Church TR, Black WC, Aberle DR, et al. Results of initial low-dose computed tomographic screening for lung cancer. *N Engl J Med*. 2013;368(21):1980-91.
120. Hori SS, Gambhir SS. Mathematical model identifies blood biomarker-based early cancer detection strategies and limitations. *Sci Transl Med*. 2011;3(109):109ra116.
121. Wang AZ, Langer R, Farokhzad OC. Nanoparticle delivery of cancer drugs. *Annu Rev Med*. 2012;63:185-98.

122. Pollock NR, Rolland JP, Kumar S, et al. A paper-based multiplexed transaminase test for low-cost, point-of-care liver function testing. *Sci Transl Med.* 2012;4(152):152ra129.
123. Posthuma-trumpie GA, Korf J, Van amerongen A. Lateral flow (118signal)assay: its strengths, weaknesses, opportunities and threats. A literature survey. *Anal Bioanal Chem.* 2009;393(2):569-82.
124. Hanahan D, Weinberg RA. The hallmarks of cancer. *Cell.* 2000;100(1):57-70.
125. Yarden Y, Sliwkowski MX. Untangling the ErbB 118signaling network. *Nat Rev Mol Cell Biol.* 2001;2(2):127-37.
126. Ullrich A, Schlessinger J. Signal transduction by receptors with tyrosine kinase activity. *Cell.* 1990;61(2):203-12.
127. Hynes NE, Lane HA. ERBB receptors and cancer: the complexity of targeted inhibitors. *Nat Rev Cancer.* 2005;5(5):341-54.
128. Dawson JP, Berger MB, Lin CC, Schlessinger J, Lemmon MA, Ferguson KM. Epidermal growth factor receptor dimerization and activation require ligand-induced conformational changes in the dimer interface. *Mol Cell Biol.* 2005;25(17):7734-42.
129. Yarden Y. The EGFR family and its ligands in human cancer: 118signaling mechanisms and therapeutic opportunities. *Eur J Cancer.* 2001;37 Suppl 4:S3-8.
130. Schlessinger J. Cell signaling by receptor tyrosine kinases. *Cell.* 2000;103(2):211-25.
131. Bethune G, Bethune D, Ridgway N, Xu Z. Epidermal growth factor receptor (EGFR) in lung cancer: an overview and update. *J Thorac Dis.* 2010;2(1):48-51.
132. Gorgoulis V, Aninos D, Mikou P, et al. Expression of EGF, TGF-alpha and EGFR in squamous cell lung carcinomas. *Anticancer Res.* 1992;12(4):1183-7.
133. Hatanpaa KJ, Burma S, Zhao D, Habib AA. Epidermal growth factor receptor in glioma: signal transduction, neuropathology, imaging, and radioresistance. *Neoplasia.* 2010;12(9):675-84.
134. Bhargava R, Gerald WL, Li AR, et al. EGFR gene amplification in breast cancer: correlation with epidermal growth factor receptor mRNA and protein expression and HER-2 status and absence of EGFR-activating mutations. *Mod Pathol.* 2005;18(8):1027-33.
135. Hanawa M, Suzuki S, Dobashi Y, et al. EGFR protein overexpression and gene amplification in squamous cell carcinomas of the esophagus. *Int J Cancer.* 2006;118(5):1173-80.
136. Maiti GP, Mondal P, Mukherjee N, et al. Overexpression of EGFR in head and neck squamous cell carcinoma is associated with inactivation of SH3GL2 and CDC25A genes. *PLoS ONE.* 2013;8(5):e63440.
137. Lee J, Moon C. Current status of experimental therapeutics for head and neck cancer. *Exp Biol Med (Maywood).* 2011;236(4):375-89.
138. Hechtman JF, Polydorides AD. HER2/neu gene amplification and protein overexpression in gastric and gastroesophageal junction adenocarcinoma: a review of histopathology, diagnostic testing, and clinical implications. *Arch Pathol Lab Med.* 2012;136(6):691-7.

139. Slamon DJ, Clark GM, Wong SG, Levin WJ, Ullrich A, Mcguire WL. Human breast cancer: correlation of relapse and survival with amplification of the HER-2/neu oncogene. *Science*. 1987;235(4785):177-82.
140. Slamon DJ, Godolphin W, Jones LA, et al. Studies of the HER-2/neu proto-oncogene in human breast and ovarian cancer. *Science*. 1989;244(4905):707-12.
141. Dawood S, Broglio K, Buzdar AU, Hortobagyi GN, Giordano SH. Prognosis of women with metastatic breast cancer by HER2 status and trastuzumab treatment: an institutional-based review. *J Clin Oncol*. 2010;28(1):92-8.
142. Nicholson RI, Gee JM, Harper ME. EGFR and cancer prognosis. *Eur J Cancer*. 2001;37 Suppl 4:S9-15.
143. Pusztai L, Mazouni C, Anderson K, Wu Y, Symmans WF. Molecular classification of breast cancer: limitations and potential. *Oncologist*. 2006;11(8):868-77.
144. Sapino A, Goia M, Recupero D, Marchiò C. Current Challenges for HER2 Testing in Diagnostic Pathology: State of the Art and Controversial Issues. *Front Oncol*. 2013;3:129.
145. Gown AM. Current issues in ER and HER2 testing by IHC in breast cancer. *Mod Pathol*. 2008;21 Suppl 2:S8-S15.
146. Ooi A, Takehana T, Li X, et al. Protein overexpression and gene amplification of HER-2 and EGFR in colorectal cancers: an immunohistochemical and fluorescent in situ hybridization study. *Mod Pathol*. 2004;17(8):895-904.
147. Atkins D, Reiffen KA, Tegtmeier CL, Winther H, Bonato MS, Störkel S. Immunohistochemical detection of EGFR in paraffin-embedded tumor tissues: variation in staining intensity due to choice of fixative and storage time of tissue sections. *J Histochem Cytochem*. 2004;52(7):893-901.
148. Arnedos M, Nerurkar A, Osin P, A'hern R, Smith IE, Dowsett M. Discordance between core needle biopsy (CNB) and excisional biopsy (EB) for estrogen receptor (ER), progesterone receptor (PgR) and HER2 status in early breast cancer (EBC). *Ann Oncol*. 2009;20(12):1948-52.
149. Wiener RS, Schwartz LM, Woloshin S, Welch HG. Population-based risk for complications after transthoracic needle lung biopsy of a pulmonary nodule: an analysis of discharge records. *Ann Intern Med*. 2011;155(3):137-44.
150. Case BA, Hackel BJ. Synthetic and natural consensus design for engineering charge within an affibody targeting epidermal growth factor receptor. *Biotechnol Bioeng*. 2016;113(8):1628-38.
151. Kruziki MA, Bhatnagar S, Woldring DR, Duong VT, Hackel BJ. A 45-Amino-Acid Scaffold Mined from the PDB for High-Affinity Ligand Engineering. *Chem Biol*. 2015;22(7):946-56.
152. Hackel BJ, Kapila A, Wittrup KD. Picomolar affinity fibronectin domains engineered utilizing loop length diversity, recursive mutagenesis, and loop shuffling. *J Mol Biol*. 2008;381(5):1238-52.
153. Liu JC, Heilshorn SC, Tirrell DA. Comparative cell response to artificial extracellular matrix proteins containing the RGD and CS5 cell-binding domains. *Biomacromolecules*. 2004;5(2):497-504.

154. Nallamsetty S, Kapust RB, Tözsér J, et al. Efficient site-specific processing of fusion proteins by tobacco vein mottling virus protease in vivo and in vitro. *Protein Expr Purif.* 2004;38(1):108-15.
155. Wu Z, Hong H, Zhao X, Wang X. Efficient expression of sortase A from in and its enzymatic characterizations. *Bioresour Bioprocess.* 2017;4(1):13.
156. Hirata IY, Sedenho Cezari MH, Nakaie CR, et al. Internally quenched fluorogenic protease substrates: Solid-phase synthesis and fluorescence spectroscopy of peptides containing *ortho*-aminobenzoyl/dinitrophenyl groups as donor-acceptor pairs. *Lett Pept Sci.* 1995;1:299–308.
157. Poreba M, Szalek A, Rut W, et al. Highly sensitive and adaptable fluorescence-quenched pair discloses the substrate specificity profiles in diverse protease families. *Sci Rep.* 2017;7:43135.
158. Poreba M, Drag M. Current strategies for probing substrate specificity of proteases. *Curr Med Chem.* 2010;17(33):3968-95.
159. Stanton P, Richards S, Reeves J, et al. Epidermal growth factor receptor expression by human squamous cell carcinomas of the head and neck, cell lines and xenografts. *Br J Cancer.* 1994;70(3):427-33.
160. Haigler H, Ash JF, Singer SJ, Cohen S. Visualization by fluorescence of the binding and internalization of epidermal growth factor in human carcinoma cells A-431. *Proc Natl Acad Sci USA.* 1978;75(7):3317-21.
161. Li N, Nguyen HH, Byrom M, Ellington AD. Inhibition of cell proliferation by an anti-EGFR aptamer. *PLoS ONE.* 2011;6(6):e20299.
162. Kruziki MA, Case BA, Chan JY, et al. Cu-Labeled Gp2 Domain for PET Imaging of Epidermal Growth Factor Receptor. *Mol Pharm.* 2016;13(11):3747-3755.
163. Thurber GM, Wittrup KD. A mechanistic compartmental model for total antibody uptake in tumors. *J Theor Biol.* 2012;314:57-68.
164. Schmidt MM, Wittrup KD. A modeling analysis of the effects of molecular size and binding affinity on tumor targeting. *Mol Cancer Ther.* 2009;8(10):2861-71.
165. Kawamoto M, Horibe T, Kohno M, Kawakami K. A novel transferrin receptor-targeted hybrid peptide disintegrates cancer cell membrane to induce rapid killing of cancer cells. *BMC Cancer.* 2011;11:359.
166. Kurien BT, Everds NE, Scofield RH. Experimental animal urine collection: a review. *Lab Anim.* 2004;38(4):333-61.
167. Zhang S, Garcia-d'angeli A, Brennan JP, Huo Q. Predicting detection limits of enzyme-linked immunosorbent assay (ELISA) and bioanalytical techniques in general. *Analyst.* 2014;139(2):439-45.
168. Pinilla-macua I, Sorkin A. Methods to study endocytic trafficking of the EGF receptor. *Methods Cell Biol.* 2015;130:347-67.
169. Sorkin A, Duex JE. Quantitative analysis of endocytosis and turnover of epidermal growth factor (EGF) and EGF receptor. *Curr Protoc Cell Biol.* 2010;Chapter 15:Unit 15.14.
170. Chen X, Zaro JL, Shen WC. Fusion protein linkers: property, design and functionality. *Adv Drug Deliv Rev.* 2013;65(10):1357-69.

171. Bergeron LM, Gomez L, Whitehead TA, Clark DS. Self-renaturing enzymes: design of an enzyme-chaperone chimera as a new approach to enzyme stabilization. *Biotechnol Bioeng.* 2009;102(5):1316-22.
172. Böttger R, Hoffmann R, Knappe D. Differential stability of therapeutic peptides with different proteolytic cleavage sites in blood, plasma and serum. *PLoS ONE.* 2017;12(6):e0178943.
173. Anido J, Matar P, Albanell J, et al. ZD1839, a specific epidermal growth factor receptor (EGFR) tyrosine kinase inhibitor, induces the formation of inactive EGFR/HER2 and EGFR/HER3 heterodimers and prevents heregulin signaling in HER2-overexpressing breast cancer cells. *Clin Cancer Res.* 2003;9(4):1274-83.
174. Wang W, Wildes CP, Pattarabanjird T, et al. A light- and calcium-gated transcription factor for imaging and manipulating activated neurons. *Nat Biotechnol.* 2017;35(9):864-871.
175. Van den berg S, Löfdahl PA, Härd T, Berglund H. Improved solubility of TEV protease by directed evolution. *J Biotechnol.* 2006;121(3):291-8.
176. Isetti G, Maurer MC. Employing mutants to study thrombin residues responsible for factor XIII activation peptide recognition: a kinetic study. *Biochemistry.* 2007;46(9):2444-52.
177. Bromfield KM, Quinsey NS, Duggan PJ, Pike RN. Approaches to selective peptidic inhibitors of factor Xa. *Chem Biol Drug Des.* 2006;68(1):11-9.
178. Gasparian ME, Ostapchenko VG, Schulga AA, Dolgikh DA, Kirpichnikov MP. Expression, purification, and characterization of human enteropeptidase catalytic subunit in *Escherichia coli*. *Protein Expr Purif.* 2003;31(1):133-9.
179. Long AC, Orr DC, Cameron JM, Dunn BM, Kay J. A consensus sequence for substrate hydrolysis by rhinovirus 3C proteinase. *FEBS Lett.* 1989;258(1):75-8.
180. Woldring DR, Holec PV, Stern LA, Du Y, Hackel BJ. A Gradient of Sitewise Diversity Promotes Evolutionary Fitness for Binder Discovery in a Three-Helix Bundle Protein Scaffold. *Biochemistry.* 2017;56(11):1656-1671.
181. Meister SW, Hendrikse NM, Löfblom J. Directed evolution of the 3C protease from coxsackievirus using a novel fluorescence-assisted intracellular method. *Biol Chem.* 2019;400(3):405-415.



Primer 27 TTTCCCTACACGACGCTCTTCCGATCTCTCGGCCAGCAATTTTGCCAAC  
 Primer 28 GTTCAGACGTGTGCTCTTCCGATCTTTTCACAGGCTGAAAGGGCTCCTC  
 Primer 29 AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTT  
 Primer 30 CAAGCAGAAGACGGCATAACGAGATATCGTGGTACTGGAGTTCAGACGTGTGCTCTTCC  
 Primer 31 CAAGCAGAAGACGGCATAACGAGATTGAGTGGTACTGGAGTTCAGACGTGTGCTCTTCC  
 Primer 32 CAAGCAGAAGACGGCATAACGAGATTCGCTGGTACTGGAGTTCAGACGTGTGCTCTTCC  
 Primer 33 CAAGCAGAAGACGGCATAACGAGATGCCATGGTACTGGAGTTCAGACGTGTGCTCTTCC  
 Primer 34 CAAGCAGAAGACGGCATAACGAGATAAAATGGTACTGGAGTTCAGACGTGTGCTCTTCC  
 Primer 35 CAAGCAGAAGACGGCATAACGAGATTGTTGGGTACTGGAGTTCAGACGTGTGCTCTTCC  
 Primer 36 CAAGCAGAAGACGGCATAACGAGATATTCCGGTACTGGAGTTCAGACGTGTGCTCTTCC  
 Primer 37 CAAGCAGAAGACGGCATAACGAGATAGCTAGGTACTGGAGTTCAGACGTGTGCTCTTCC  
 Primer 38 CAAGCAGAAGACGGCATAACGAGATGTATAGGTACTGGAGTTCAGACGTGTGCTCTTCC  
 Primer 39 CAAGCAGAAGACGGCATAACGAGATTGCTGAGGTACTGGAGTTCAGACGTGTGCTCTTCC  
 Primer 40 CAAGCAGAAGACGGCATAACGAGATGTCGTCGTACTGGAGTTCAGACGTGTGCTCTTCC  
 Primer 41 CAAGCAGAAGACGGCATAACGAGATCGATTAGTACTGGAGTTCAGACGTGTGCTCTTCC  
 Primer 42 CAAGCAGAAGACGGCATAACGAGATGCTGTAGTACTGGAGTTCAGACGTGTGCTCTTCC  
 Primer 43 CAAGCAGAAGACGGCATAACGAGATATTATAGTACTGGAGTTCAGACGTGTGCTCTTCC  
 Primer 44 GAAGGGACCAAATCCGATGCCATAACGYAVTARYTGTGTGACCGTCGCTCTCGTTTGT  
 Primer 45 GAAGGGACCAAATCCGATGCCATAACGYAVTARYTGTGTGACCGTCGCTCTCGTTTGT  
 Primer 46 GAAGGGACCAAATCCGATGCCATAACGYAVTACATGTGTGACCGTCGCTCTCGTTTGT  
 Primer 47 GAAGGGACCAAATCCGATGCCATAACGYACAARYTGTGTGACCGTCGCTCTCGTTTGT  
 Primer 48 GAAGGGACCAAATCCGATGCCATAACGYACAACATGTGTGACCGTCGCTCTCGTTTGT  
 Primer 49 GAAGGGACCAAATCCGATGCCATAACGYACAACATGTGTGACCGTCGCTCTCGTTTGT  
 Primer 50 GAAGGGACCAAATCCGATGCCATAACBACAARYTGTGTGACCGTCGCTCTCGTTTGT  
 Primer 51 GAAGGGACCAAATCCGATGCCATAACBACAACATGTGTGACCGTCGCTCTCGTTTGT  
 Primer 52 GAAGGGACCAAATCCGATGCCATAACAAVTARYTGTGTGACCGTCGCTCTCGTTTGT  
 Primer 53 GAAGGGACCAAATCCGATGCCATAACAACAVTARYTGTGTGACCGTCGCTCTCGTTTGT  
 Primer 54 GAAGGGACCAAATCCGATGCCATAACAACAARYTGTGTGACCGTCGCTCTCGTTTGT  
 Primer 55 GAAGGGACCAAATCCGATGCCATAACAACAACATGTGTGACCGTCGCTCTCGTTTGT  
 Primer 56 TATGGCATCGGATTTGGTCCCTC  
 Primer 57 TAATACTGAGTCGGCGTTTATCTCCACCCAGAGACCCACTGTTGCGCCTCTTGATTAGT  
 Primer 58 TAATACTGAGTCGGCGTTAAVTCTCCACCCAGAGACCCACTGTTGCGCCTCTTGATTAGT  
 Primer 59 TAATACTGAGTCGGCGTTACATCTCCACCCAGAGACCCACTGTTGCGCCTCTTGATTAGT  
 Primer 60 AACGCCGACTCAGTATTATGGGGCGGTMMAMRRGSWGTTTATGGTGRAGCCTGAGGAGCCC  
 Primer 61 AACGCCGACTCAGTATTATGGGGCGGTMMAMRRGSWGTTTCCGGTGRAGCCTGAGGAGCCC  
 Primer 62 AACGCCGACTCAGTATTATGGGGCGGTMMAMRRGSWGTTTGGGGTGRAGCCTGAGGAGCCC  
 Primer 63 AACGCCGACTCAGTATTATGGGGCGGTTGRRRGSWGTTTATGGTGRAGCCTGAGGAGCCC  
 Primer 64 AACGCCGACTCAGTATTATGGGGCGGTTGRRRGSWGTTTCCGGTGRAGCCTGAGGAGCCC  
 Primer 65 AACGCCGACTCAGTATTATGGGGCGGTTGRRRGSWGTTTGGGGTGRAGCCTGAGGAGCCC  
 Primer 66 AACGCCGACTCAGTATTAATRGGCGGTMMAMRRGSWGTTTATGGTGRAGCCTGAGGAGCCC  
 Primer 67 AACGCCGACTCAGTATTAATRGGCGGTMMAMRRGSWGTTTCCGGTGRAGCCTGAGGAGCCC  
 Primer 68 AACGCCGACTCAGTATTAATRGGCGGTMMAMRRGSWGTTTGGGGTGRAGCCTGAGGAGCCC  
 Primer 69 AACGCCGACTCAGTATTAATRGGCGGTTGRRRGSWGTTTATGGTGRAGCCTGAGGAGCCC  
 Primer 70 AACGCCGACTCAGTATTAATRGGCGGTTGRRRGSWGTTTCCGGTGRAGCCTGAGGAGCCC  
 Primer 71 AACGCCGACTCAGTATTAATRGGCGGTTGRRRGSWGTTTGGGGTGRAGCCTGAGGAGCCC  
 Primer 72 AGGAGATATACATATGGCTAGCGGTGAATCCCTTTTAAGGGACCACG  
 Primer 73 GATGATGGTATGGTGGGATCCCTGGGAGTAGACTAACTCGTTCATAAGC  
 Primer 74a AGTCAGCTGCAGCTACCCGAAACGGGAGGTCCATGGTCAGTC  
 Primer 74b GACTGACCATGGACCTCCCCTTTCGGGTAGCTGCAGCTGACT  
 Primer 75a AGTCAGCTGCAGACCGAGGGGTTGGGGCCACCATGGTCAGTC  
 Primer 75b GACTGACCATGGTGGCCCAACCCCTCGGTCTGCAGCTGACT  
 Primer 76 AGTCAGGGATCCCAAGCTAAACCTCAAATTCC  
 Primer 77 AGTCAGCCGCGTTTACTTCTGTAGCTACAAAG

Supplemental Table 2-1. List of gBlocks and oligonucleotide primers used.

<u>Name</u>	<u>Sequence</u>
1	AGCTAGGCTAGCGTTTCCGATGTTCCGAGGGACCTGGAG
2	CTAGCTGGATCCCTGGGGTGGTTTGTCAATTTCTGTTTCG
3	AGCTAGGCTAGCGTTTCTGATGTTCCGAGGGACCTGGAAG
4	CTAGCTGGATCCCTGGGATGGTTTGTCAATTTCTGTTTCG
5	AGCTAGGCTAGCGTTTCTGATGTTCCGAGGGACCTGGAAG
6	CTAGCTGGATCCCTGGGATGGTTTGTCAATTTCTGTTTCG
7	AGCTAGGCTAGCGTGAGCGACGTTCCAAGAGATCTGGAAGTCGTG
8	CTAGCTGGATCCGGTACGATAATTAATGCTGATCGGACGGCTGGAC
9	AGCTAGGCTAGCGCACAGGCCAACTACGCCAAAGAAATGTG
10	CTAGCTGGATCCGTTGGGTGCTTGTGCGTCGTTCAATTTCTG
11	AGCTAGGCTAGCAAATTTTGGGCGACTGTATCGCGGGGCGACTC
12	CTAGCTGGATCCCGGACGCACGCGGGTCACGTAATAAATATGATAC
13	AGCTAGGCTAGCGGTGAATCCCTTTTTTAAGGGACCAC
14	CTAGCTGGATCCCTGGGAGTAGACTAACTCGTTCATAAGCTG
15	AGCTAGGCTAGCCAAGCTAAACCTCAAATTCGAAAAGATAAATCAAAAAGTG
16	CTAGCTGGATCCCTTGACTTCTGTAGCTACAAAGATTTTACGTG
17	AGCTAGGGTACCGGTGAATCCCTTTTTTAAGGGACCAC
18	CTAGCTGAGCTCCTGGGAGTAGACTAACTCGTTCATAAGCTG
19	AGCTAGGGTACCCAAGCTAAACCTCAAATTCGAAAAGATAAATCAAAAAGTG
20	CTAGCTGAGCTCTTGACTTCTGTAGCTACAAAGATTTTACGTG
21	AGCTAGGGTACCGTTTCCGATGTTCCGAGGGACCTGGAG
22	CTAGCTGAGCTCCTGGGGTGGTTTGTCAATTTCTGTTTCG
23	AGCTAGGGTACCGTTTCTGATGTTCCGAGGGACCTGGAAG
24	CTAGCTGAGCTCCTGGGATGGTTTGTCAATTTCTGTTTCG
25	AGCTAGGGTACCGTGAGCGACGTTCCAAGAGATCTGGAAGTCGTG
26	CTAGCTGAGCTCGGTACGATAATTAATGCTGATCGGACGGCTGGAC
27	AGCTAGGGTACCAAATTTTGGGCGACTGTATCGCGGGGCGACTC
28	CTAGCTGAGCTCCGGACGCACGCGGGTCACGTAATAAATATGATAC

Supplemental Table 3-1. **List of oligonucleotide primers used for PCR amplification.**



## Supplemental Figures

---

### Mathematical Model

$$V_{\text{cell}} = (4/3) * \pi * (d_{\text{cell}}/2)^3$$

$$V_{\text{tumor}} = (4/3) * \pi * (d_{\text{tumor}}/2)^3$$

$$N_{\text{cells/tumor}} = (V_{\text{tumor}}/V_{\text{cell}}) * (1 - \phi)$$

$$N_{\text{EGFR/tumor}} = N_{\text{cells/tumor}} * N_{\text{EGFR/cell}}$$

$$n_{\text{EGFR/tumor}} = N_{\text{EGFR/tumor}}/N_{\text{Av}}$$

$$M_{\text{EGFR}} = n_{\text{EGFR/tumor}}/V_{\text{tumor}}$$

$$M_{\text{fusions}} = M_{\text{EGFR}} * x_{\text{bound}}$$

$M_{\text{fusions}}$ ,  $k_{\text{cat}}$ ,  $K_M$  and Michaelis-Menten kinetics were then used to numerically calculate  $[R]_{\text{tumor}}$  after 1 hour, accounting for substrate depletion and using starting substrate concentration  $[S]_0$

$$D \text{ (tumor:blood dilution factor)} = V_{\text{tumor}}/V_{\text{blood}}$$

$$[R]_{\text{blood}} = [R]_{\text{tumor}} * D$$

$$C \text{ (blood:urine concentrating factor)} = V_{\text{blood}}/V_{\text{urine}}$$

$$[R]_{\text{urine}} = [R]_{\text{blood}} * C$$

Supplemental Figure 3-1. **Mathematical model for the extension of the *in vivo* synthetic reporter approach to detecting aberrant receptor expression.** Relevant variables are V (volume), d (diameter), N (number), n (mole), M (molarity), and [R] (synthetic reporter concentration).

### Constants

$$k_{\text{cat}} = 0.3 \text{ s}^{-1}$$

$$K_M = 65 \text{ } \mu\text{M}$$

$$d_{\text{cell}} = 20 \text{ } \mu\text{m}$$

$$N_{\text{Av}} = 6.02 \times 10^{23} \text{ molecules/mol}$$

$$V_{\text{urine}} = 21 \text{ } \mu\text{L}$$

$$V_{\text{blood}} = 1.5 \text{ mL}$$

$$N_{\text{EGFR/cell}} = 2 \times 10^6$$

$$\phi \text{ (void fraction)} = 0.2$$

$$x_{\text{bound}} \text{ (fraction of EGFR bound to fusion)} = 0.02$$

### Variables

$$d_{\text{tumor}}$$

$$[S]_0$$