Engineering the MS2 Capsid Using Systematic Protein Fitness Landscapes

By

Emily C Hartman

A dissertation submitted in partial satisfaction of the

requirements for the degree of

Doctor of Philosophy

in

Chemistry

in the

Graduate Division

of the

University of California, Berkeley

Committee in Charge:

Professor Matthew Francis, Co-Chair
Professor Danielle Tullman-Ercek, Co-Chair
Professor Michael Marletta
Professor John Dueber

Spring 2019

Abstract

Engineering the MS2 Capsid Using Systematic Protein Fitness Landscapes

by

Emily C. Hartman

Doctor of Philosophy in Chemistry

University of California, Berkeley

Professor Matthew Francis, Co-Chair
Professor Danielle Tullman-Ercek, Co-Chair

Self-assembling proteins are emerging as compelling solutions in both drug delivery and vaccine development. Typically, a self-assembling protein, such as a virus-like particle, is repurposed as a drug delivery vehicle by decorating a native, well-characterized viral capsid through chemical or genetic modification. While this was successful in several applications, it does not allow the physical properties of the particle itself to be adjusted in a rational manner. Indeed, there are many instances in which native capsid properties—such as stability, size, binding, chemical reactivity, among others—are non-ideal for the application at hand.

Herein, we developed and employed a new technique to study the mutability of self-assembling virus-like particles. This technique allows facile generation of highly targeted libraries, as well as simultaneous evaluation of many variants in a single pool. Furthermore, this technique is compatible with many different direct functional selections, such as heat, acid, or chemical challenges, enabling granular insight into how mutations affect chemical and physical properties. We evaluated how single amino acid mutations affect self-assembly of a model virus-like particle (**Chapter 2**). We then applied this technique to study the two-amino acid mutability of a small and flexible loop (**Chapter 3**). We also studied how N-terminal extensions alter the stability and chemical reactivity of the virus-like particle (**Chapter 4**). Finally, we sought to understand how mutations can affect the quaternary geometry of a self-assembling particle (**Chapter 5**). In sum, by allowing the simultaneous evaluation of many variants in a single pool, this work has generated the most systematic data available regarding the effects of individual amino acid substitutions on the resulting properties of virus-like particles.

Dedication

To my parents, Scott and Cathleen.

Table of Contents

List of Figures and Tables

List of Appendices

Chapter 4 Supplementary Figures

List of Abbreviations

1D-AFL	One-dimensional apparent fitness landscape
2D-AFL	Two-dimensional apparent fitness landscape
2PCA	2-pyridinecarboxaldehyde
abTYR	Tyrosinase enzyme isolated from *Agaricus bisporus*
AFL	Apparent Fitness Landscape
AFS	Apparent Fitness Score
ASA	Accessible Surface Area
CP	Coat Protein
HiPerX	High-Performing N-terminal Extensions
$K_3Fe(CN)_6$	Potassium ferricyanide
MI	Mutability Index
NNK	A degenerate codon. N = all four bases, K = G/T only
SEC	Size-exclusion chromatography
SERF	Self-encoding removable fragment
SyMAPS	Systematic Mutagenesis and Assembled Particle Selection
T	Triangulation number
TEM	Transmission Electron Microscopy
Tm	Melting temperature
TR-RNA	Translational operator
VLP	Virus-like particle
WT	Wild type

Acknowledgements

To the Francis and Tullman-Ercek labs, thank you not only for the many wonderful hours discussing science together, but also for your mentorship, support, and friendship, over the past five years. Thank you to the undergraduates who have worked with me—it's been wonderful to watch you grow into strong and independent scientists.

To Matt and Danielle, thank you for the time, effort, and support (plus hours of videochat!) that made this project successful. Your guidance and insight have been crucial throughout my PhD, and I am lucky to have had both of you as mentors.

The NDSEG fellowship was instrumental in allowing me to pursue this project, and I am grateful for funding for this project.

Thank you to the *Berkeley Science Review* and the Chemistry Departmental Climate Group. Both student groups taught me the important of balance during graduate school. The work you are doing is incredibly important, and I'm proud to have been a part of both of these efforts.

To my wonderful friends, thank you for your encouragement and friendship over the past five years, if it was a hike, phone call, or just a quick cup of coffee on a tough day. Special thanks to friends in my cohort and at Berkeley, who have been with me through every step of this PhD.

To my family, I can't express how much your support, patience, and love has meant to me. You've kept me grounded and sane, and I absolutely am where I am today because of you.

To Ben, I am lucky to be able to write two separate acknowledgements to you. One as a scientist who has had an incredibly impact on my work—so many important insights were from our discussions about my project. And of course, one as my partner. I love you and am so grateful for your encouragement and support.

Chapter 1: Introduction

The following is adapted from Hartman and Tullman-Ercek; Curr. Opin. Syst. Biol., 2019 with permission

1.A. Self-assembling proteins and virus-like particles

Self-assembling proteins perform many critical functions throughout biological systems, acting as chaperones, cages, scaffolds, and nanoscale machines, and more. Self-assembling proteins are becoming increasingly important in industrial and health applications as well, including as drug delivery or imaging vehicles[1–5], nanoscale bioreactors[6–10], and light harvesting devices[11,12], among many others. In recent years, researchers have made great strides in predicting how a *de novo* protein sequence will fold into self-assembled structures[3,13,14]. However, to date, our ability to predict how amino acid substitutions will impact the self-assembly of native structures remains remarkably limited.

In this work, we sought to expand our understanding of how mutations alter the assembly, stability, and reactivity of self-assembled protein cages. We selected virus-like particles as a model to study self-assembly. VLPs can be derived from bacteriophages (i.e. Qβ, MS2, PP7, and P22), plant viruses (i.e. TMV, CCMV, BMV) and human viruses (i.e. HBV, HPV), among many others[15]. Virus-like particles (VLPs) are closed-shell protein containers that tend to be composed of a few proteins or even a single protein, making them a relatively simple model. VLPs can be overexpressed to high titers in bacteria, and closed shell structures form spontaneously during expression. VLPs are not infectious, as they do not contain a genome. In addition, many VLPs encapsulate available negatively charged molecules, including nucleic acids, during assembly, providing a useful genotype-to-phenotype link if assembly is permitted.

While VLPs are an excellent model to study self-assembly, they are also widely studied for a broad array of applications, meaning a better understanding of VLP mutability has immediate impact these other areas of research. The diverse pool of potential applications of virus-like particles were recently described in a review by Glasgow and Tullman-Ercek[15]. These applications typically take advantage of several unique properties of these particles, including the ability to chemical or genetically modify the particle at spatially defined positions; an interior protein core that can protect cargo; and pores that permit diffusion in and out of the protein shell. All of these applications are grounded in the ability of VLPs to self-assemble spontaneously. However, a full understanding of the subtle energetics involved in the self-assembly process is still lacking, hindering the design of new closed-shell particles with desirable chemical or physical properties.

1.B. MS2 Bacteriophage

The MS2 bacteriophage is a well-studied particle with a long scientific history. In the past fifty years, the MS2 bacteriophage has been studied as an infectious bacteriophage virus, a noninfectious VLP, and, perhaps most famously, as an RNA tag. The MS2 bacteriophage was first described in 1961[16,17]. At 3,569 nucleotides, its genome is one of the smallest described to date[18], and it was also the first complete genome to be sequenced[19].

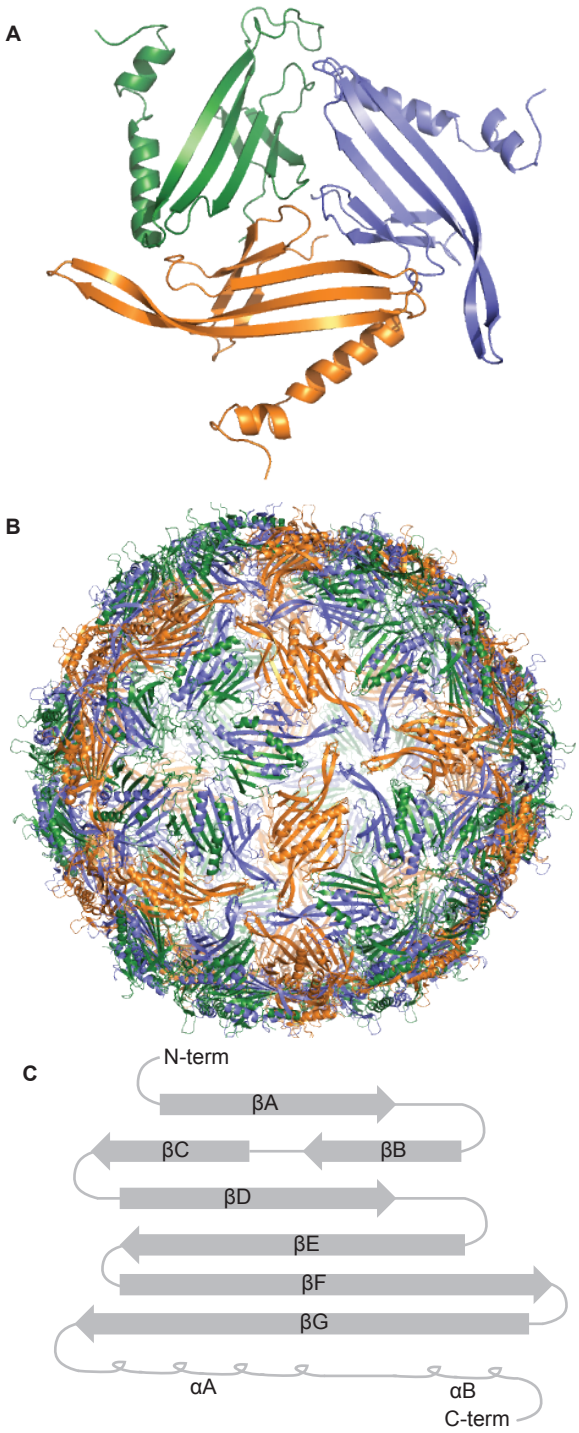


Figure 1.1. The structure of the MS2 coat protein (CP). A) The asymmetric unit of the MS2 CP. The monomer in the A form is shown in blue, the B form in green, and the C form in orange. B) The complete VLP structure of the MS2 CP with the same color structure as shown in (A). C) The secondary structure of a monomer of the MS2 CP is shown using accepted alphabetic nomenclature.

The viable MS2 bacteriophage natively infects male specific *Escherichia coli* (F+, F', or HFr)[20] and is closely-related to several other well-studied bacteriophage virions, including Qβ[20]. The MS2 bacteriophage genome encodes for only four proteins: maturation, coat, replicase, and lysis proteins. The infectious virion particle contains only two of these proteins: a single copy of the maturation protein (also called Protein A), which binds bacterial pili during infection, along with 178 copies of the coat protein. Recently, two high-resolution cryoEM structures of the complete infectious MS2 virion were published, which for the first time solved the structure of the interior genome in the infectious virion, as well as the structure of the maturation protein[18,21].

MS2 can also form virus-like particles, which are composed of 180 copies of the coat protein (CP) but do not contain the genome and thus are noninfectious. The VLP spontaneously self-assembles when the single coat protein gene is overexpressed from a plasmid in a bacterial host, such as *E. coli*. To assemble into closed-shell structure, a coat protein adopts three conformations (A, B, and C) to form a quasi-equivalent T=3 shell[22]. Structurally, the VLP is composed primarily of β-sheets and has 32 pores of 2 nm patterned around the structure (**Figure 1.1A,B**)[23]. Each β-sheet or turn has a widely-used name, which will be used throughout this thesis (**Figure 1.1C**). The MS2 VLP is also surprisingly robust and can tolerate temperatures up to 68 ºC, acidity to pH 2, and basicity to pH 12 without unfolding.

Finally, MS2 can act as an RNA tag

to study RNA localization. Often referred to as MS2 tagging, this technique allows study of RNA localization[24], purification of RNA–protein complexes[25], or recruitment nucleic acid to a containers like extracellular vesicles[26], among other applications.This technique uses a nonassembling variant of the MS2 coat protein to bind and track the location of RNA within a cell, repurposing a tight protein–RNA interaction used to package the native genome. During replication, the MS2 coat protein binds tightly to a specific RNA stem loop sequence, called the translation repressor (TR-RNA), which in turn represses transcription as the concentration of coat protein rises intracellularly[18]. The coat protein also undergoes a conformational change upon binding, which allows it to achieve the necessary asymmetry to form a closed shell structure. Not only has this high-affinity interaction has been studied as a model of sequence-specific RNA–protein interactions, it has also been has used to study and track RNA in a sequence-specific manner. This technique has been used for several decades continues to be widely used today[24].

1.B.i. MS2 VLPs as a model of self-assembly

We chose to study protein self-assembly using the MS2 VLP as a model protein shell. The assembly, structure, and utility of the MS2 VLP have been well characterized[27-32], making it an ideal candidate to study self-assembly. MS2 VLPs are attractive targets due to their promise as delivery vehicles for small molecules, nucleic acids, or proteins to host cells[8,28,31,32]. Alternatively, the MS2 VLPs have be used to integrate light absorbing chromophores with photocatalysts[33] or position dye molecules near the surface of gold nanoparticles housed within the capsid shells[34]. The interior cavity is available to load cargo, protecting it from the external environment, while the exterior can be chemically modified to display targeting groups such as peptides or antibodies[1,30]. These VLPs are surprisingly long-lived in the bloodstream of mice, stable to serum, and accumulate in several tissues of interest[1,5].

Beginning in the 1990s, researchers have employed protein engineering techniques to learn about the MS2 VLP in great detail. These efforts have successfully been used to describe the protein–RNA binding interaction, which in turn informed the many applications of MS2 tagging. Protein engineering techniques also improved display of foreign epitopes, showing the promise of MS2 VLP as a vaccine display platform. The work in this thesis seeks to expand on these efforts, using comprehensive and systematic techniques to develop a more complete understanding of the influences of amino acids on the assembly behavior of this closed-shell structure.

1.B.ii. Protein engineering to study MS2–RNA binding interactions

Early protein engineering efforts with the MS2 bacteriophage sought to better describe the MS2–RNA binding interaction. These studies interrogated the function and behavior of MS2 mutants, and they largely relied either on plaque assays[35] or, in the early 1990s, a clever two-plasmid system to probe RNA binding efficiency[36]. Plaque assays either evaluated infectivity or, in later reports, the ability of overexpressed mutants to interrupt plaque formation by binding the translational repressor[37]. The two-plasmid system more directly assayed coat protein–translational repressor interactions. In this screen, the TR-RNA sequence was inserted upstream of a β-galactosidase enzyme. If the coat protein variant

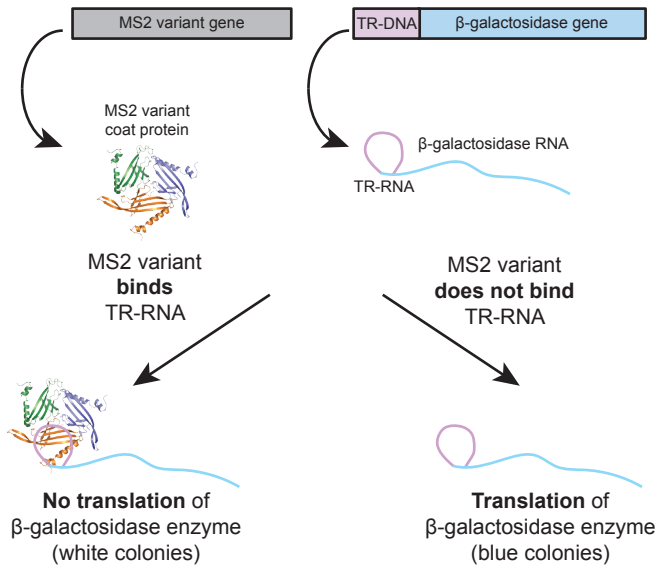


Figure 1.2. Screen for MS2 CP–RNA binding. MS2 variant and β-galactosidase genes are transformed into *E. coli*. If MS2 variant binds the TR-RNA, translation of β-galactosidase is blocked, resulting in white colonies. If the variant does not bind the TR-RNA, then translation of β-galactosidase proceeds, resulting in blue colonies.

bound to the transcribed TR-RNA, translation was inhibited, and levels of β-galactosidase reduced. Thus, β-galactosidase expression levels—which could be assayed with blue/white colony screening—correlated with the strength and persistence of protein–RNA interactions (**Figure 1.2**). This screen allowed labs to screen for RNA binding efficiency across a fairly wide array of coat protein mutants.

One of the earliest studies of MS2 using protein engineering techniques was published in 1989[36]. The MS2 coat protein contains two cysteines, and both were thought to be important for the coat protein to bind RNA. David Peabody tested nine substitutions at C46 and eleven at C101, then quantified the effect of each mutant on RNA binding using the two-plasmid system described above. Blue/white screening, followed by Sanger sequencing, was used to evaluate RNA binding. Interestingly, neither cysteine was required for binding to the TR-RNA, and many of the single amino acid mutants (and one double mutant) preserved the translation repressor property of the wild-type coat protein. It should be noted that these mutants were not assayed for VLP formation. Wild-type like protein expression was confirmed for 17 of the 20 variants. A follow-up study showed that while C46S and C101S both bound the TR-RNA, repressing β-galactosidase expression, neither serine mutant successfully formed VLPs[38].

A flurry of studies soon followed, further characterizing the nature of the protein–RNA interaction using the two-plasmid system. These studies largely focused on altering or challenging the specificity and affinity of the RNA binding pocket[39–41]. Of particular note, the Peabody lab had found that assembly properties of a variant could confound their efforts to probe the RNA–protein binding interaction, as assembly-deficient variants were more available to bind nucleic acid than assembled variants[41]. Deleting a region of the FG loop resulted in assembly-deficient variants that could still bind RNA, and variants bearing this deletion could then be used to probe nucleic acid binding more accurately.

In 1996, Peabody began fusing the MS2 coat protein monomers into a single chain dimer, allowing researchers to probe the function of two equivalent RNA binding sites separately across an axis of symmetry[42]. In 1999, this technique was used to probe mutations of the threonine at position 45, which is known to bind RNA but is not mutable in the MS2 coat protein monomer[43]. Interestingly, this fusion strategy, in which the assembled VLP contains only 90 mutations instead of the 180 mutations without the fusion, corrected

the folding and assembly defects of mutations at position 45 and allowing an in-depth analysis of the importance of this position in RNA binding. Additional mutagenesis and repressor analysis suggested that T45 likely interacted with the TR-RNA at a looped adenine (A–4) rather than a bulged adenine (A–10), adding more nuance to the coat protein–RNA interactions.

Around the same time, Peter Stockley and collaborators began using crystallography, transmission electron microscopy, and other techniques to study the effect of mutations on the coat protein. In the early 1990s, Stockley evaluated a number of variants to determine their effect on several properties, including RNA binding[20,44], melting temperature[45], and structure[46]. These studies began to reveal a high tolerance to mutation, often without altering the melting temperature of the parent VLP[45]. Crystallography further showed how the CP structure could compensate for mutation. For example, Stockley crystallized CP[P78N], which replaces a highly conserved proline that undergoes a *cis–trans* isomerization during assembly with an asparagine residue. The mutated VLP was stable up to 66 ºC (a 2 degree decrease from wild type)[45]. Crystallography showed significant conformational changes in the FG loop, which contains the P78 residue[46], including a normal *trans* bond at P78N. These studies showed the diverse ways that the MS2 VLP could compensate for mutation, even at a critical position like P78.

Up to the early 2000s, most studies on MS2 relied on rational mutagenesis informed by VLP structure. In 2001, the Peabody lab applied random mutagenesis to the MS2 coat protein[47]. Mutagenic PCR was coupled with a medium-throughput assembly assay to screen mutants for VLP formation. Bacteria expressing mutated capsids were transformed on plates. A layer of agarose containing lysis buffer was overlaid and allowed to cool on the plate. Overnight, cells lysed, and protein was allowed to diffuse. Because a 28 kDa dimer has different diffusive properties compared to an assembled VLP, assembly-defective variants diffused more rapidly than well-assembled VLPs. A final step transferred protein onto a nitrocellulose membrane, and the radius of diffusion for each variant was probed with anti-MS2 antibodies. This method allowed screening of tens to hundreds of variants for assembly in parallel. One variant that was incapable of forming assembled capsids was further mutated to see if additional substitutions could rescue assembly—an early study of epistasis in the MS2 CP. A few hundred colonies were examined, but no complete rescue was found.

Throughout these efforts, variants with altered physical properties were reported. For example, Stockley found that CP[E76D] did not disassemble under strongly acidic conditions that dissociate the wild-type MS2 CP into dimers. Further study, including crystallography, suggested that a hydrogen bond between E76 and the backbone of T71 could be important for the acid stability of the VLP[37], a finding that is confirmed in this work (see **Chapter 2** and **Chapter 3**).

In 2003, the Peabody lab found a variant that was more stable to urea and, interestingly, pressure than wild-type MS2 (CP[D11N]). Residue D11 appears a few positions before the AB loop, which extends in a loop away from the VLP. The reason for the increased stability of the variant is not known, though it is possible that asparagine alters the local hydrogen bonding network, conferring additional stability to these stressors. Several variants were less stable to urea and high pressure (CP[M88V] and CP[T45S])[48], though

these are perhaps less useful than CP[D11N].

The Stockley lab rationally introduced disulfide bonds into the FG loop, mimicking the disulfide bonds found in Qβ, which is known to confer unusually high melting temperature (~85 ºC)[49]. Interestingly, Stockley found that the melting temperature of MS2 increases from 68 ºC to 73 ºC following introduction of this presumed disulfide bond at the five-fold axis. While a minor improvement in thermal stability, this effort showed that VLP properties have the potential to be rationally tuned.

Recently, the Tullman-Ercek group identified a single amino acid variant of the MS2 CP that confers a stable shift in quaternary geometry of the VLP[50]. This variant, CP[S37P], changes the diameter of the VLP from 27 nm to 17 nm and altered the icosahedral symmetry from T = 3 (180 monomers) to T = 1 (60 monomers). This study used error-prone PCR in combination with a size selection to enrich for well-formed virus-like particles. More information on this study can be found in **Chapter 5**.

From these studies, we can see that both the physical and the chemical properties of the MS2 VLP can be altered in exciting and surprising ways with amino acid substitutions, providing strong motivation for the comprehensive and systematic mutagenesis efforts described in this thesis.

1.B.iii. Protein engineering to display foreign epitopes on the MS2 VLP

Since the early 1990s, the MS2 VLP has also been used as a surface display platform for foreign peptides. In 1993, Stockley first inserted peptides into the MS2 CP at the AB loop, which protrudes out from the VLP[51], though the insertion appeared to require significant optimization of expression and purification conditions. Shortly thereafter, Peabody found that peptide insertions this position were better tolerated when the monomer was fused into a single chain dimer[52]. This technique was used previously to study mutations at residue 45[43], a position that is not tolerant to mutation in the absence of the single chain dimer fusion. Peabody tested the assembly of the VLP with a synthetic peptide FLAG (DYKDDDDK) inserted at the AB loop and N terminus in the monomer and single chain dimer fusion. The study reported that the peptide was best tolerated at the AB loop of the coat protein in the second copy of the single chain dimer fusion. This technique allowed 90 copies of the FLAG peptide to be displayed while still preserving assembly.

A follow-up study more rigorously evaluated the diversity of 6-, 8- and 10-mer peptides that could be inserted into the AB loop in the single chain dimer. Surprisingly, a fairly wide set of insertions were tolerated[53], though the thermal stability of the CP is compromised with some insertions[54]. It is also interesting to note that the symmetry of the viral particle changes with the dimer fusion, and particles made from the single chain dimer fusion are octahedral rather than icosahedral[55].

In more recent years, many of these peptide insertion studies rely on the ability of the MS2 CP to internalize and protect its own nucleic acid—the same property that is used throughout this thesis to study capsid assembly and mutability. For example, in 2017, a library of all possible overlapping 10-mer peptides from the dengue virus (DENV-3) polyprotein was inserted at the AB loop in the single chain dimer (90 copies)[56]. These displayed peptides were then used to determine which segments of the DENV-3 polyprotein were recognized by IgGs from patients previously infected with dengue virus.

IgGs were incubated with MS2 bearing these peptide insertions, then enriched via affinity chromatography. RNA from MS2 displaying peptides that bound to the IgGs could be isolated and sequenced, providing information on which peptides were immunogenic and stimulating antibody production in patients. Several interesting epitopes were identified across multiple patients, including several that were highly conserved but buried in the assembled virion. This same technique, called "biopanning", was used to evaluate antibody responses in ovarian cancer patients[57]. From these studies, MS2 VLPs have emerged as a powerful alternative to traditional phage display, which uses non-icosahedral, infectious bacteriophages such as M13.

In addition to biopanning, peptide display on MS2 can be used to identify promising vaccine candidates[58]. The highly repetitive structure of viral capsids is known to stimulate an immune response, and the display of immunogenic peptides on the surface further enhances that effect[59–62]. Perhaps the most promising vaccine candidate is for HPV, which has proceeded through mouse models, as well as dose, adjuvant, and stability studies[60,61,63–65]. Taken together, these studies have developed the MS2 particle as a surface display platform and promising vaccination strategy. They also provide excellent context for our work to extend the N terminus of the MS2 bacteriophage in the absence of the single chain dimer fusion, both to identify permitted extensions and to enhance the chemical reactivity and surface display capabilities of the particles (see **Chapter 4**).

1.B.iv. Summary of the MS2 VLP

In the past 30 years, protein engineering has allowed researchers to learn a great deal about capsid assembly and utility. These studies also show that even one amino acid variants can dramatically change the chemical and physical properties of the MS2 VLP. This thesis expands on these efforts by developing a new technique to probe the mutability of the coat protein itself by generating and evaluating many mutations in a single experiment. These efforts have expanded our understanding of mutability and self-assembly and identified variants of the MS2 CP with highly tuned physical and chemical properties.

1.C. Protein fitness landscapes

In this work, we sought to study VLPs derived from the MS2 CP using protein fitness landscapes. Protein fitness landscapes measure the relationship between a mutation and its phenotype, or fitness. With the MS2 CP, fitness can be assembly, stability, or other properties. By using this technique, we are able to generate a more comprehensive understanding of how mutations affect VLP physical and chemical properties and shape available evolutionary paths. In the following sections, we describe fitness landscapes in more detail, including numerous ways that this technique can be used.

1.C.i. Introduction to protein fitness landscapes

Protein fitness landscapes are mapped by generating comprehensive sets of variants, then evaluating how those variations affect a measurable phenotypic read-out (**Figure 1.3**). This read-out is typically not a direct measurement of protein folding or function, but rather some proxy, such as growth[66–71] or infectivity[72–78], though more direct

measurements of fluorescence[79] and self-assembly[50,80–82] have also been used. High-throughput sequencing before and after the selection is used to identify mutants that are enriched following the selection[83]. To date, all protein fitness landscapes measured via deep mutational scanning[83] rely on a genotype-to-phenotype link and high-throughput sequencing. Fitness landscapes can be used to study evolution, identify proteins with new and useful properties, or quantify mutability.

Though protein fitness landscapes are described as an unbiased way to study mutability, mutability landscapes are frequently quite different from mutability maps generated from phylogenetic, evolutionary, or even *in vitro* simulated evolutionary analyses. In fact, protein fitness landscapes are particularly useful because they operate outside of the paradigm of evolution, allowing identification of proteins and functionalities that may not be accessible by other techniques. In essence, fitness landscapes can identify proteins with new and useful properties, even with minimal changes to the genetic code. Other techniques, such as *de novo* protein design[3], are also emerging ways to circumvent traditional evolutionary constraints.

Protein fitness landscapes are just one of many synthetic biology techniques that are useful to study the shape and dynamics of protein evolution and mutability. These include phage-assisted continuous evolution (PACE), which speeds up directed evolution by linking protein evolution to phage survival[84–86]; eVOLVER, which enables high-throughput, parallel evaluation of a wide array of selection conditions[87]; and many others[88,89]. Several recent reviews have thoroughly described how protein fitness landscapes can be used to study evolution[90,91], elucidating the shape of mutational space and the frequency of epistasis, or coevolution. These efforts seek to assess questions about protein fitness

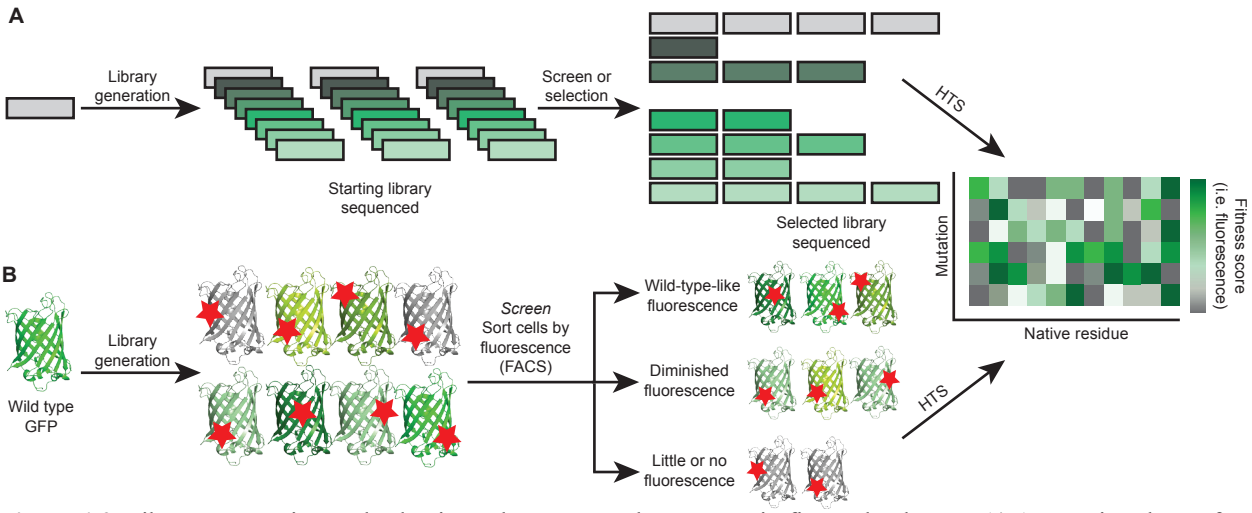


Figure 1.3. Library generation and selection scheme to produce a protein fitness landscape. A) A generic scheme for mapping a protein fitness landscape is shown. A comprehensive set of mutations is generated, then a screen or selection enriches for a phenotype of interest. High-throughput sequencing is used to quantify variant abundance in the starting and selected libraries, and differences in percent abundance are used to quantify the fitness of every mutation. B) An example scheme detailing how to map the protein fitness landscape of GFP is shown. All possible one amino acid mutations of GFP are generated and expressed in bulk. Cells are sorted using fluorescence activated cell sorting (FACS), then pools of each fluorescence level are sequenced. The relative abundance of each variant in each pool is used to quantify the mutability of each residue across the GFP backbone, giving rise to a protein fitness landscape.

experimentally, complementing years of computational and theoretical work[92–94].

Interestingly, many differences exist between mutability quantified by evolutionary studies and deep scanning mutagenesis. There are likely several contributing factors to this difference. One obvious reason why protein fitness landscapes are different from evolutionary landscapes is due to the structure of the genetic code. On average, single nucleotide mutations can access ~6 other amino acids, a marked drop from comprehensive codon mutagenesis, which can access all 20 amino acids[95]. This difference has several consequences. First, evolutionary paths become longer to access two and three base pair mutations and therefore non-adjacent amino acids[96]. In addition, the number and type of amino acid substitutions are also limited, as not all physical properties are accessible by single nucleotide substitutions. In several studies, mutability quantified via error prone PCR appeared to correlate better with phylogenetic studies[79,97] than comprehensive codon mutagenesis studies[68,70,81,96]. Indeed, it is intuitive to argue that error prone PCR is a better mimic of evolution than deep scanning mutagenesis, as evolution and error prone PCR both largely proceed via single base pair mutations and are generally constrained by the structure of the genetic code.

However, a survey of the literature suggests that several other factors contribute to these differences in mutability, including epistasis, binding partners, and selection conditions. In the following section, we explore these differences in more detail, discuss what we can learn about evolution and protein function, and describe why protein fitness landscapes are particularly well-suited as a tool to study the mutability and self-assembly of VLPs.

1.C.ii. Learning about epistasis from protein fitness landscapes

In nature, replacing a native residue must be beneficial to be fixed. Neutral mutations are less likely to be fixed than beneficial mutations, and deleterious mutations are not retained[98]. This is one of the primary constraints of evolution, and this requirement profoundly shapes available mutational paths. In addition, the behavior conferred by a mutation, and how likely it is to be fixed, will change based on previous and subsequent mutations—an effect known as epistasis.

In 2006, Weinreich *et al.* published a pivotal study that evaluated the available evolutionary paths between wild-type β-lactamase and a clinically-relevant, high-resistance variant of β-lactamase that increased antibiotic resistance by a factor of 100,000[98]. The researchers asked how the mutational space between these two important variants could actually be traversed.

Weinreich *et al.* calculated the probability of each of the 120 mutational trajectories linking the two alleles. Nearly all of these trajectories are statistically unlikely under the assumption that a mutation must be beneficial to be fixed. For example, one mutation, G238S, enhanced hydrolysis (a benefit to antibiotic resistance) but increases protein aggregation (a detriment). A second mutation, M182T, reduced hydrolyses (a detriment) but also increases stability (a benefit). While both mutations carry benefits and detriments, M182T only provided a net benefit in the G238S genetic background by the detrimental aggregation caused by G238S. M182T in the wild-type genetic background would not be retained. Thus, a particular order was required to achieve the beneficial combination of

M182T / G238S.

In this study, the researchers found that only a few of the 120 possible paths were statistically likely. This led to the now much-repeated conclusion that “it now appears that intramolecular interactions render many mutational trajectories selectively inaccessible, which implies that replaying the protein tape of life might be surprisingly repetitive”[98].

Since this study, a number of researchers have calculated the prevalence of epistasis in two amino acid mutational space using protein fitness landscapes[79,81,94,99,100]. Most of these studies sample the two-residue mutational space or map a complete domain or region of a protein, as complete epistatic networks quickly become a combinatorial challenge. In contrast to evolutionary studies, these studies can probe mutational spaces that extend beyond local minima and maxima. By employing a variety of mutagenesis strategies, even combinations with no viable evolutionary paths can be generated, studied, and evaluated for their potential utility. Indeed, in these studies, forcing a second mutation, even if the first is detrimental, can reveal the hidden effects of epistasis, giving additional shape to protein fitness landscapes.

In this work, we use protein fitness landscapes to study the effect of epistasis on MS2 VLPs. We first map the complete one dimensional landscape of the MS2 coat protein (**Chapter 2**), describing how single amino acid mutations affect self-assembly. In **Chapter 3**, we build on these results by comparing how mutations in a highly mutable loop in MS2 are impacted by a second mutation. We find epistatic effects in nearly every combination of residues in this six amino acid loop, and charged residues are particularly tightly regulated in this loop.

In general, most studies of epistasis find significant positive and negative epistasis, identifying combinations of mutations that are unexpectedly beneficial or detrimental to protein fitness. In addition, a surprising number of viable two amino acid mutations arise from cases in which both individual mutations are not permitted—a seemingly inaccessible variant.

A few studies of higher order epistasis—or the coevolution of up to four mutations—have been carried out as well[96,101]. A high prevalence of epistasis was identified, as in the studies with fewer mutations, yet these studies also found that few functional mutations were completely inaccessible through evolution. In fact, in one study, a higher percentage of two amino acid mutations were inaccessible than four amino acid mutations, meaning that round-about mutational paths may exist that take advantage of the full intragenic epistasis available in nature. Thus, while we can see a significant number of unattainable paths via controlled epistasis studies, several studies of higher epistasis indicate that evolution may be able to avoid local minima in interesting and surprising ways[96,101].

As an example, in 2015, Podgornaia and Laub mapped the complete mutational landscape of four residues in a protein kinase PhoQ[96]. These four residues govern recognition of the substrate, PhoP. All four positions were altered to be encoded by NNS codons, and all combinations were evaluated for viability. Of 160,000 possible combinations, only 1,659, or about 1%, were identified as functional, comprised of 16 single mutants, 100 double, 544 triple, and 988 quadruple mutants. In addition to finding many instances of epistasis, the predictive power of the single mutational landscape was surprisingly poor. Of the 490 mutations that were expected to be functional in the absence

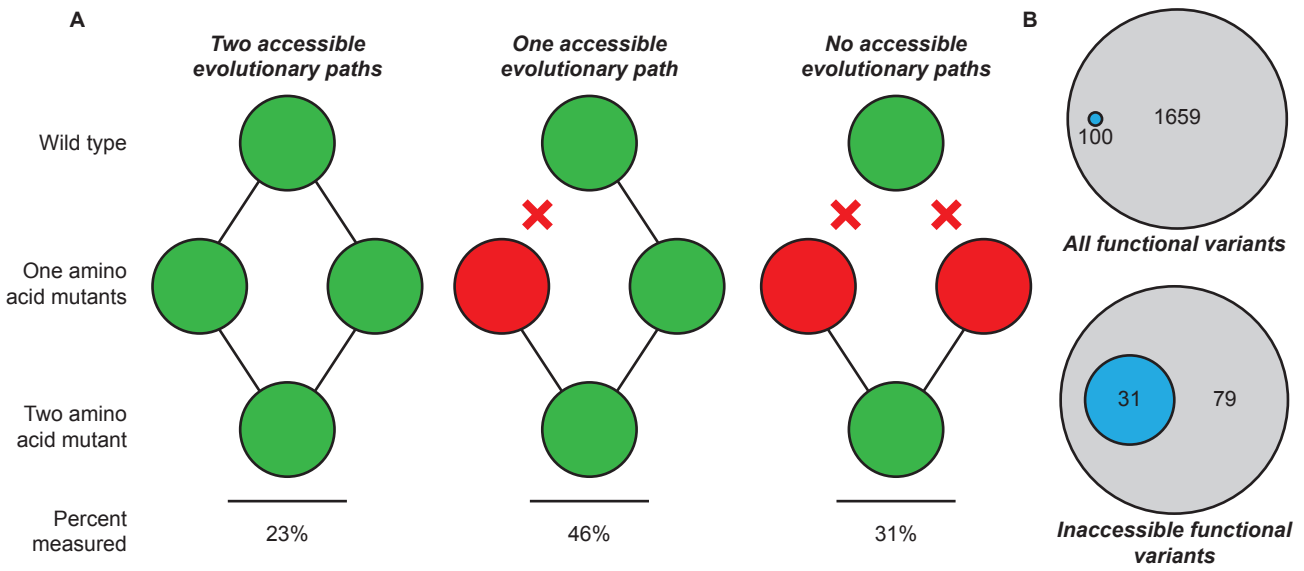


Figure 1.4. Evolutionary paths of functional two amino acid mutations of PhoQ. A) Functional mutations are indicated in green, while nonfunctional mutations are indicated in red. Viable evolutionary paths are shown with a black line, while nonviable paths are shown with a red X. In total, 46% of functional PhoQ double mutants were comprised of one functional and one nonfunctional single mutants, 31% were comprised of two nonfunctional single mutants, and 23% were comprised of two functional single mutants. B) Two amino acid mutants represented 100 of 1659 functional variants but 31 of 79 functional variants with no evolutionary path. Two amino acid variants are indicated in blue, and all variants are indicated in gray. Adapted from Podgornaia *et al*.

of any epistatic effects, only 104 of these were functional, and an additional 1,555 were viable.

Importantly, because all possible 1, 2, 3, and 4 amino acid mutations were mapped, the researchers could visualize which paths were accessible (or inaccessible) to the protein via evolution. They found that nearly 40% of the functional variants with no viable evolutionary path were two amino acid mutations, representing 31% of the functional two amino acid mutants (**Figure 1.4A**). However, overall, only 5% of the functional variants were inaccessible without passing through a nonfunctional intermediate, or 79 of 1,659 overall (**Figure 1.4B**). Had this study only evaluated two amino acid mutations, the percent of inaccessible variants would have appeared far higher. This result supports the hypothesis that nature takes advantage of higher order epistasis to access functional variants—though these effects are more challenging to study using existing protein engineering techniques.

1.C.iii. Evaluating how selection conditions alter protein fitness landscapes

Changing environments can also alter which mutations are beneficial in a particular context. For example, a mutation that is beneficial when bacteria are challenged by an antibiotic—but detrimental in most other circumstances—can result in evolutionary paths that achieve seemly inaccessible variants. This is also true in lab-based selections: selection conditions can dramatically alter the mutability map of a given protein. For example, in growth-based selections for antibiotic resistance, enzyme expression level and amount of antibiotic can impact how much growth is affected by a change in enzyme

efficiency[90]. Clever selection schemes, such as the one employed by the Ostermeier lab, in which variants are isolated and challenged to 13 different concentrations of antibiotic, can achieve more comprehensive evaluation of fitness compared to traditional bulk growth experiments[66]. Even so, selection conditions can still affect protein fitness landscapes.

However, the changing mutability of a protein can also be considered an advantage of protein fitness landscape studies: researchers can study how protein mutability changes under a variety of conditions. For example, in 2017, Wrenbeck *et al.* evaluated three different substrates for an amidase enzyme[102]. The three different substrates were relatively similar, though the largest had a branched methyl unit (**Figure 1.5A**). Surprisingly, only 26 of the 598 beneficial mutations were beneficial for all three substrates (**Figure 1.5B**). Beneficial mutations were also not constrained to the active site, but rather appeared throughout the protein at distances between 6 and 27 angstroms.

Because so few mutations were beneficial across multiple substrates, the researchers concluded that mutations are likely affecting enzyme specificity or activity on a particular substrate, rather than global enzyme expression or stability. In essence, the substrate completely changed the mutability map of the enzyme, and a beneficial mutation on one substrate had little correlation to its effect on other substrates. Taken together, this protein fitness landscape study suggests that across evolution, even slight changes to substrate pools may dramatically alter which mutations are viable. Furthermore, this implies that evolutionary paths are likely more dynamic than static and can change depending on environmental conditions.

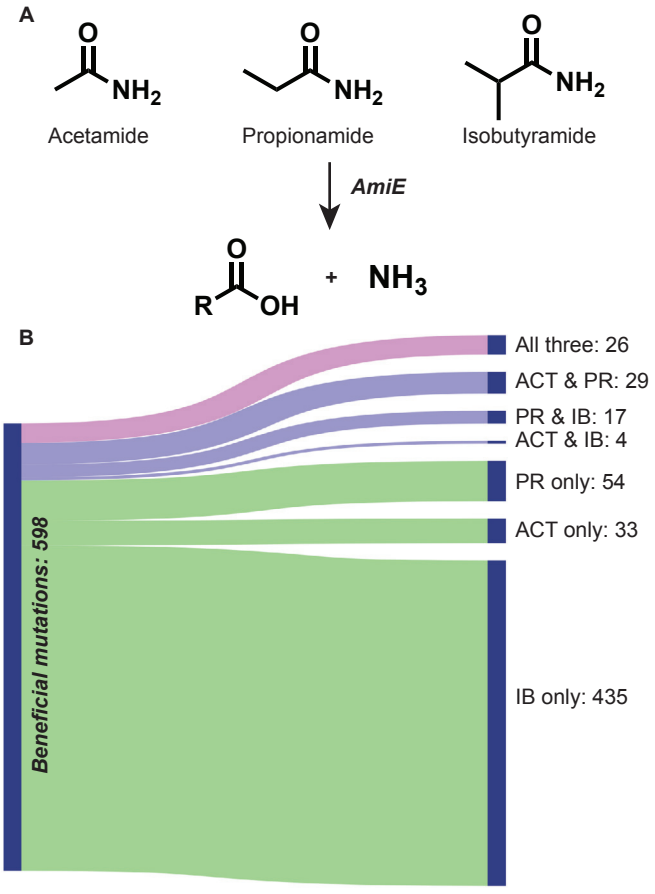


Figure 1.5. Dependent mutability of AmiE based on available substrate. A) AmiE produces ammonia from acetamide (ACT), propionamide (PR), or isobutyramide (IB). B) Of 598 beneficial mutations, only 26 are beneficial to AmiE with all three substrates. Adapted from Wrenbeck *et al* and created using SankeyMATIC.

Throughout our work, we find that changing selection conditions alter which variants are permitted in a given fitness landscapes of the MS2 CP. More importantly, when we apply direct functional selections, such as acidic, thermal, or chemical modification stress, we are able to identify variants with highly tuned chemical and physical properties. As such, the ability to change selection conditions and study how different variants react to those conditions is a significant advantage of protein fitness

landscapes—an advantage that is highlighted throughout this work.

1.C.iv. The role of binding partners in shaping protein fitness landscapes

In addition to small molecules, proteins interact with many biological binding partners in their native context, including other proteins, nucleic acids, or small molecules. For example, several proteins with mapped fitness landscapes have many critical cellular binding partners. Hsp90 has been found to interact with around 3% of the proteome in high-throughput sequencing experiments[70], and non-covalent binding partners are necessary for ubiquitin to function in degradation pathways[68]. These binding partners can constrain mutation throughout evolution, even if the same mutations are permitted in *in vitro* evolutionary studies, including protein fitness landscapes. In addition, the evolution of the binding partner itself can dictate how mutational constraints change over time.

Ubiquitin function is critical for eukaryotes and is highly conserved across many species. Only three residues vary between yeast and human ubiquitin[67]. However, in *in vitro*, growth-based mutability studies, ubiquitin is observed to be far more mutable than is suggested by evolutionary studies[67,68]. Because of prior biophysical, biochemical, and structural studies of ubiquitin, the researchers mapping the protein fitness landscape were able to interpret mutability in interesting ways. Ten positions that were least tolerant to mutation were on the protein surface, at or adjacent to places where covalent attachment is mediated.

In a follow-up study, the same protein fitness landscape was mapped for ubiquitin, but under different chemical stressors[67]. Caffeine, DTT, hydroxyurea, and MG132 (a proteasome inhibitor) were all used to alter the environment in which ubiquitin functioned. These chemical stressors changed the mutability landscape of ubiquitin, primarily sensitizing new positions to mutation. Some of these sensitized sites were shared across multiple stressors, while others were specific to individual chemical environments. Among other explanations, changes to critical binding partners are a central hypothesis for the altered mutability. Presumably across a wide variety of environments in evolution, the many binding partners and functions of a protein like ubiquitin constrain its mutability far more than is seen in lab-based protein fitness landscape studies.

In our work, the self-assembly of the MS2 coat protein was isolated and studied, independent from its role in infectivity. We expect that a similar study with an infectivity selection, which requires protein-genomic RNA packaging and incorporation of the maturation protein, Protein A, among other diverse functions, would lead to a far different picture of protein mutability. In this case, we would expect infectivity to be a more restrictive selection, decreasing the number of permitted mutations. For example, by using lab-based functional selections, we identified several variants with properties that are likely detrimental for infectivity—namely, a smaller capsid size with altered quaternary geometry (**Chapter 5**, **Figure 1.6**)[50]. Because of the small interior diameter, the native genome does not fit in the smaller virus-like particle, meaning that the variant would likely be impossible to access using phylogenetic or evolutionary analyses. Thus, selections in which a protein is isolated from some or all of its native binding partners are useful tools to identify new variants with functions that may not be accessible by other methods.

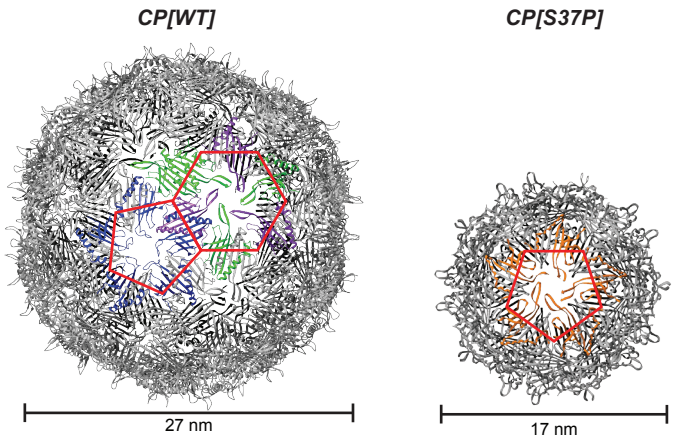


Figure 1.6. A one amino acid variant of the MS2 bacteriophage coat protein. CP[S37P] forms a smaller capsid with altered quaternary geometry. The 17 nm capsid is too small to fit the native RNA genome of the MS2 bacteriophage, making it nearly impossible to identify via evolutionary studies.

1.C.v. Conclusions from protein fitness landscapes.

Protein fitness landscapes are uniquely well-suited for comprehensive, high-resolution mapping of mutability. These studies are not constrained by the structure of the genetic code, can identify variants that are well-folded even if there is no obvious evolutionary path, and can track reactions to binding partners or changing selection conditions. Protein fitness landscapes also provide unique insight into the prevalence of epistasis under different selection conditions. Furthermore, by isolating proteins from their native binding partners and applying well-chosen direct functional selections, variants with highly useful changes to their physical and chemical properties can be identified, even if those variants are inaccessible by evolution. Taken together, fitness landscapes operate outside of the traditional evolutionary paradigm, making them an excellent tool to study VLP assembly and mutability.

1.D. Overview and Outlook

In this thesis, we use protein fitness landscapes to study the self-assembly and mutability of virus-like particles. To do so, we developed a new method, called Systematic Mutagenesis and Assembled Particle Selection (SyMAPS), which combines highly targeted libraries with direct functional selections. High-throughput sequencing is used as a read-out for self-assembly, allowing the assembly state of many mutants to be quantified all in one pool. Using this technique, we quantified the first protein fitness landscape of a native viral coat protein (**Chapter 2**); evaluated the effect of epistasis on a highly mutable loop (**Chapter 3**); altered the chemical reactivity of coat protein (**Chapter 4**); and sought variants with altered quaternary structures (**Chapter 5**). With each new dataset and project, we add to our understanding of how mutations affect the physical properties and self-assembly of a native viral coat protein.

1.E. References

1. Farkas, M. E., Aanei, I. L., Behrens, C. R., Tong, G. J., Murphy, S. T., Neil, J. P. O. & Francis, M. B. PET Imaging and Biodistribution of Chemically Modified Bacteriophage MS2. *Mol. Pharm.* **10**, 69–76 (2013).

2. Qazi, S., Miettinen, H. M., Wilkinson, R. A., McCoy, K., Douglas, T. & Wiedenheft, B. Programmed Self-Assembly of an Active P22-Cas9 Nanocarrier System. *Mol. Pharm*. **13**, 1191–1196 (2016).

3. Butterfield, G. L., Lajoie, M. J., Gustafson, H. H., Sellers, D. L., Nattermann, U., Ellis, D., Bale, J. B.,

Ke, S., Lenz, G. H., Yehdego, A., Ravichandran, R., Pun, S. H., King, N. P. & Baker, D. Evolution of a designed protein assembly encapsulating its own RNA genome. *Nature* **552**, 415–420 (2017).

4. ElSohly, A. M., Netirojjanakul, C., Aanei, I. L., Jager, A., Bendall, S. C., Farkas, M. E., Nolan, G. P. & Francis, M. B. Synthetically Modified Viral Capsids as Versatile Carriers for Use in Antibody-Based Cell Targeting. *Bioconjug. Chem.* **26**, 1590–1596 (2015).
5. Aanei, I. L., Elsohly, A. M., Farkas, M. E., Netirojjanakul, C., Regan, M., Taylor Murphy, S., O'Neil, J. P., Seo, Y. & Francis, M. B. Biodistribution of antibody-MS2 viral capsid conjugates in breast cancer models. *Mol. Pharm.* **13**, 3764–3772 (2016).
6. Kim, E. Y. & Tullman-Ercek, D. Engineering nanoscale protein compartments for synthetic organelles. *Curr. Opin. Biotechnol.* **24**, 627–32 (2013).
7. Glasgow, J. E., Asensio, M. A., Jakobson, C. M., Francis, M. B. & Tullman-Ercek, D. Influence of Electrostatics on Small Molecule Flux through a Protein Nanoreactor. *ACS Synth. Biol*. **4**, 1011–1019 (2015).
8. Glasgow, J. E., Capehart, S. L., Francis, M. B. & Tullman-Ercek, D. Osmolyte-mediated encapsulation of proteins inside MS2 viral capsids. *ACS Nano* **6**, 8658–8664 (2012).
9. Azuma, Y., Herger, M. & Hilvert, D. Diversification of Protein Cage Structure Using Circularly Permuted Subunits. *J. Am. Chem. Soc.* **140**, 558–561 (2018).
10. Tetter, S. & Hilvert, D. Enzyme Encapsulation by a Ferritin Cage. *Angew. Chemie - Int. Ed.* **56**, 14933–14936 (2017).
11. Dedeo, M. T., Finley, D. T. & Francis, M. B. *Viral capsids as self-assembling templates for new materials. Progress in molecular biology and translational science* **103**, (Elsevier Inc., 2011).
12. Miller, R. a., Presley, A. D. & Francis, M. B. Self-assembling light-harvesting systems from synthetically modified tobacco mosaic virus coat proteins. *J. Am. Chem. Soc.* **129**, 3104–3109 (2007).
13. Hsia, Y., Bale, J. B., Gonen, S., Shi, D., Sheffler, W., Fong, K. K., Nattermann, U., Xu, C., Huang, P.-S., Ravichandran, R., Yi, S., Davis, T. N., Gonen, T., King, N. P. & Baker, D. Design of a hyperstable 60-subunit protein icosahedron. *Nature* **535**, 136–139 (2016).
14. Bale, J. B., Gonen, S., Liu, Y., Sheffler, W., Ellis, D., Thomas, C., Cascio, D., Yeates, T. O., Gonen, T., King, N. P. & Baker, D. Accurate design of megadalton-scale two-component icosahedral protein complexes. *Science* **353**, 389–394 (2016).
15. Glasgow, J. & Tullman-Ercek, D. Production and applications of engineered viral capsids. *Appl. Microbiol. Biotechnol.* **98**, 5847–58 (2014).
16. Davis, J. E., Strauss, J. H. & Sinsheimer, R. Bacteriophage MS2: another RNA phage. *Science* **134**, 1427 (1961).
17. Strauss, J. H. & Sinsheimer, R. L. Purification and properties of bacteriophage MS2 and of its ribonucleic acid. *J. Mol. Biol.* **7**, 43–54 (1963).
18. Koning, R. I., Gomez-Blanco, J., Akopjana, I., Vargas, J., Kazaks, A., Tars, K., Carazo, J. M. & Koster, A. J. Asymmetric cryo-EM reconstruction of phage MS2 reveals genome structure *in situ*. *Nat. Commun.* **7**, 12524 (2016).
19. Fiers, W., Contreras, R., Duerinck, F., Haegeman, G., Iserentant, D., Merregaert, J., Jou, W. M., Molemans, F., Raeymaekers, A., Berghe, A. Van den, Volckaert, G. & Ysebaert, M. Complete nucleotide sequence of bacteriophage MS2 RNA. *Nature* **260**, 500–507 (1976).
20. Stockley, P. G., Stonehouse, N. J. & Valegård, K. Molecular mechanism of RNA phage morphogenesis. *Int. J. Biochem.* **26**, 1249–1260 (1994).
21. Dai, X., Li, Z., Lai, M., Shu, S., Du, Y., Zhou, Z. H. & Sun, R. *In situ* structures of the genome and

genome-delivery apparatus in a single-stranded RNA virus. *Nature* **541**, 112–116 (2016).

22. Mannige, R. V. & Brooks, C. L. Periodic table of virus capsids: Implications for natural selection and design. *PLoS One* **5**, 1–7 (2010).
23. Golmohammadi, R., Valegård, K., Fridborg, K. & Liljas, L. The Refined Structure of Bacteriophage MS2 at 2·8 Å Resolution. *J. Mol. Biol.* **234**, 620–639 (1993).
24. George, L., Indig, F. E., Abdelmohsen, K. & Gorospe, M. Intracellular RNA-tracking methods. *Open Biology* **8**, 180104 (2018).
25. Zhou, Z. & Reed, R. Purification of functional RNA-protein complexes using MS2-MBP. *Curr. Protoc. Mol. Biol.* **Chapter 27**, 27.3.1-27.3.7 (2003).
26. Hung, M. E. & Leonard, J. N. A platform for actively loading cargo RNA to elucidate limiting steps in EV-mediated delivery. *J. Extracell. Vesicles* **5**, (2016).
27. Wu, N. C., Dai, L., Olson, C. A., Lloyd-Smith, J. O. & Sun, R. Adaptation in protein fitness landscapes is facilitated by indirect paths. *Elife* **5**, 1–21 (2016).
28. Wu, W., Hsiao, S. C., Carrico, Z. M. & Francis, M. B. Genome-free viral capsids as multivalent carriers for taxol delivery. *Angew. Chemie - Int. Ed.* **48**, 9493–9497 (2009).
29. Giessen, T. W. & Silver, P. A. A Catalytic Nanoreactor Based on *in Vivo* Encapsulation of Multiple Enzymes in an Engineered Protein Nanocompartment. *ChemBioChem* **17**, 1931–1935 (2016).
30. Carrico, Z. M., Romanini, D. W., Mehl, R. A. & Francis, M. B. Oxidative coupling of peptides to a virus capsid containing unnatural amino acids. *Chem. Commun. (Camb)*. 1205–1207 (2008).
31. Galaway, F. A. & Stockley, P. G. MS2 viruslike particles: A robust, semisynthetic targeted drug delivery platform. *Mol. Pharm.* **10**, 59–68 (2013).
32. Ashley, C. E., Carnes, E. C., Phillips, G. K., Durfee, P. N., Buley, M. D., Lino, C. A., Padilla, D. P., Phillips, B., Carter, M. B., Willman, C. L., Brinker, C. J., Caldeira, J. D. C., Chackerian, B., Wharton, W. & Peabody, D. S. Cell-specific delivery of diverse cargos by bacteriophage MS2 virus-like particles. *ACS Nano* **5**, 5729–5745 (2011).
33. Stephanopoulos, N., Carrico, Z. M. & Francis, M. B. Nanoscale integration of sensitizing chromophores and porphyrins with bacteriophage MS2. *Angew. Chemie - Int. Ed.* **48**, 9498–9502 (2009).
34. Capehart, S. L., ElSohly, A. M., Obermeyer, A. C. & Francis, M. B. Bioconjugation of Gold Nanoparticles through the Oxidative Coupling of ortho -Aminophenols and Anilines. *Bioconjug. Chem.* **25**, 1888–1892 (2014).
35. Remaut, E., Waele, P. De, Marmenout, A., Stanssens, P. & Fiers, W. Functional expression of individual plasmid-coded RNA bacteriophage MS2 genes. *EMBO J.* **1**, 205–209 (1982).
36. Peabody, D. S. Translational repression by bacteriophage MS2 coat protein does not require cysteine residues. *Nucleic Acids Res.* **17**, 6017–6027 (1989).
37. Lago, H., Fonseca, S. A., Murray, J. B., Stonehouse, N. J. & Stockley, P. G. Dissecting the key recognition features of the MS2 bacteriophage translational repression complex. *Nucleic Acids Res.* **26**, 1337–1344 (1998).
38. Peabody, D. S. Translational Repression by Bacteriophage MS2 Coat Protein Expressed from a Plasmid. *J. Biol. Chem.* **265**, 5684–5689 (1990).
39. Lim, F., Spingola, M. & Peabody, D. S. Altering the RNA Binding Specificity of a Translational Repressor. *J. Biol. Chem*. 9006–9010 (1994).
40. Spingola, M. & Peabody, D. S. MS2 coat protein mutants which bind Qβ RNA. *Nucleic Acids Res.* **25**, 2808–2815 (1997).

41. Lim, F. & Peabody, D. S. Mutations that increase the affinity of a translational repressor for RNA. *Nucleic Acids Res.* **22**, 3748–3752 (1994).

42. Peabody, D. S. & Lim, F. Complementation of RNA binding site mutations in MS2 coat protein heterodimers. *Nucleic Acids Res.* **24**, 2352–2359 (1996).

43. Peabody, D. S. & Chakerian, A. Asymmetric contributions to RNA binding by the Thr45 residues of the MS2 coat protein dimer. *J. Biol. Chem.* **274**, 25403–25410 (1999).

44. Stockley, P. G., Stonehouse, N. J., Walton, C., Walters, D. A., Medina, G., Macedo, J. M. B., Hill, H. R., Goodman, S. T. S., Talbot, S. J., Tewary, H. K., Golmohammadi, R., Liljas, L. & Valegård, K. Molecular mechanism of RNA-phage morphogenesis. *Biochem. Soc. Trans.* **21**, 627–634 (1993).

45. Stonehouse, N. J. & Stockley, P. G. Effects of amino acid substitution on the thermal stability of MS2 capsids lacking genomic RNA. *FEBS Lett.* **334**, 355–359 (1993).

46. Stonehouse, N. J., Valegård, K., Golmohammadi, R., van den Worm, S., Walton, C., Stockley, P. G. & Liljas, L. Crystal structures of MS2 capsids with mutations in the subunit FG loop. *J. Mol. Biol.* **256**, 330–9 (1996).

47. Peabody, D. S. Isolation of viral coat protein mutants with altered assembly and aggregation properties. *Nucleic Acids Res.* **29**, 113e–113 (2001).

48. Lima, S. M. B., Peabody, D. S., Silva, J. L. & De Oliveira, A. C. Mutations in the hydrophobic core and in the protein-RNA interface affect the packing and stability of icosahedral viruses. *Eur. J. Biochem.* **271**, 135–145 (2003).

49. Ashcroft, A. E., Lago, H., Macedo, J. M. B., Horn, W. T., Stonehouse, N. J. & Stockley, P. G. Engineering thermal stability in RNA phage capsids via disulphide bonds. *J. Nanosci. Nanotechnol.* **5**, 2034–41 (2005).

50. Asensio, M. A., Morella, N. M., Jakobson, C. M., Hartman, E. C., Glasgow, J. E., Sankaran, B., Zwart, P. H. & Tullman-Ercek, D. A Selection for Assembly Reveals That a Single Amino Acid Mutant of the Bacteriophage MS2 Coat Protein Forms a Smaller Virus-like Particle. *Nano Lett.* **16**, 5944–5950 (2016).

51. Mastico, R. A., Talbot, S. J. & Stockley, P. G. Multiple presentation of foreign peptides on the surface of an RNA-free spherical bacteriophage capsid. *J. Gen. Virol.* **74**, 541–548 (1993).

52. Peabody, D. S. Subunit fusion confers tolerance to peptide insertions in a virus coat protein. *Arch. Biochem. Biophys.* **347**, 85–92 (1997).

53. Peabody, D. S., Manifold-Wheeler, B., Medford, A., Jordan, S. K., do Carmo Caldeira, J. & Chackerian, B. Immunogenic Display of Diverse Peptides on Virus-like Particles of RNA Phage MS2. *J. Mol. Biol.* **380**, 252–263 (2008).

54. Caldeira, J. C. & Peabody, D. S. Thermal stability of RNA phage virus-like particles displaying foreign peptides. *J. Nanobiotechnology* **9**, 22 (2011).

55. Plevka, P., Tars, K. & Liljas, L. Crystal packing of a bacteriophage MS2 coat protein mutant corresponds to octahedral particles. *Protein Sci.* **17**, 1731–1739 (2008).

56. Frietze, K. M., Pascale, J. M., Moreno, B., Chackerian, B. & Peabody, D. S. Pathogen-specific deep sequence-coupled biopanning: A method for surveying human antibody responses. *PLoS One* **12**, e0171511 (2017).

57. Frietze, K. M., Roden, R. B. S., Lee, J.-H., Shi, Y., Peabody, D. S. & Chackerian, B. Identification of Anti-CA125 Antibody Responses in Ovarian Cancer Patients by a Novel Deep Sequence-Coupled Biopanning Platform. *Cancer Immunol. Res.* **4**, 157–164 (2016).

58. O'Rourke, J. P., Peabody, D. S. & Chackerian, B. Affinity selection of epitope-based vaccines using a

bacteriophage virus-like particle platform. *Curr. Opin. Virol.* **11**, 76–82 (2015).

59. Ord, R. L., Caldeira, J. C., Rodriguez, M., Noe, A., Chackerian, B., Peabody, D. S., Gutierrez, G. & Lobo, C. a. A malaria vaccine candidate based on an epitope of the Plasmodium falciparum RH5 protein. *Malar. J.* **13**, 326 (2014).

60. Tumban, E., Peabody, J., Tyler, M., Peabody, D. S. & Chackerian, B. VLPs Displaying a Single L2 Epitope Induce Broadly Cross-Neutralizing Antibodies against Human Papillomavirus. *PLoS One* **7**, 1–10 (2012).

61. Tumban, E., Muttil, P., Escobar, C. A. A., Peabody, J., Wafula, D., Peabody, D. S. & Chackerian, B. Preclinical refinements of a broadly protective VLP-based HPV vaccine targeting the minor capsid protein, L2. *Vaccine* **33**, 3346–53 (2015).

62. Crossey, E., Frietze, K., Narum, D. L., Peabody, D. S. & Chackerian, B. Identification of an immunogenic mimic of a conserved epitope on the plasmodium falciparum blood stage antigen ama1 using virus-like particle (VLP) peptide display. *PLoS One* **10**, (2015).

63. Tumban, E., Peabody, J., Peabody, D. S. & Chackerian, B. A universal virus-like particle-based vaccine for human papillomavirus: Longevity of protection and role of endogenous and exogenous adjuvants. *Vaccine* **31**, 4647–4654 (2013).

64. Tyler, M., Tumban, E., Peabody, D. S. & Chackerian, B. The use of hybrid virus-like particles to enhance the immunogenicity of a broadly protective HPV vaccine. *Biotechnol. Bioeng.* **111**, 2398–2406 (2014).

65. Tyler, M., Tumban, E., Dziduszko, A., Ozbun, M. A., Peabody, D. S. & Chackerian, B. Immunization with a consensus epitope from human papillomavirus L2 induces antibodies that are broadly neutralizing. *Vaccine* **32**, 4267–4274 (2014).

66. Firnberg, E., Labonte, J. W., Gray, J. J. & Ostermeier, M. A comprehensive, high-resolution map of a Gene's fitness landscape. *Mol. Biol. Evol.* **31**, 1581–1592 (2014).

67. Mavor, D., Barlow, K., Thompson, S., Barad, B. A., Bonny, A. R., Cario, C. L., Gaskins, G., Liu, Z., Deming, L., Axen, S. D., Caceres, E., Chen, W., Cuesta, A., Gate, R. E., Green, E. M., Hulce, K. R., Ji, W., Kenner, L. R., Mensa, B., *et al*. Determination of ubiquitin fitness landscapes under different chemical stresses in a classroom setting. *Elife* **5**, (2016).

68. Roscoe, B. P., Thayer, K. M., Zeldovich, K. B., Fushman, D. & Bolon, D. N. A. Analyses of the effects of all ubiquitin point mutants on yeast growth rate. *J. Mol. Biol.* **425**, 1363–1377 (2013).

69. Hietpas, R. T., Jensen, J. D. & Bolon, D. N. A. Experimental illumination of a fitness landscape. *Proc. Natl. Acad. Sci. U. S. A.* **108**, 7896–7901 (2011).

70. Bank, C., Hietpas, R. T., Jensen, J. D. & Bolon, D. N. A. A Systematic Survey of an Intragenic Epistatic Landscape. *Mol. Biol. Evo.* **32**, 229–238 (2015).

71. Garst, A. D., Bassalo, M. C., Pines, G., Lynch, S. A., Halweg-Edwards, A. L., Liu, R., Liang, L., Wang, Z., Zeitoun, R., Alexander, W. G. & Gill, R. T. Genome-wide mapping of mutations at single-nucleotide resolution for protein, metabolic and genome engineering. *Nat. Biotechnol.* **35**, 48–55 (2016).

72. Al-Mawsawi, L. Q., Wu, N. C., Olson, C. A., Shi, V. C., Qi, H., Zheng, X., Wu, T.-T. & Sun, R. High-throughput profiling of point mutations across the HIV-1 genome. *Retrovirology* **11**, 124 (2014).

73. Ferguson, A. L., Mann, J. K., Omarjee, S., Ndung, T. & Walker, B. D. Resource Translating HIV Sequences into Quantitative Fitness Landscapes Predicts Viral Vulnerabilities for Rational Immunogen Design. *Immunity* **38**, 606–617 (2012).

74. Lauring, A. S. & Andino, R. Exploring the Fitness Landscape of an RNA Virus by Using a Universal Barcode Microarray. *J. Virol.* **85**, 3780–3791 (2011).

75. Visher, E., Whitefield, S. E., McCrone, J. T., Fitzsimmons, W. & Lauring, A. S. The Mutational

Robustness of Influenza A Virus. *PLoS Pathog.* **12**, e1005856 (2016).

76. Betancourt, A. J. Genomewide patterns of substitution in adaptively evolving populations of the RNA bacteriophage MS2. *Genetics* **181**, 1535–1544 (2009).

77. Acevedo, A., Brodsky, L. & Andino, R. Mutational and fitness landscapes of an RNA virus revealed through population sequencing. *Nature* **505**, 686–690 (2013).

78. Wu, N. C., Young, A. P., Al-Mawsawi, L. Q., Olson, C. A., Feng, J., Qi, H., Chen, S.-H., Lu, I.-H., Lin, C.-Y., Chin, R. G., Luan, H. H., Nguyen, N., Nelson, S. F., Li, X., Wu, T.-T. & Sun, R. High-throughput profiling of influenza A virus hemagglutinin gene at single-nucleotide resolution. *Sci. Rep.* **4**, 4942 (2015).

79. Sarkisyan, K. S., Bolotin, D. A., Meer, M. V., Usmanova, D. R., Mishin, A. S., Sharonov, G. V., Ivankov, D. N., Bozhanova, N. G., Baranov, M. S., Soylemez, O., Bogatyreva, N. S., Vlasov, P. K., Egorov, E. S., Logacheva, M. D., Kondrashov, A. S., Chudakov, D. M., Putintseva, E. V., Mamedov, I. Z., Tawfik, D. S., *et al*. Local fitness landscape of the green fluorescent protein. *Nature* **533**, 397–401 (2016).

80. Hartman, E. C., Jakobson, C. M., Favor, A. H., Lobba, M. J., Álvarez-Benedicto, E., Francis, M. B. & Tullman-Ercek, D. Quantitative characterization of all single amino acid variants of a viral capsid-based drug delivery vehicle. *Nat. Commun.* **9**, 1385 (2018).

81. Hartman, E. C., Lobba, M. J., Favor, A. H., Robinson, S. A., Francis, M. B. & Tullman-Ercek, D. Experimental Evaluation of Coevolution in a Self-Assembling Particle. *Biochemistry* (2018). *ASAP*.

82. Brauer, D. D., Hartman, E. C., Bader, D. L. V., Mertz, Z. N., Tullman-Ercek, D. & Francis, M. B. Systematic Engineering of a Protein Nanocage for High-Yield, Site-Specific Modification. *ChemRxiv* (2018).

83. Fowler, D. M. & Fields, S. Deep mutational scanning: a new style of protein science. *Nat. Methods* **11**, 801–807 (2014).

84. Esvelt, K. M., Carlson, J. C. & Liu, D. R. A system for the continuous directed evolution of biomolecules. *Nature* **472**, 499–503 (2011).

85. Carlson, J. C., Badran, A. H., Guggiana-Nilo, D. A. & Liu, D. R. Negative selection and stringency modulation in phage-assisted continuous evolution. *Nat. Chem. Biol*. **10**, 216–222 (2014).

86. Dickinson, B. C., Packer, M. S., Badran, A. H. & Liu, D. R. A system for the continuous directed evolution of proteases rapidly reveals drug-resistance mutations. *Nat. Commun.* **5**, 5352 (2015).

87. Wong, B. G., Mancuso, C. P., Kiriakov, S., Bashor, C. J. & Khalil, A. S. Precise, automated control of conditions for high-throughput growth of yeast and bacteria with eVOLVER. *Nat. Biotechnol*. **36**, 614–623 (2018).

88. Plesa, C., Sidore, A. M., Lubock, N. B., Zhang, D. & Kosuri, S. Multiplexed gene synthesis in emulsions for exploring protein functional landscapes. *Science* **359**, 343–347 (2018).

89. Meyer, M. M., Hochrein, L. & Arnold, F. H. Structure-guided SCHEMA recombination of distantly related β-lactamases. *Protein Eng. Des. Sel.* **19**, 563–570 (2006).

90. Hartl, D. L. What can we learn from fitness landscapes? *Curr. Opin. Microbiol.* **21**, 51–57 (2014).

91. Payne, J. L. & Wagner, A. The causes of evolvability and their evolution. *Nat. Rev. Genet.* **20**, 24–38 (2018).

92. de Visser, J. A. G. M. & Krug, J. Empirical fitness landscapes and the predictability of evolution. *Nat. Rev. Genet.* **15**, 480–490 (2014).

93. Sailer, Z. R. & Harms, M. J. High-order epistasis shapes evolutionary trajectories. *PLoS Comput. Biol.* **13**, e1005541 (2017).

94. Bershtein, S., Segal, M., Bekerman, R., Tokuriki, N. & Tawfik, D. S. Robustness–epistasis link shapes the fitness landscape of a randomly drifting protein. *Nature* **444**, 929–932 (2006).

95. Pines, G., Winkler, J. D., Pines, A. & Gill, R. T. Refactoring the genetic code for increased evolvability. *MBio* **8**, (2017).

96. Podgornaia, A. I. & Laub, M. T. Pervasive degeneracy and epistasis in a protein-protein interface. *Science* **347**, 673–677 (2015).

97. Romero, P. A., Tran, T. M. & Abate, A. R. Dissecting enzyme function with microfluidic-based deep mutational scanning. *Proc. Natl. Acad. Sci.* **112**, 7159–7164 (2015).

98. Weinreich, D. M., Delaney, N. F., DePristo, M. A. & Hartl, D. L. Darwinian evolution can follow only very few mutational paths to fitter proteins. *Science* **312**, 111–114 (2006).

99. Melamed, D., Young, D. L., Gamble, C. E., Miller, C. R. & Fields, S. Deep mutational scanning of an RRM domain of the Saccharomyces cerevisiae poly(A)-binding protein. *RNA* **19**, 1537–1551 (2013).

100. Jacquier, H., Birgy, A., Le Nagard, H., Mechulam, Y., Schmitt, E., Glodt, J., Bercot, B., Petit, E., Poulain, J., Barnaud, G., Gros, P.-A. & Tenaillon, O. Capturing the mutational landscape of the beta-lactamase TEM-1. *Proc. Natl. Acad. Sci.* **110**, 13067–13072 (2013).

101. Wu, N. C., Dai, L., Olson, C. A., Lloyd-Smith, J. O. & Sun, R. Adaptation in protein fitness landscapes is facilitated by indirect paths. *Elife* **5**, (2016).

102. Wrenbeck, E. E., Azouz, L. R. & Whitehead, T. A. Single-mutation fitness landscapes for an enzyme on multiple substrates reveal specificity is globally encoded. *Nat. Commun.* **8**, 1–10 (2017).

Chapter 2: Quantitative characterization of all single amino acid variants of a viral capsid-based drug delivery vehicle

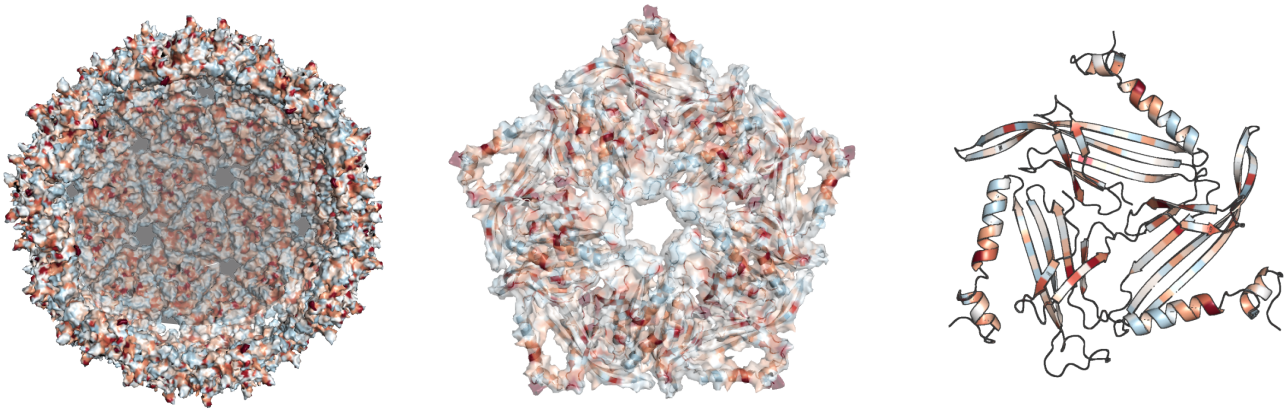
The following is adapted from Hartman, Jakobson, Favor, Lobba, Álvarez-Benedicto, Francis, and Tullman-Ercek; Nature Communications, 2018 with permission

Short summary:

We developed and validated a new way to study the mutability of self-assembling proteins. We used this technique to evaluate how all single amino acid mutations of a viral capsid (and drug delivery vehicle) alters its assembly.

Abstract

Self-assembling proteins are critical to biological systems and industrial technologies, but predicting how mutations affect self-assembly remains a significant challenge. Here, we report a new technique, termed SyMAPS (Systematic Mutation and Assembled Particle Selection), that can be used to characterize the assembly competency of all single amino acid variants of a self-assembling viral structural protein. SyMAPS studies on the MS2 bacteriophage coat protein revealed a high-resolution fitness landscape that challenges some conventional assumptions of protein engineering. An additional round of selection identified a previously unknown variant (CP[T71H]) that is stable at neutral pH but less tolerant to acidic conditions than the wild-type coat protein. The capsids formed by this variant could be more amenable to disassembly in late endosomes or early lysosomes—a feature that is advantageous for delivery applications. In addition to providing a mutability blueprint for virus-like particles, SyMAPS can be readily applied to other self-assembling proteins.



2.A. Introduction

Protein self-assembly relies on optimally-balanced energetics that arise in part from the complex interplay of amino acids in apposed protein monomers[1]. Remarkable progress has been achieved using computational methods to understand these interactions, yielding several compelling examples of designed closed-shell structures[2–5] and encapsulated enzymes[6,7]. However, the subtly-cooperative nature of these interactions still makes it difficult to predict how particular amino acid substitutions will affect self-assembly behavior. Furthermore, single amino acid substitutions can lead to significant changes to the structure and function of a protein or protein assembly, often leading to surprising outcomes, further complicating computational predictions[8–10]. As such, engineering particles with specifically desired assembly properties remains a challenging goal.

Protein fitness landscapes can provide a useful complementary tool by describing the ways in which systematic changes in primary sequence alter the resulting self-assembly competency[11–13]. In a protein fitness landscape, all possible variants of a protein sequences are ordered such that primary sequences differ only by single amino acid mutations, and variant effect on a functional output is quantified[12]. To date, most quantified fitness landscapes have been determined for enzymes or proteins with a straightforward selection or screen, where fitness is defined as catalytic activity[14,15], binding[16,17], growth[18], or fluorescence[19]. Fitness landscapes have also been explored for a variety of viruses, including human pathogens like HIV, influenza, polio, and others, using infectivity as the selection criterion[20–25]. Recently, a synthetic icosahedral protein assembly was engineered to encapsulate its own genome, and its fitness landscape was evaluated[26]. Several amino acid substitutions yielded improved genome packaging, serum stability, and circulation behavior compared to the original synthetic nucleocapsid. Viral fitness landscapes can also be used to predict vaccine candidates and characterize viral evolution[27], though the infectivity selection strategy combines protein structure, replication, and attachment into a single selection step. Viral fitness landscapes can be generated using bioinformatics techniques, but this requires vast sequence information[21], ruling out its use for little-sequenced viruses, such as bacteriophages or zoonotic pathogens.

Here, we describe a library generation and single-step selection strategy—termed SyMAPS (Systematic Mutation and Assembled Particle Selection)—to study the structure a self-assembling protein capsid composed of a noninfectious viral structural protein, or virus-like particle (VLP). This selection does not rely on infectivity, clinical abundance, or serum stability, and therefore enables experimental characterization of all single amino acid variants of MS2 bacteriophage coat protein (MS2 CP). The resulting fitness landscape is a fundamental roadmap to altering the MS2 CP to achieve tunable chemical and physical properties. After recapitulating the results of many previous investigations in a single experiment[28–30], we separately calculated the effect of ten physical properties on the Apparent Fitness Landscape (AFL), and we evaluated the validity of several common protein engineering assumptions. An additional round of selection identified a previously unknown variant, CP[T71H], that exhibits acid-sensitive properties that are promising for engineering controlled endosomal release of cargo in targeted drug delivery. The library of MS2 variants can be subjected to future selections to address any number of additional engineering goals. In addition, SyMAPS is a straightforward approach that can be applied

more broadly to assess the fitness landscapes of the coat proteins of clinically-relevant pathogens, including hepatitis B and human papillomavirus virions[31].

2.B. Results

2.B.i. Generating a virus-like particle fitness landscape

MS2 bacteriophage is a well-studied single-stranded RNA virus[32]. MS2 VLPs are composed of 180 copies of a single coat protein, which adopts three conformations (A, B, and C) to form a quasi-equivalent T=3 protein shell. The assembly, structure, and utility of this particle have been well characterized[28,29,33–37], making it an ideal candidate to map its fitness landscape. Additionally, MS2 VLPs are an attractive target as they are promising vehicles to deliver small molecule, nucleic acid, or protein cargo to host cells[28,35,36,38]. As MS2 VLPs are used primarily as scaffolding in targeted drug delivery, selections based on infectivity are less relevant in this context. In addition, infectivity selections can fail to identify variants that are noninfectious but have useful physical properties. For instance, we recently discovered with a MS2 CP variant that confers a smaller capsid geometry but is incapable of encapsulating its native genome[10]. As scaffolds, MS2 VLPs are biocompatible, homogeneous, and stable to high temperatures (Tm = 68 °C) and a wide pH range (pH 3-10)[39–41]. The interior cavity is available to load cargo, protecting it from the external environment, while the exterior can be chemically modified to display targeting groups such as peptides or antibodies[29,42]. Studies have shown that these VLPs are surprisingly long-lived in the bloodstream of mice, stable to serum, and accumulate in several tissues of interest[42,43]. When overexpressed in a bacterial host, the MS2 CP can spontaneously self-assemble into virus-like particles (VLPs).

While the MS2 CP is best-known for a sequence-specific, high-affinity protein–RNA interaction between the CP interior and a short RNA stem loop known as the Translational Operator (TR-RNA)[44], these particles can also nucleate nonspecifically on available negatively-charged material to form a VLP. This has been demonstrated for DNA[45], negatively charged proteins[38,46], anionic polymers[38], and anionic nanoparticles[47]. To measure the fitness landscape of the MS2 CP, we harnessed this behavior to establish a genotype-to-phenotype link. This results in the encapsulation of a sample of the mRNA strands inside the cell, including the strands that encode the coat protein itself, if capsid assembly can occur. We then selected for well-formed particles using a size-based purification to separate assembled particles from dimers, truncations, aggregates, and unencapsidated nucleic acids. This strategy was validated using a non-assembling CP variant, a truncated variant, and wild-type MS2 (**Supplementary Fig. 2.1**).

To generate a SyMAPS fitness landscape, we first synthesized and characterized a targeted library of all single amino acid variant of the MS2 CP. Every codon in the MS2 *cp* gene was separately swapped for a degenerate NNK codon, which encodes for all twenty amino acids and one stop codon, using a modified version of EMPIRIC cloning[48]. By coupling comprehensive codon mutagenesis with high-throughput sequencing, we quantified MS2 CP variant abundance before and after our size-based selection (**Figure 2.1**). The targeted library contained 2580 possible variants. Of these, over 90 percent were identified in our plasmid library prior to applying selective pressure. Following selective pressure, 15 percent of the library was not identified during sequencing—meaning, these

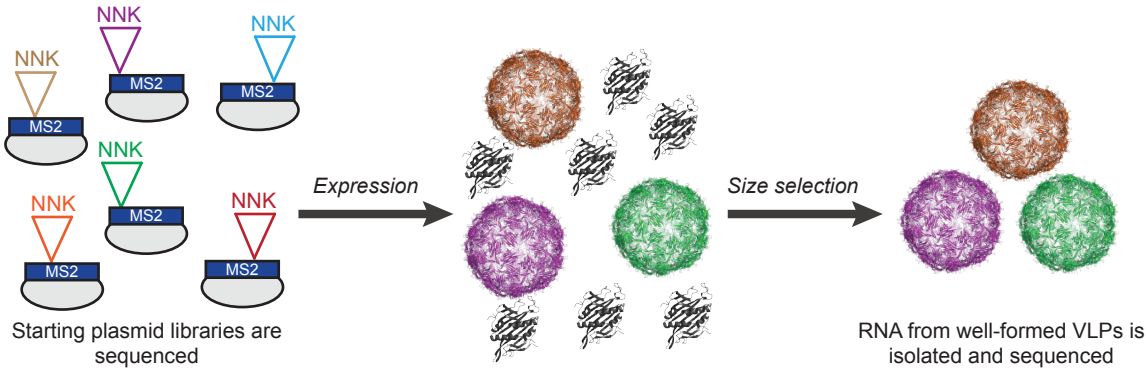


Figure 2.1. The SyMAPS approach to understanding VLP self-assembly. The NNK codon = (Any)(Any)(G/T), and results in the systematic incorporation of all 20 amino acids and a stop codon in each position of the sequence. FPLC SEC was used as a size selection, enriching for well-formed VLPs.

variants were selected against and thus are not expected to express and/or assemble into well-formed particles. Around 75 percent were identified in our assembled virus-like particle library, though relative abundances changed between these two conditions.

Under this selective pressure, we expected assembly-competent variants to increase in relative abundance after a size selection, while non-assembling variants were expected to decrease in abundance. By comparing the percent abundance of every variant across three biological replicates, we generated a quantitative Apparent Fitness Landscape (AFL), which contains an Apparent Fitness Score (AFS) for every variant at every residue across the backbone of the self-assembling MS2 CP (**Figure 2.2**, **Supplementary Data 2.1**). Positive apparent fitness scores (blue) indicate increased variant abundance following selection, while negative scores (red) indicate decreased abundance. Standard deviations were calculated across the three biological replicates (**Supplementary Fig. 2.2**).

2.B.ii. Fitness landscape reflects biophysical expectations

To validate our fitness landscape, we interpreted silent and nonsense mutations as positive and negative controls, respectively. Silent mutations encode for wild type amino acids and should exhibit a wild type-like phenotype, resulting in a positive AFS. In contrast, nonsense mutations insert stop codons in the coding region of the MS2 CP, and these mutations should produce truncations and result in a negative AFS. We found that the average AFS of silent mutations in our fitness landscape was 0.63, with a standard deviation of 0.3. The average AFS of nonsense mutations was –2.66 with a standard deviation of 1.6 (**Figure 3A**), significantly lower than the mean AFS of silent mutations ($p < 10\text{-}4$ by Student's two-tailed T-test).

We also evaluated the reproducibility of the three replicates by generating scatterplots of each replicate and finding the trendline and R^2 values of each comparison (**Supplementary Fig. 2.3A-C**). We find good correlation between the three replicates with R^2 values consistently around 0.64. While some variation exists between individual datapoints, few scores with an absolute value greater than 0.5 are inverted. However, these results do suggest that positive and negative scores should be considered as binary results, as it is unlikely that a score such as 0.5 is fundamentally different from a score

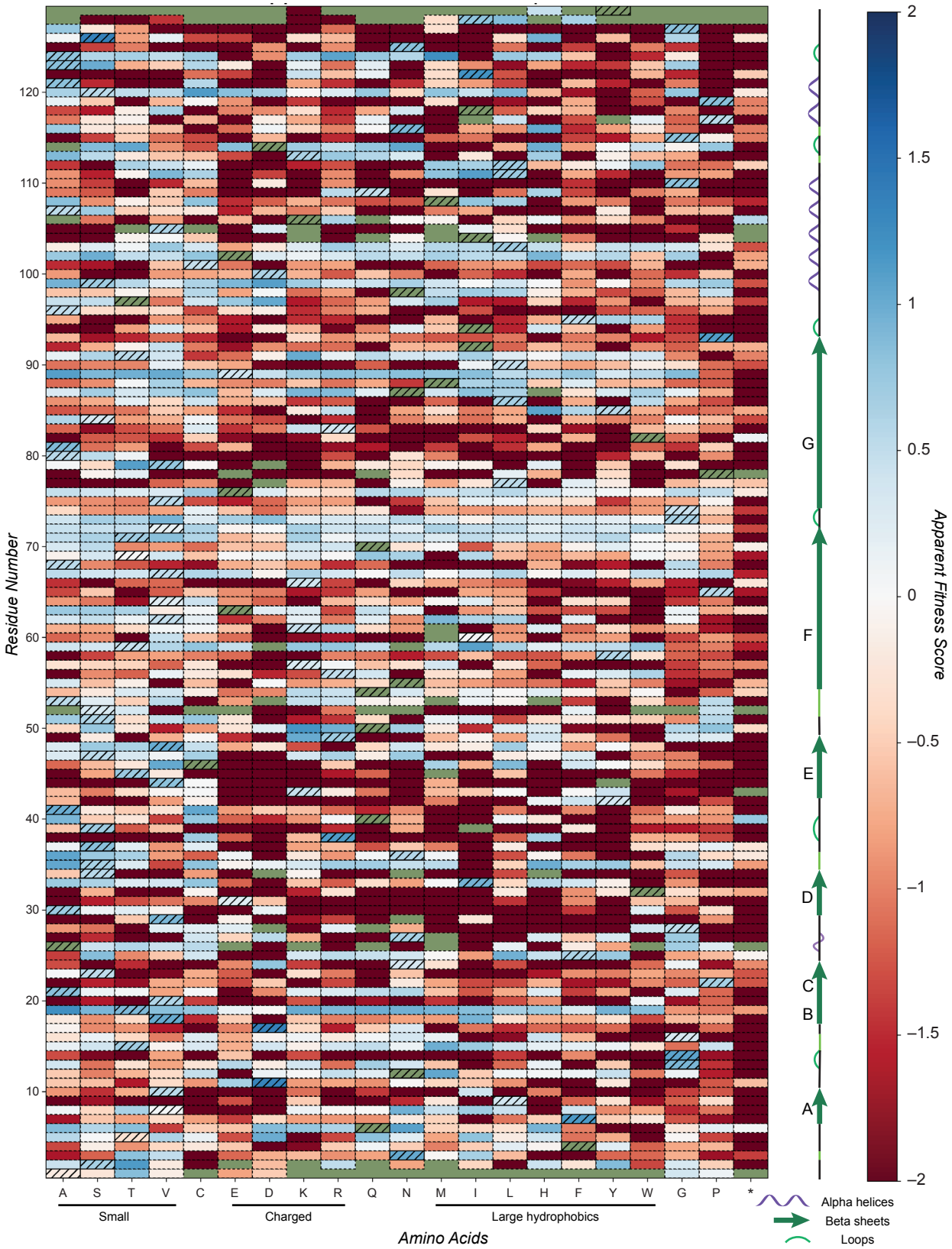


Figure 2.2. Apparent Fitness Scores (AFS, n=3) for all single amino acid variants of the MS2 coat protein (MS2 CP). Wild-type residues are indicated with hatches, and missing values are green. Dark red variants were sequenced before selection but absent following selection.

such as 1.5 in this dataset.

A specific region of the MS2 CP with secondary structure that should produce characteristic mutability patterns was validated. The MS2 CP is rich in β-sheets, and β-sheet G is positioned such that one face of the sheet is oriented towards solvent, while the other is oriented towards the protein core (**Figure 2.3C**). Odd-numbered positions, which face solvent, were expected to accommodate charged residues much more easily than the even-numbered, core-facing residues. The AFL shows that lysine and arginine are indeed only tolerated at water-facing positions (**Figure 2.3B**). In contrast, leucine is tolerated at every position along this β-sheet, while glycine and proline, which are both expected to disrupt β-sheets, are poorly tolerated at 5 of 6 and 6 of 6 positions in this region, respectively. These results match biological intuition and validate the selection strategy.

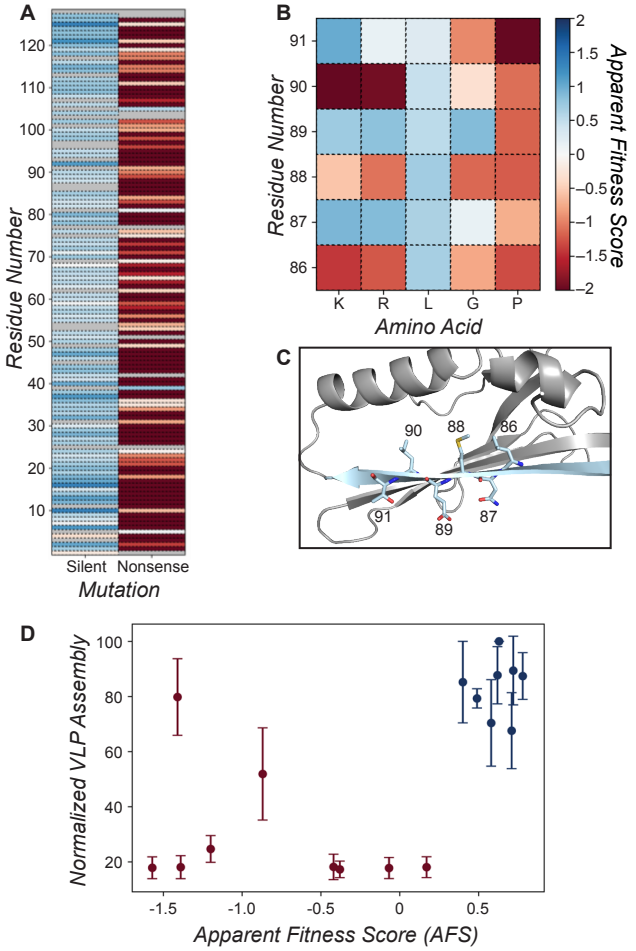


Figure 2.3. Validation of Apparent Fitness Landscape (AFL). A) Synonymous and nonsense mutation values are plotted against residue number. B,C) Beta sheet G (light blue) shows predicted mutability patterns. D) The assembly assay (n=3) correlates well with AFS values. Red points were identified as non-assembling in the AFL, and blue points represent mutants with favorable assembly. Error bars indicate standard deviation.

2.B.iii. Mutability Index identifies highly mutable residues

To quantify the mutability of each residue, the Shannon entropy[49]—a measure of diversity at a given residue—was calculated as a proxy for mutability. Comparing Shannon entropy before and after our selection yielded the Mutability Index (MI) of each residue (**Figure 2.4A,B**, **Supplementary Data 2.2**). This mutability score enables the identification of attractive sites to rationally engineer properties of interest, since positions with higher MI tolerate a wider array of amino acids at that position.

Across the backbone of the MS2 CP, we see a range of MI values, though the average Shannon entropy before selection is higher than the average post-selection Shannon entropy, as expected (**Supplementary Data 2.3**). On the exterior of the VLP, several residues are identified as highly mutable, including residue 19. Previous studies successfully installed unnatural amino acids at position 19, validating this finding[29]. On the VLP interior, several positions are highly mutable (**Figure 2.4C**). One of these residues is position 87, where a reactive cysteine is commonly installed to load small molecule cargo, such as drugs or

imaging agents[50]. Perhaps the most mutable region of the MS2 CP contains residues facing the pore, in what is known as the FG loop. In some ways, this is surprising: like other quasi-equivalent structures, the MS2 CP must engage in a conformational change to form three monomeric structures known as the A, B, and C monomers in order to form a closed icosahedral structure[51]. During this process, the FG loop engages in a critical conformational shift, including a *cis/trans* isomerization at Pro-78[52]. The loop is highly flexible, and this flexibility is critical for VLP assembly[34]. Previous work in our lab successfully mutated this region to alter the kinetics of small molecule transport, and others also have successfully manipulated this region[53,54]. Indeed, this mutability indicates that region may be a useful site to insert small, flexible peptides into the MS2 CP.

Our experimental results can be used to understand the parameters that are critical for self-assembly of biomacromolecules like virus-like particles. To this end, the MI of each residue was compared to its Accessible Surface Area (ASA), which measures the exposure of the residue to solvent. ASA has been used previously to identify mutable, exterior-facing residues[55]. Interestingly, ASA did not globally correlate with MI (**Supplementary Fig. 2.4A,B**). However, trends in a region of secondary structure did match between these two metrics. In β-sheet G, which shows a clear difference in mutability between solvent-facing and core-facing residues, similar patterns were observed between MI and ASA, as calculated by PDBePISA (**Supplementary Fig. 2.4C**)[56]. Although accessibility does contribute to mutability, additional factors contribute to the complex quaternary structure of this VLP. This result underscores the importance of experimentally generating protein fitness landscapes, particularly for self-assembling structural proteins like VLPs.

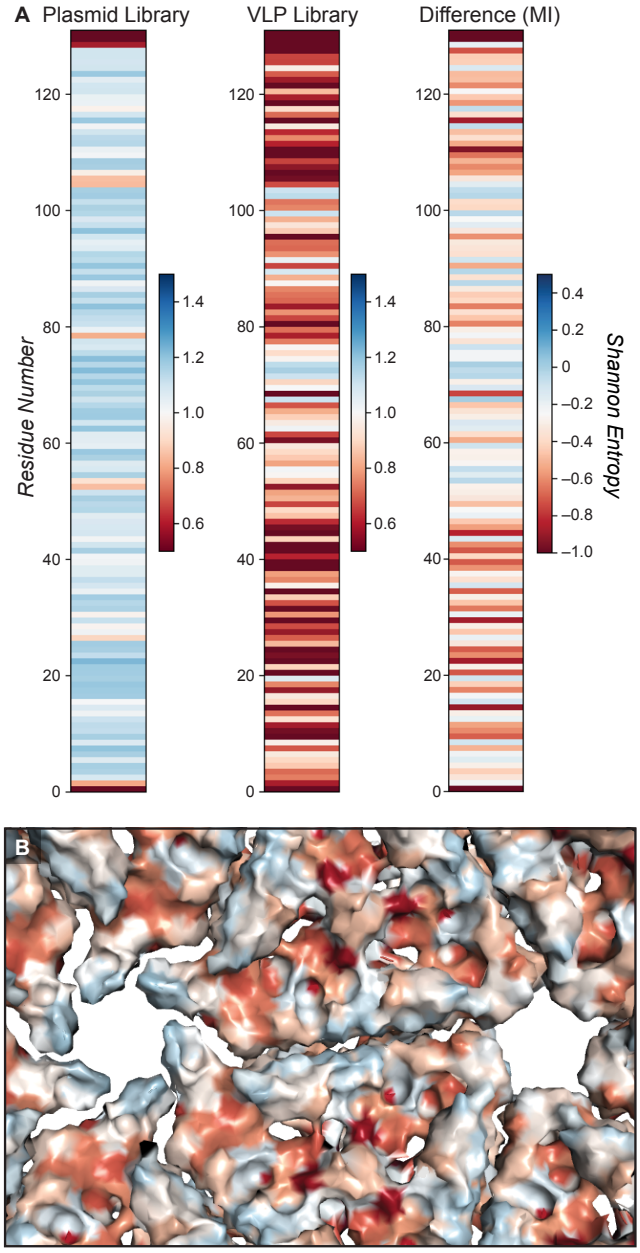


Figure 2.4. Mutability of MS2 CP. A) The Shannon Entropy (SE) of the plasmid library and VLP library are shown, along with the Mutability Index (MI, VLP SE – plasmid SE). C) The MI is also shown from the interior perspective of assembled MS2 CP (PDB ID = 2MS2) using the same coloring scheme as part A.

2.B.iv. Apparent Fitness Score confirms VLP assembly

To be a valuable engineering tool, the Apparent Fitness Landscape should enable a priori insight of whether a MS2 CP variant will self-assemble into a well-formed VLP, even if the mutation is not at an intuitive residue. To confirm this, sixteen example variants with an AFS from –1.6 to 0.8 were selected, with an emphasis on proline mutations, which are often difficult to insert into a complex structure, and cysteine mutations, which are useful for bioconjugation. These variants were evaluated for VLP formation. Three of the four proline mutations were expected to form VLPs and one was not. CP[V67P] has an AFS of 0.6 despite inserting a proline into the middle of a β-sheet—a change canonically expected to be highly disruptive to protein structure. CP[E76P] was permitted but CP[L77P] was not, and both are close to the conserved Pro78, which adopts a *cis* conformation in the B monomer structure. CP[S99P] inserts a proline at the base of an alpha helix. We also selected six cysteine variants patterned across the CP. In addition, several selected variants (CP[Q6V] and CP[Q50V]) have low AFS, though the residue is highly mutable, and one (CP[N24D]) has a high AFS at a poorly mutable residue.

All sixteen variants were characterized in an assembly assay (**Figure 2.3D**). Fourteen of the sixteen variants showed similar trends when comparing apparent fitness to the assembly assay. False positives—that is, variants expected to assemble that do not—would be problematic for applying fitness landscapes to engineering applications, but were not observed among the tested subset.

The assembly assay identified two variants that formed VLPs even though their AFS was low (CP[T91C] AFS = –1.4; CP[Q50C] AFS = –0.87). Upon further analysis by transmission electron microscopy (TEM), both false negatives (CP[T91C] and CP[Q50C]) were observed to form non-VLP, fibril-like aggregates in addition to wild type-shaped VLPs (**Supplementary Fig. 2.5**). Several other variants (CP[P78L] and CP[N36C]) also formed non-VLP aggregates, though these did not form wild-type-shaped VLPs, and their AFS was low.

A similar fibril-like phenotype has previously been reported when high concentrations of the VLP assembly-inducing oligonucleotide TR-DNA were applied to MS2 CP dimers during an *in vitro* reassembly assay[36]. TR-DNA induces a conformational switch from the C/C to A/B type dimers[57], and in this reported case, high concentrations of TR-DNA likely resulted in a dimer imbalance that produced non-VLP aggregates. Here, we hypothesize that these CP[T91C] and CP[Q50C] variations resulted in a similar imbalance in abundance of C/C or A/B type conformations, yielding the rod- or fibril-like structures. Furthermore, in subsequent studies, we identify additional, potentially useful variants with a stable shift to this fibril phenotype (see **Chapter 5**). These non-VLP aggregates likely were not enriched in our size-based selection, yielding a low AFS value. This result shows that variants forming non-VLP structures may be penalized in our selection.

2.B.v. Interpretation of Apparent Fitness Scores

We sought to screen a wider set of protein variants to interpret the behavior of positive and negative scores more clearly. To do so, we first optimized a small-scale assembly screen, which could be applied to a wider set of variants. We compared six methods of cell lysis—which is the slowest step in the large-scale assembly assay—on

wild-type MS2, assaying each for VLP yield by HPLC. We found that freeze-thaw followed by a short round of sonication yielded the highest and most consistent amount of VLP (**Supplementary Fig. 2.6A**).

We next generated and performed an assembly screen on all variants at position 49 and 91 using the optimized small-scale assembly assay. Both are RNA-binding residues, though residue 91 is more mutable than residue 49. While this assembly screen did not perfectly mirror the assembly selection (both lysis and purification were changed to enable higher throughput), a relatively large subset of the AFL could be tested in these experiments. We found that variants with a score higher than 0.2 generally formed assembled VLPs (**Supplementary Fig. 2.6B**). Scores between 0.2 and –0.2 were challenging to interpret, and these variants ranged from well-assembled to poorly-assembled. Finally, a number of variants with a score less than –0.2 appeared to form some amount of assembled VLP, though the majority did not form VLPs.

From these results, we can conclude that positive scores generally correspond to well-assembled VLPs, but negative scores likely arise from a number of different outcomes. Thus, we interpret variants with a negative AFS value as corresponding to capsids that exhibit low expression, assemble poorly, and/or are unstable toward protein purification conditions. All of these properties would likely limit their utility for delivery applications. As with any selection, the resulting quantitative fitness score could also be influenced by limited growth of the host cells, although this selection strategy eliminated attachment, whole genome encapsulation, among other variables required for infectivity selections.

Of particular concern was whether mutations to the RNA binding pocket[58,59] in the VLP interior could yield a stable VLP that poorly binds RNA and thus is not detected by high-throughput sequencing. While we cannot rule out this possibility completely, we think it is unlikely for several reasons: 1) We are relying on passive nucleic acid encapsulation[38] rather than the specific and high-affinity interaction between the MS2 CP and the TR-RNA stem loop; 2) Simple negative charge is sufficient to stimulate VLP reassembly *in vitro*; 3) Recent structural data suggest many contacts between the MS2 CP and its genomic RNA[33], making individual point mutations less likely to influence the overall avidity of binding interactions; and 4) In the SyMAPS dataset, critical RNA-binding residues on average have high Mutability Index scores. Indeed, of the eleven RNA binding residues, only two fell lower than the 30th percentile in mutability (**Supplementary Fig. 2.7**). An additional three residues had mutability scores between the 30th and 70th percentile, and the remaining six had a MI higher than 70 percent of other residues.

Sequencing read errors are the most likely reason to generate false positives; indeed, two of the 129 nonsense mutations have AFS values higher than zero, even though nonsense mutations cannot form well-formed VLPs. Upon closer inspection, both of the wild-type codons at these positions are one base pair away from TAG, the stop codon present in an NNK codon (residue 40, CAG; residue 106, AAG), making these positions more susceptible to sequencing read errors. We generated a higher-stringency AFL that only analyzed codons that are two or more base pairs away from wild-type, essentially eliminating the effect of sequencing read errors (**Supplementary Fig. 2.8**). We find that mutability trends hold in the stringent AFL, leading us to conclude that sequencing read errors minimally affect the AFL.

2.B.vi. Apparent Fitness Landscape shows complexity of self-assembly

Ten physical properties of amino acids were evaluated to determine how each property contributes to mutability across the VLP (**Figure 2.5**). Size-based parameters (volume, molecular weight, length, and steric bulk) seem to group by region, where a set of several proximally positioned residues show similar preferences. In contrast, polarity, polar area, and flexibility show a distinct banding patterns across the protein, indicating a strong preference based on the local environment of a single residue.

This analysis can also be used to understand the physical preferences of an individual residue. For example, residue S99 is a relatively mutable residue at the base of an alpha helix. We can see that position 99 prefers smaller, flexible, or polar residues, but is biased against hydrophobic residues. In contrast, position E89—one of the most mutable residues in the VLP—has minimal preferences for any particular property except flexibility, which is disfavored.

We can further use these analyses to identify locations to insert non-native amino acids, which tend to be larger than native amino acids. Several residues seem to prefer larger amino acids, including residue T19, a position where non-native residues have been installed. Combining these analyses with the Mutability Index and AFL, we anticipate that residues Q54 or T59 may be additional locations where we can install non-native amino acids. Taken together, the complexity of these results highlights the importance of measuring a fitness landscape directly.

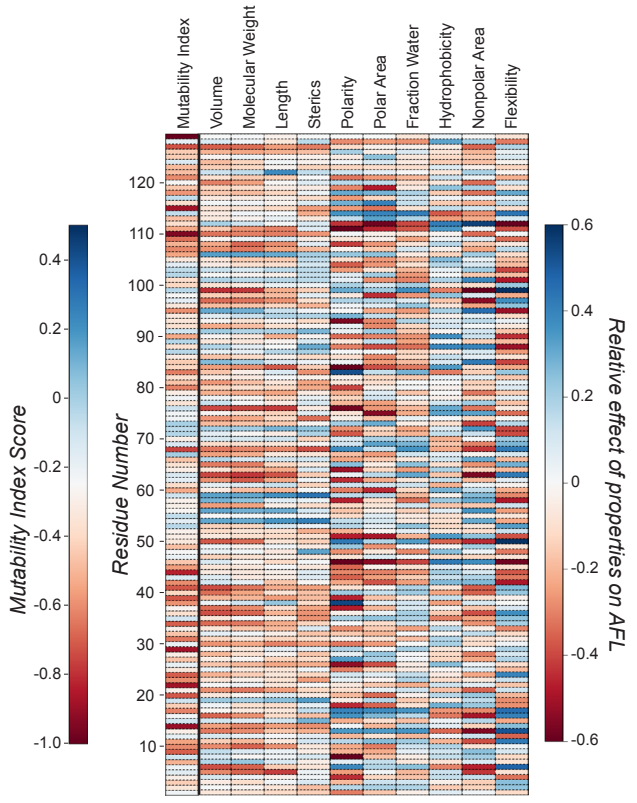


Figure 2.5. The effect of physical properties on the Apparent Fitness Landscape (AFL). Each value represents the sum of the Apparent Fitness Score (AFS) multiplied by a scalar corresponding to the indicated property. Blue scores mean that the position showed a preference for high property values, while red scores show a preference against high property values. Mutability Index is shown as a reference.

2.B.vii. Engineering an acid sensitive MS2 CP variant

We hypothesized that variants in this library may have useful differences in physical properties compared to the wild-type MS2 CP. To uncover specific variants that exhibit desired traits, we applied additional selective pressures to the library. MS2 VLPs have been used to deliver a variety of cargo to cells other than the canonical *E. coli* hosts, including sensitive biomolecules such as proteins and RNA[36,60]. In these cells, release from the VLP is presumed to occur via an acid-triggered cargo release mechanism in the late endosome or lysosome[35]. However, an *in vitro* acid screen revealed that the

CP[WT] maintains approximately 90% soluble VLP at the acidity of lysosomes (pH 4.5). We reasoned that delivery applications could benefit from an MS2 CP variant that is less tolerant to these acidic conditions.

We exposed the library of MS2 CP variants to pH 5 at physiological temperature (37 °C) for four hours, mimicking the conditions of a human endosome or lysosome. Under these conditions, two variants (CP[T71H] and CP[E76C]) were predicted to form particles but were found to have a lower percent abundance following the selective pressure (**Supplementary Fig. 2.9A**). Other possible hits were eliminated either because the variant was not predicted to form assembled VLPs (CP[A41G]) or was predicted to be temperature sensitive rather than acid sensitive (CP[T71R]).

We constructed both variants and conducted an *in vitro* acid screen (**Figure 2.6**). Indeed, CP[T71H] exhibited a reduction in well-formed particles between pH 4.5 and 3.5 compared to CP[WT], confirming that we successfully identified a variant that is less stable to mildly acidic conditions than the CP[WT] using SyMAPS. At this pH range—which is near the acidity of late lysosomes—we observed aggregation and precipitation of CP[T71H] (**Supplementary Fig. 2.9B,C**), and these precipitated VLPs were morphologically irregular by transmission electron microscopy (TEM). We then combined CP[T71H] with CP[E76C] to form a double mutant that contained both predicted acid-sensitive variants. The acid sensitivity effect was additive, yielding even more acid sensitive variants that aggregated near pH 5. However, expression yields and storage stability were compromised, suggesting that additional rounds of optimization are still needed.

We hypothesize that the decreased stability of the protein cage in mildly acidic conditions could enhance the rate of cargo release in endosomes and lysosomes. Thus, we predict that CP[T71H] or an optimized double mutant may improve endosomal release without compromising the circulation stability and cargo protection properties that make the MS2 CP an ideal drug delivery vehicle.

2.C. Discussion

Many patterns observed in our data challenge conventional protein engineering assumptions. Smaller or hydrophobic amino acids, including valine and, surprisingly, glycine, were among the best-tolerated amino acids across the CP and may be useful to remove undesirable chemical functionalities (**Figure 2.7A**). Negative charges, bulky residues, and proline were poorly tolerated across the entire gene, and nonsense mutations were still much more detrimental, as expected. Overall, we see that the structure is more immutable than mutable, though a surprising number of locations tolerated

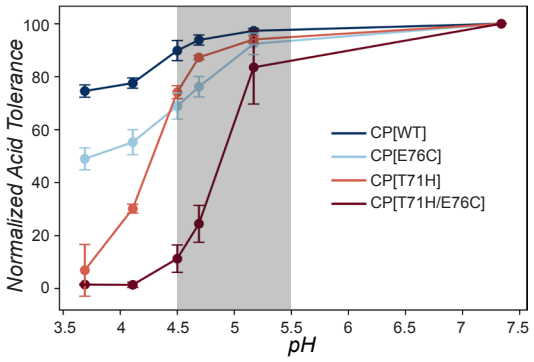


Figure 2.6. Reduced acid tolerance of MS2 CP[T71H] and CP[T71H/E76C]. Both CP[T71H] and CP[T71/E76C] are less acid tolerant than CP[WT] and CP[E76C]. Normalized acid tolerance (n=3) indicates HPLC SEC peak height relative to pH 7.3. The highlighted region is ideal for endosomal disassembly. Error bars indicate standard deviation.

modifiable amino acids like cysteine and lysine. This may be useful for increasing the capacity of the MS2 CP to carry small molecule cargo.

Several conventional protein engineering techniques are directly evaluated in this system. For example, alanine scanning is often used to identify amino acids that contribute to stability or function in a protein, as alanine is assumed to be a neutral mutation for most native residues. To evaluate this assumption, we plotted the AFS value for alanine compared to every other amino acid, then calculated the Pearson correlation for each scenario. Our results indicate that, surprisingly, alanine shows minimal correlation in mutability to most other amino acids and shows particularly poor or negative correlations to bulky amino acids, such as phenylalanine and tryptophan (**Supplementary Fig. 2.10A**). Indeed, our results show that alanine scanning may artificially inflate the chances of identifying a bulky amino acid as critical to protein stability, as the mutation likely is not neutral.

We used the same correlations to show that while lysine and arginine AFS values correlate well, and glutamate and aspartate AFS values correlate well, the two types of charges correlate poorly to one another (**Supplementary Fig. 2.10B**). This result indicates that positively and negatively charged residues, which are often evaluated together, actually behave quite differently in this system. In addition, we see more nuanced variation in substitutability between valine, isoleucine, and leucine, three hydrophobic residues that typically are considered quite similar. While isoleucine AFS values correlate well with valine and leucine AFS values, valine and leucine AFS values correlate less well with one another (**Supplementary Fig. 2.10C**), indicating that the position of the methyl branch is likely important in determining amino acid substitution patterns.

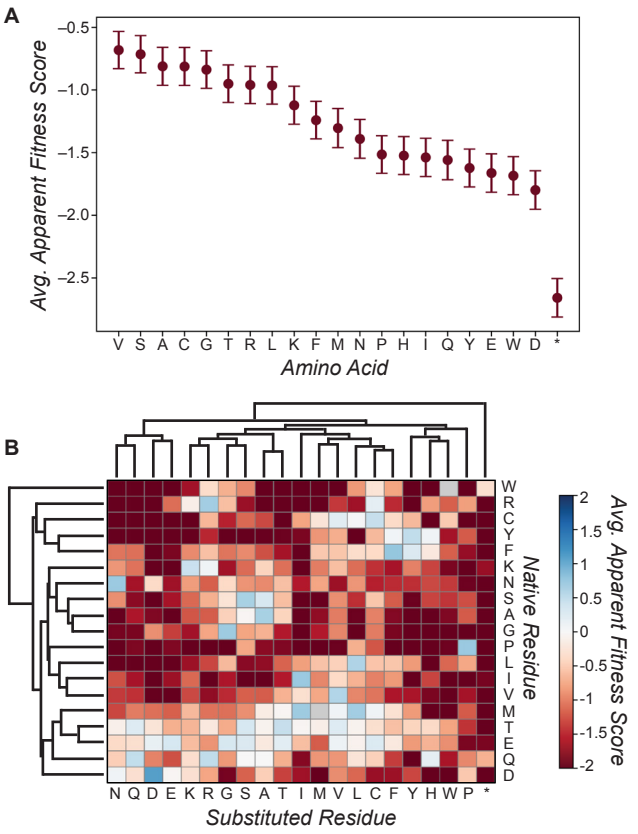


Figure 2.7. Average AFS values (n=3) of amino acids across MS2 CP. A) The average AFS and standard error values are indicated for every native residue mutated to the indicated amino acid. B) Averages from part A are separated by native residue and clustered by similarity.

The influence of the native amino acid on substitution at a particular residue in the MS2 CP was evaluated (**Figure 2.7B**). Substituted residues of similar chemistries cluster together, as would be expected. The MS2 CP has ten native positively-charged residues and nine native negatively-charged residues. On average, negatively-charged residues permitted more non-native substitutions than positively-charged residues. Several residues did not tolerate any single amino acid well, including glycine (nine native residues), proline (six

native residues), and leucine (seven native residues). Surprisingly, threonine (nine native residues) tolerated many substitutions, perhaps because it is both bulky and hydrophilic in nature. While only two cysteines are present in the native sequence, it is interesting to note that serine was not tolerated as a mutation for either cysteine. One cysteine did not tolerate any mutations, while the other tolerated residues with similar hydrophobicity index to cysteine, such as valine and leucine. No one substituted amino acid was well-tolerated at every native amino acid.

In one step, we recapitulated and expanded on many studies that characterized where mutations can be installed in the MS2 CP, both on the interior and exterior of the delivery vehicle. In this study, 530 MS2 variants were characterize that permit MS2 VLP formation. Included in this number are 35 cysteine mutants and 28 lysine mutants, which indicates that additional reactive groups may be able to be installed to increase the modification rates and carrying capacity of the MS2 CP as a delivery vehicle. In addition to selecting for assembly alone, we selected for a variant that is less tolerant to acidic conditions than wild type. This previously unknown variant, CP[T71H], is less tolerant to pH 4, which is near the pH of late endosomes or early lysosomes. Moving forward, we and others can use AFL and MI values generated in this study to guide where mutations of interest can be installed in the MS2 CP. These results may also be directly applicable to identify mutable residues in structurally-related VLPs, such as Qβ. This map saves significant time and effort and allows researchers to produce MS2 CP variants with rationally engineered, highly tunable chemical or physical properties.

2.D. Methods

2.D.i. Entry Vector Generation (EMPIRIC Cloning)

To generate libraries with single amino acid mutations, we modified a cassette ligation strategy developed by the Bolon lab in 2011[48]. This strategy uses a plasmid with self-encoded removable fragments (SERF) that are surrounded by inverted BsaI restriction sites, so BsaI digestion removes both the SERF and BsaI sites. These plasmids will be referred to as entry vectors, and the SERF contains constitutively expressed GFP to enable green/white screening. We divided the MS2 CP into five segments of 26 codons in length—a decision that allowed us to mutagenize the MS2 CP by purchasing 100 basepair single stranded DNA primers rather than full-length double stranded DNA, which can be more expensive. Five golden gate compatible entry vectors were synthesized with constitutively-active GFP swapped for a 26-codon segment of the MS2 gene[48,61] (**Supplementary Data 2.4**). Inverted BsaI cut sites were then introduced into the vectors using QuikChange mutagenesis[62].

2.D.ii. CP Variant Library Generation (EMPIRIC cloning)

Single stranded DNA primers were purchased that spanned the length of each 26-codon region of MS2 with appropriate cut sites for each corresponding entry vector (**Supplementary Data 2.4**). Primers for each entry vector were resuspended, pooled, and diluted to a final concentration of 50 ng/μL. The reverse strand was filled in using a touchdown PCR[63] with 10-mers directed to the golden gate cut sites. The amplified, double-stranded DNA was purified using a PCR Clean-up Kit (Promega, Cat# A9282),

then diluted to 1-5 ng/μL. These mixtures were cloned into their respective entry vectors using EMPIRIC cloning[48], which relies on established Golden gate cloning techniques[61]. The ligated plasmids were transformed into chemically competent DH10B *E. coli* and plated on large (245 x 245 x 20mm, #7200134, Fisher) LB-A plates with 32 μg/mL chloramphenicol. Colony number varied, but every transformation yielded a number of colonies that was at least three times the theoretical library size. This protocol was repeated in full for three total biological replicates that are fully independent from library generation through selection.

2.D.iii. CP Variant Library Expression and Purification

Colonies were scraped from the plates, combined into LB, and allowed to grow for 2 h. The pools of variants were combined by OD600 into 1 L of 2xYT (Teknova, Cat: Y0210) to generate a library of CP variants. These variants were allowed to grow to an OD of 0.6, when they were induced with 0.1 percent arabinose. Expression proceed overnight at 37 °C, and cultures were harvested and sonicated. Two rounds of ammonium sulfate precipitation at 50 percent saturation were followed by FPLC size exclusion chromatography purification (**Supplementary Fig. 2.11A**). Fractions 12-22 were harvested (**Supplementary Fig. 2.11B**).

2.D.iv. Sample Preparation for High-throughput Sequencing

Plasmid DNA was extracted prior to the assembly selection using a Zyppy Plasmid Miniprep Kit (Zymo, Cat# D4036). RNA was extracted from the CP library after the assembly selection following previously-published protocols[64], and cDNA was synthesized using the Superscript III first strand cDNA synthesis kit from Life (cat: 18080051, polyT primer). cDNA and plasmids were then amplified with two rounds of PCR to add barcodes (10 cycles) and the Illumina sequencing handles (8 cycles), following Illumina 16S Metagenomic Sequencing Library Preparation recommendations (**Supplementary Data 2.4**). Libraries were combined and analyzed by 300 PE MiSeq in collaboration with the UC Davis Sequencing Facilities. 20.6 million reads passed filter, and an overall Q30 > 79%.

2.D.v. FPLC Size Exclusion Chromatography (SyMAPS size Selection)

MS2 libraries and mutants were purified on an Akta Pure 25 L Fast Protein Liquid Chromatography (FPLC) system with a HiPrep Sephacryl S-500 HR column (GE Healthcare Life Sciences, Cat# 28935607) Size Exclusion Chromatography (SEC) column via isocratic flow with 10 mM phosphate pH 7.2, 200 mM sodium chloride, and 2 mM sodium azide. Fractions containing MS2 coat protein were harvested. FPLC SEC traces of SyMAPS replicate 1, 2, and 3, and their comparison to the wild-type CP protein, can be found in **Supplementary Fig. 2.11**.

2.D.vi. FPLC Anion Exchange

Individual CP variants were purified on an Akta Pure 25L with a hand-packed DEAE Sepharose Fast Flow column (GE Healthcare Life Sciences, Cat# 17070901). Variants were eluted with 20 mM taurine ph 9.

2.D.vii. HPLC SEC

MS2 CP variants were analyzed via isocratic flow on an Agilent 1290 Infinity HPLC with an Agilent Bio SEC-5 column (5 um, 2000A, 7.8x300nm) with isocratic flow with 10 mM phosphate pH 7.2, 200 mM sodium chloride, and 2 mM sodium azide. Wild type MS2 has a characteristic elution peak at 11.2 minutes, and peak height was used as a proxy for VLP formation.

2.D.viii. Transmission Electron Microscopy (TEM)

Samples were prepared for TEM analysis by applying an analyte solution (A280 of approximately 1) to carbon-coated copper grids for 2 min, followed by triple rinsing with dd-H_2O. The grids were then exposed to a 1.6% aqueous solution of uranyl acetate for 1 min as a negative stain. Images were obtained at the Berkeley Electron Microscope Lab using a FEI Tecnai 12 transmission electron microscope with 120 kV accelerating voltage.

2.D.ix. Agarose Gel Electrophoresis

PCR products were analyzed in a 1% agarose gel in TAE buffer (40mM Tris, 20mM acetic acid, and 1mM EDTA) with 2X SYBR Safe DNA Gel Stain (ThermoFisher Scientific, Cat# S33102) for 30 minutes at 120 volts. Agarose gels were imaged on a BioRad GelDoc EZ Imager.

2.D.x. SDS-PAGE Analysis

NuPAGE 4-12% Bis-Tris Protein Gels (Invitrogen, Cat# NP0323BOX) were used. Gels were run with 1x MES buffer for 45 min at 160 volts. Samples were loaded with Laemmli sample buffer and imaged with a Coomassie stain on a BioRad GelDoc EZ Imager.

2.D.xi. Strains

Escherichia coli DH10B Competent Cells from ThermoFisher Scientific were used for all experiments. Overnight growth from a single colony was grown for 16-20h at 37 °C shaking at 200RPM in LB-Lennox media (VWR, Cat# AAH26760) with chloramphenicol at 34 mg/L. Expressions were sub- cultured 1:100 into 2xYT media (Teknova, Cat# Y0210) with 34 mg/L chloramphenicol and allowed to express overnight at 37 °C shaking at 200RPM.

2.D.xii. Individual Variant Cloning

Individual variants were cloned using a method adapted from above. Briefly, overlap extension PCR (**Supplementary Data 2.4**) yielded a double stranded fragment that spanned the length of one entry vector chunk. Each fragment was cloned into its respective Entry Vector using standard Golden gate cloning techniques. Cloned plasmids were transformed into DH10B cells. Individual colonies were sequenced prior to expression.

2.D.xiii. Individual Variant Assembly-Competency Screen

Sixteen selected mutants were individually then expressed in 50 mL cultures of

2xYT as described. These expressions were lysed by sonication, precipitated twice with 50 % ammonium sulfate, and evaluated by HPLC SEC and TEM. Peak height percent compared to wild type at 11.2 minutes was used as a proxy for VLP formation.

2.D.xiv. Individual Variant Expression and Purification

Individual variants identified as potentially acid sensitive were expressed as described previously, precipitated with 50 percent ammonium sulfate, and purified with FPLC anion exchange as described above. Fractions containing MS2 CP variants were buffer exchanged into 10 mM phosphate pH 7.2 with an Amicon Ultra-15 Centrifugal Filter with a 100 kDa membrane filter. Concentrated variants were frozen for further analysis.

2.D.xv. Lysis method comparisons

Six lysis methods were compared for a 5 mL assembly screen. Expressions were conducted on a 5 mL scale in 2xYT as described above. Cells were pelleted and resuspended in 1 mL of 10 mM phosphate 200 mM NaCl with 2 mM NaN_3 at pH 7.3 and subjected to each lysis method: two minutes of sonication with and without four cycles of freeze-thaw; four cycles of freeze-thaw followed by passage through a fine-gauge needle; and thirty seconds of vortexing with glass beads with and without the same freeze-thaw protocol. Samples were pelled following lysis, precipitated with 50% ammonium sulfate, then resuspended and evaluated for VLP formation by HPLC SEC. Equal volume was loaded from each lysate. Subsequent small-scale assembly screens were conducted with freeze-thaw followed by two minutes of sonication for lysis and HPLC SEC for VLP quantification.

2.D.xvii. High-throughput Sequencing Data Processing

Data were trimmed using Trimmomatic[65] with a 2-unit sliding quality window of 30 and a minimum length of 30.

java -jar trimmomatic-0.36.jar PE input_forward_HTS001.fastq.gz input_reverse_HTS001.fastq.gz s1_pe s1_se s2_pe s2_se SLIDINGWINDOW:2:30 MINLEN:30

Reads were merged with FLASH (Fast Length Adjustment of SHort reads)[66] with a maximum over- lap of 167 basepairs.

flash -M 167 s1_pe s2_pe -o HTS001

Reads were then aligned to the wild-type MS2 CP reference gene with Burrows-Wheeler Aligner (BWA-MEM)[67].

bwa mem -p Reference/ref.fasta HTS001.extendedFrags.fastq > HTS001.sam

Reads were sorted and indexed with Samtools[68].

samtools view -bT Reference/ref.fasta HTS001.sam -o HTS001.bam samtools sort -o HTS001_sort.bam HTS001.bam

samtools index HTS001_sort.bam

The Picard function CleanSam was used to filter unmapped reads.

java -Xmx4g -jar picard.jar CleanSam I=HTS001_sort.bam O=HTS001_filt.bam

Reads longer or shorter than the expected length of the MS2 CP were removed.

samtools view -b -F 4 HTS001_filt.bam > HTS001_map.bam
samtools view HTS001_map.bam | grep "393M" | sort | less -S > HTS001.txt

Reads were proceed to generate an Apparent Fitness Landscape using code written in-house.

APPARENT FITNESS LANDSCAPE (AFL) DEFINITIONS

m: one of the 21 mutations encoded by the NNK codon, including the 20 canonical amino acids and a stop codon.
$m \in$ {A, S, T, V, C, E, D, K, R, Q, N, M, I, L, H, F, Y, W, G, P, * }

$A_{p,m}$: an abundance score, indexed by position, p, and mutation, m.

CA_I : position indices corresponding to each entry vector, where CA_1 = 1-26, CA_2 = 27-52, CA_3 = 53-78, CA_4 = 79-104, CA_5 = 105-129.

$PA_{I,p,m}$: a percent abundance score, indexed by position, p, and mutation, m, within positional indices, i.

$RPA_{R,p,m}$: a relative percent abundance score, indexed by position, p, and mutation, m, corresponding to replicate, R.

$f_{p,m}$: a fitness score, indexed by position, p, and mutation, m.

APPARENT FITNESS LANDSCAPE (AFL) CALCULATIONS

Aligned, trimmed sequences were analyzed with code written in-house. Following data processing described above, a textfile was produced for each experiment that contains one sequencing read per line. This file was read into python, and lines that did not begin with ATG and end with TAA were discarded.

Each line was compared to the wild-type MS2 CP, and the total number of codons containing mutations was counted. Wild-type reads, or lines without any mutations, were discarded. If more than one codon in a given line contained a mutation, then any codons with single base pair mutations (as opposed to 2- or 3- base pair changes to a single codon) were assumed to be sequencing errors and discarded. Lines with one mutated codon were kept.

Every non-wild-type codon in every read was counted into a codon abundance array. Codons that did not correspond to NNK codons, meaning, the codon ended in either C or A, were discarded. The remaining codons were translated and combined by amino acid identity to generate A.

We divided A into five submatrices corresponding to the length of each primer set (see EMPIRIC cloning). These submatrices are represented by matrices A_I. The grand sum, or the sum of all counts at every amino acid along every position, was calculated:

$$CA_I = \sum A_{I,p,m}$$

We next divided A_I by its grand sum CA_I, generating a matrix of percent abundances, PA_I:

$$PA_I = \left(\frac{A_I}{CA_I} \right)$$

The submatrices were remerged into a parent percent abundance array, $PA \in \mathbb{R}^{129 \times 21}$. These calculations were repeated for each biological replicate of VLP and plasmid libraries, generating six PA

matrices ($PA_{R,L}$) where R indicates biological replicate, and L indicates either the VLP (V) or Unselected (U) library. We calculated relative percent abundances, $RPA_{R,p,m}$ by dividing $PA_{R,V}$ by $PA_{R,U}$ for each replicate:

$$RPA_R = \left(\frac{PA_{R,V}}{PA_{R,U}}\right)$$

We calculated the mean and standard deviation across three RPA replicates. All nan values, which indicate variants that were not identified in the plasmid library, were ignored. Scores of zero, which indicate variants that were sequenced in the Unselected library but absent in the VLP library, were replaced with an arbitrary score of 0.0001:

$$RPA_{p,m} = \left(max(RPA_{p,m}, 0.0001)\right)$$

We calculated the log10 of the RPA array to calculate the final $f_{p,m}$ array.

$$f_{p,m} = log_{10}(RPA_{p,m})$$

$f_{p,m}$ is plotted in Figure 2 and is available as a supplemental csv file. Error was propagated to generate standard deviations, which are plotted in Figure S2:

$$\sigma(f_{p,m}) = \left(0.434\frac{\sigma(RPA_{p,m})}{\mu(RPA_{p,m})}\right)$$

We calculated the average AFS value for each amino acid by finding the mean $f_{p,m}$ value for every mutation, m. These values are plotted in Figure 7a. Error indicates SEM values.

MUTABILITY INDEX DEFINITIONS

Shannon Entropy can be used to calculate diversity at a given residue. Here, differential Shannon Entropy is determined to generate a Mutability Index, or MI.

$P_{p,m}$: a probability score, indexed by position, p, and mutation, m.

SE_p : a Shannon Entropy score, indexed by position.

MI_p : an score of mutability, indexed by position.

MUTABILITY INDEX CALCULATIONS

Shannon entropy is defined as:

$$ShannonEntropy = -\sum P \log(P)$$

where P refers to a given probability. We first calculate the probability of a given codon occurring within a single residue:

$$P_{p,m} = \frac{A_{p,m}}{\sum(A_p)}$$

Any zero values in the resulting average array were replaced with .00001.

$$P_{p,m} = max(P_{p,m},\ .00001)$$

We calculated the Shannon entropy at every residue, generating $SE \in \mathbb{R}^{129\times 1}$.

$$SE_p = -\sum P_{p,m} \cdot log_{10}(P_{p,m})$$

Shannon Entropy values were averaged across three biological replicates for each library. The difference between the Unselected plasmid library (U) and the VLP library (V) generated the Mutability Index, $MI \in \mathbb{R}^{129\times 1}$, which is used as a proxy for mutability and is available as a supplemental csv file.

$$MI = SE_V\ -\ SE_U$$

AVERAGE MUTABILITY CLUSTERING DEFINITIONS

$\mu_{n\to m}$: the average value of all fitness scores corresponding to substituting a native residue, n , with a mutant residue, m , averaged over all occurrences of such a mutation throughout the protein backbone.

$n \in$ {A, S, T, V, C, E, D, K, R, Q, N, M, I, L, F, Y, W, G, P }

It is worth noting that the set of native residues, n, does not contain histidine as there is no native occurrence of histidine in MS2's backbone

$m \in$ {A, S, T, V, C, E, D, K, R, Q, N, M, I, L, H, F, Y, W, G, P, * }

In consistency with our other analyses, the symbol " * " here represents a mutation to a stop-codon.

AVERAGE MUTABILITY CLUSTERING ANALYSIS

We begin by constructing an array, Υ, containing average substitution scores, $\mu_{n\to m}$, where rows correspond to native residues, and columns correspond to the substituted residue:

$$\Upsilon = \begin{bmatrix} \mu_{A\to A} & \mu_{A\to S} & \mu_{A\to T} & \mu_{A\to V} & \cdots & \mu_{A\to W} & \mu_{A\to G} & \mu_{A\to P} & \mu_{A\to *} \\ \mu_{S\to A} & \mu_{S\to S} & \mu_{S\to T} & \cdots & & \cdots & \mu_{S\to G} & \mu_{S\to P} & \mu_{S\to *} \\ \mu_{T\to A} & \mu_{T\to S} & & & & & & \mu_{T\to P} & \mu_{T\to *} \\ \mu_{V\to A} & \vdots & & \ddots & & & & \vdots & \mu_{V\to *} \\ \vdots & & & & \ddots & & & & \vdots \\ \mu_{Y\to A} & \vdots & & & & \ddots & & \vdots & \mu_{Y\to *} \\ \mu_{W\to A} & \mu_{W\to S} & & & & & & \mu_{W\to P} & \mu_{W\to *} \\ \mu_{G\to A} & \mu_{G\to S} & \mu_{G\to T} & \cdots & & \cdots & \mu_{G\to G} & \mu_{G\to P} & \mu_{G\to *} \\ \mu_{P\to A} & \mu_{P\to S} & \mu_{P\to T} & \mu_{P\to V} & \cdots & \mu_{P\to W} & \mu_{P\to G} & \mu_{P\to P} & \mu_{P\to *} \end{bmatrix}_{19\times 21}$$

We then input this array into MATLAB's "clustergram" function, to create a graphical object (visualized in Figure 7b) via a hierarchal clustering algorithm:

This hierarchal clustering algorithm sequentially reorders all columns and rows (respectively) from the input array in a manner which minimizes the difference between values in adjacent array entries – where the "difference" between the values of array entries is defined by a two-dimensional euclidean distance calculation.

PHYSICAL PROPERTY PREFERENCE DEFINITIONS

Physical properties for all amino acids were obtained from previous literature[5-13]. These values were tabulated and normalized to between 0 and 1 to allow comparison of relative preference, and apparent fitness scores were normalized from -1 to 1. The tolerance of a given residue for each physical property was obtained by the summation of the fitness for every amino acid multiplied by its normalized physical value. This results in an overall negative score for residues where the given property is detrimental (such as highly polar amino acids in hydrophobic regions) and a positive score if the property is well tolerated (such as residue flexibility in loop regions).

a: one of the 20 canonical amino acids,
$a \in$ {A, S, T, V, C, E, D, K, R, Q, N, M, I, L, H, F, Y, W, G, P }

ε: one of the 10 physical property indices used in our analysis
(volume, molecular weight, length, sterics, polarity, polar area,
fraction water, hydrophobicity, non-polar area, flexibility)

$f_{p,a}$: a fitness scores, indexed by position, p, and amino acid, a.

R_p : a vector containing the fitness scores of each amino acid for a given position, p

$$R_p = \begin{bmatrix} f_{A,p} & f_{S,p} & f_{T,p} & \cdots & f_{G,p} & f_{P,p} \end{bmatrix}_{1\times 20}$$

ξ_ε : a vector containing the physical property indices, $\varphi_{a,\varepsilon}$ corresponding to a given property, ε, and amino acid, a.

$$\xi_\varepsilon = \begin{bmatrix} \varphi_{A,\varepsilon} & \varphi_{S,\varepsilon} & \varphi_{T,\varepsilon} & \cdots & \varphi_{G,\varepsilon} & \varphi_{P,\varepsilon} \end{bmatrix}_{1\times 20}$$

$\mu(R_p)$: the mean value of the fitness scores for a given position

$\sigma(R_p)$: the standard deviation of the fitness scores for a given position

$\mu(\xi_\varepsilon)$: the mean value of the amino acid indices for a given physical property

$\sigma(\xi_\varepsilon)$: the standard deviation of the amino acid indices for a given physical property

PHYSICAL PROPERTY PREFERENCE STANDARDIZATION

We proceed to produce standardized fitness scores, $\tilde{f}_{p,a}$, by taking the difference from a given position's mean, and dividing by a position's standard deviation:

$$\tilde{f}_{p,a} = \left(\frac{f_{p,a} - \mu(R_p)}{\sigma(R_p)} \right)$$

Combining these standardized fitness scores into an array, $F \in \mathbb{R}^{129\times 20}$

$$F = \begin{bmatrix} \left(\frac{f_{1,A}-\mu_1}{\sigma_1}\right) & \left(\frac{f_{1,S}-\mu_1}{\sigma_1}\right) & \cdots & \left(\frac{f_{1,P}-\mu_1}{\sigma_1}\right) \\ \left(\frac{f_{2,A}-\mu_2}{\sigma_2}\right) & \left(\frac{f_{2,S}-\mu_2}{\sigma_2}\right) & & \left(\frac{f_{2,P}-\mu_2}{\sigma_2}\right) \\ \vdots & & \ddots & \vdots \\ \left(\frac{f_{128,A}-\mu_{128}}{\sigma_{128}}\right) & \left(\frac{f_{128,S}-\mu_{128}}{\sigma_{128}}\right) & & \left(\frac{f_{128,P}-\mu_{128}}{\sigma_{128}}\right) \\ \left(\frac{f_{129,A}-\mu_{129}}{\sigma_{129}}\right) & \left(\frac{f_{129,S}-\mu_{129}}{\sigma_{129}}\right) & \cdots & \left(\frac{f_{129,P}-\mu_{129}}{\sigma_{129}}\right) \end{bmatrix}_{129\times 20}$$

$$= \begin{bmatrix} \tilde{f}_{A,1} & \tilde{f}_{S,1} & \cdots & \tilde{f}_{G,1} & \tilde{f}_{P,1} \\ \tilde{f}_{A,2} & & & & \tilde{f}_{P,2} \\ \vdots & & \ddots & & \vdots \\ \tilde{f}_{A,128} & & & & \tilde{f}_{P,128} \\ \tilde{f}_{A,129} & \tilde{f}_{S,129} & \cdots & \tilde{f}_{G,129} & \tilde{f}_{P,129} \end{bmatrix}_{129\times 20}$$

Similarly, we produced standardized property indices $[\varphi_{a,\varepsilon}]_{scaled,0}$, by taking the difference from a given property's mean index value, and dividing by the associated standard deviation:

$$[\varphi_{a,\varepsilon}]_{scaled,0} = \left(\frac{\varphi_{a,\varepsilon} - \mu(\xi_\varepsilon)}{\sigma(\xi_\varepsilon)} \right)$$

Next, we subtracted the minimum value of $[\varphi_{a,\varepsilon}]_{scaled,0}$ for a given property, setting the minimum value to zero:

$$[\varphi_{a,\varepsilon}]_{scaled,1} = [\varphi_{a,\varepsilon}]_{scaled,0} - min\left\{ [\varphi_{A,\varepsilon}]_{scaled,0} , [\varphi_{S,\varepsilon}]_{scaled,0} , \cdots , [\varphi_{P,\varepsilon}]_{scaled,0} \right\}$$

Finally, we divide each value of $[\varphi_{a,\varepsilon}]_{scaled,1}$ for a given property by the maximum value of its associated set, thus setting the max value to 1, and producing a set of standardized indices, $\bar{\varphi}_{a,\varepsilon}$, fit between 0 and 1:

$$\bar{\varphi}_{a,\varepsilon} = \left(\frac{[\varphi_{a,\varepsilon}]_{scaled,1}}{max\left\{ [\varphi_{A,\varepsilon}]_{scaled,0} , [\varphi_{S,\varepsilon}]_{scaled,0} , \cdots , [\varphi_{P,\varepsilon}]_{scaled,0} \right\}} \right)$$

Combining these standardized indices into an array, $\Phi \in \mathbb{R}^{10\times 20}$

$$\Phi = \begin{bmatrix} \tilde{\varphi}_{A,volume} & \tilde{\varphi}_{S,volume} & \cdots & \tilde{\varphi}_{G,volume} & \tilde{\varphi}_{P,volume} \\ \tilde{\varphi}_{A,weight} & & & & \tilde{\varphi}_{P,weight} \\ \vdots & & \ddots & & \vdots \\ \tilde{\varphi}_{A,\ n.p.-area} & & & & \tilde{\varphi}_{P,\ n.p.-area} \\ \tilde{\varphi}_{A,flexibility} & \tilde{\varphi}_{S,flexibility} & \cdots & \tilde{\varphi}_{G,flexibility} & \tilde{\varphi}_{P,flexibility} \end{bmatrix}_{10\times 20}$$

PHYSICAL PROPERTY PREFERENCE CALCULATIONS

We can produce an array, $\Psi \in \mathbb{R}^{129\times 10}$, with entries representing each position's preference for a given physical property by the following operation:

$$\Psi = F \cdot \Phi^T$$

With individual entries, $\psi_{p,\epsilon}$, corresponding to a summation of the following product over all 20 canonical amino acids, where ϵ corresponds to a given physical property, and p corresponds to a position in the protein backbone:

$$\psi_{p,\epsilon} = \sum_{i=1}^{20} \tilde{f}_{p,a_i} \cdot \tilde{\varphi}_{a_i,\epsilon}$$

Such that:

$$\Psi = \begin{bmatrix} \left(\sum_{i=1}^{20} \tilde{f}_{1,a_i} \cdot \tilde{\varphi}_{a_i,vol.}\right) & \cdots & \left(\sum_{i=1}^{20} \tilde{f}_{1,a_i} \cdot \tilde{\varphi}_{a_i,flex.}\right) \\ \vdots & \ddots & \vdots \\ \left(\sum_{i=1}^{20} \tilde{f}_{129,a_i} \cdot \tilde{\varphi}_{a_i,vol.}\right) & \cdots & \left(\sum_{i=1}^{20} \tilde{f}_{129,a_i} \cdot \tilde{\varphi}_{a_i,flex.}\right) \end{bmatrix}_{129\times 10}$$

$$= \begin{bmatrix} \left(\tilde{f}_{A,1} \cdot \tilde{\varphi}_{A,vol.} + \cdots + \tilde{f}_{P,1} \cdot \tilde{\varphi}_{P,vol.}\right) & \cdots & \left(\tilde{f}_{A,1} \cdot \tilde{\varphi}_{A,flex.} + \cdots + \tilde{f}_{P,1} \cdot \tilde{\varphi}_{P,flex.}\right) \\ \vdots & \ddots & \vdots \\ \left(\tilde{f}_{A,129} \cdot \tilde{\varphi}_{A,vol.} + \cdots + \tilde{f}_{P,129} \cdot \tilde{\varphi}_{P,vol.}\right) & \cdots & \left(\tilde{f}_{A,129} \cdot \tilde{\varphi}_{A,flex.} + \cdots + \tilde{f}_{P,129} \cdot \tilde{\varphi}_{P,flex.}\right) \end{bmatrix}_{129\times 10}$$

$$= \begin{bmatrix} \psi_{1,volume} & \psi_{1,weight} & \cdots & \psi_{1,n.p.\ area} & \psi_{1,flexibility} \\ \psi_{2,volume} & & & & \psi_{p,flexibility} \\ \vdots & & \ddots & & \vdots \\ \psi_{128,volume} & & & & \psi_{128,flexibility} \\ \psi_{129,volume} & \psi_{129,weight} & \cdots & \psi_{129,n.p\ area} & \psi_{129,flexibility} \end{bmatrix}_{129\times 10}$$

Thus, we have obtained an array containing information about position-wise preferences for various physical properties, wherein rows index to the positions in the protein backbone, and columns index the various physical properties in question.

2.E. References

1. Rocklin, A. G. J., Chidyausiku, T. M., Goreshnik, I., Ford, A., Lemak, A., Carter, L., Ravichandran, R., Mulligan, V. K., Arrowsmith, C. H. & Baker, D. Global analysis of protein folding using massively parallel design, synthesis and testing. *Science*. **357**, 168–175 (2017).
2. Hsia, Y., Bale, J. B., Gonen, S., Shi, D., Sheffler, W., Fong, K. K., Nattermann, U., Xu, C., Huang, P.-S., Ravichandran, R., Yi, S., Davis, T. N., Gonen, T., King, N. P. & Baker, D. Design of a hyperstable 60-subunit protein icosahedron. *Nature* **535**, 136–139 (2016).
3. Bale, J. B., Gonen, S., Liu, Y., Sheffler, W., Ellis, D., Thomas, C., Cascio, D., Yeates, T. O., Gonen, T., King, N. P. & Baker, D. Accurate design of megadalton-scale two-component icosahedral protein complexes. *Science*. **353**, 389–394 (2016).
4. Jorda, J., Leibly, D. J., Thompson, M. C. & Yeates, T. O. Structure of a novel 13 nm dodecahedral nanocage assembled from a redesigned bacterial microcompartment shell protein. *Chem. Commun.* **52**, 5041–5044 (2016).
5. Lai, Y.-T., Reading, E., Hura, G. L., Tsai, K.-L., Laganowsky, A., Asturias, F. J., Tainer, J. A., Robinson, C. V. & Yeates, T. O. Structure of a designed protein cage that self-assembles into a highly porous cube. *Nat. Chem.* **6**, 1065–1071 (2014).
6. Azuma, Y., Zschoche, R., Tinzl, M. & Hilvert, D. Quantitative Packaging of Active Enzymes into a Protein Cage. *Angew. Chemie - Int. Ed.* **55**, 1531–1534 (2016).
7. Frey, R., Mantri, S., Rocca, M. & Hilvert, D. Bottom-up Construction of a Primordial Carboxysome Mimic. *J. Am. Chem. Soc.* **138**, 10072–10075 (2016).
8. Eriksson, A. E., Baase, W. A., Zhang, X.-J., Heinz, D. W., Blaber, M., Baldwin, E. P. & Matthews, B. W. Response of a Protein Structure to Cavity-Creating Mutations and Its Relation to the Hydrophobic Effect. *Science*. **255**, 178–183 (1992).
9. Cordonnier, A., Montagnier, L. & Emerman, M. Single amino-acid changes in HIV envelope affect viral tropism and receptor binding. *Nature* **340**, 571–4 (1989).
10. Asensio, M. A., Morella, N. M., Jakobson, C. M., Hartman, E. C., Glasgow, J. E., Sankaran, B., Zwart, P. H. & Tullman-Ercek, D. A Selection for Assembly Reveals That a Single Amino Acid Mutant of the Bacteriophage MS2 Coat Protein Forms a Smaller Virus-like Particle. *Nano Lett.* **16**, 5944–5950 (2016).
11. Wright, S. The roles of mutation, inbreeding, crossbreeding and selection in evolution. *Proceedings of the Sixth International Congress on Genetics* **1**, 356–366 (1932).
12. Romero, P. A. & Arnold, F. H. Exploring protein fitness landscapes by directed evolution. *Nat. Rev. Mol. Cell Biol.* **10**, 866–876 (2009).
13. Maynard Smith, J. Natural Selection and the Concept of a Protein Space. *Nature* **225**, 563–564 (1970).
14. Firnberg, E., Labonte, J. W., Gray, J. J. & Ostermeier, M. A comprehensive, high-resolution map of a gene's fitness landscape. *Mol. Biol. Evol.* **31**, 1581–1592 (2014).
15. Wrenbeck, E. E., Azouz, L. R. & Whitehead, T. A. Single-mutation fitness landscapes for an enzyme on multiple substrates reveal specificity is globally encoded. *Nat. Commun*. **8**, 1–10 (2017).
16. Fowler, D. M., Araya, C. L., Fleishman, S. J., Kellogg, E. H., Stephany, J. J., Baker, D. & Fields, S. High-resolution mapping of protein sequence-function relationships. *Nat. Methods* **7**, 741–746 (2010).
17. Tinberg, C. E., Khare, S. D., Dou, J., Doyle, L., Nelson, J. W., Schena, A., Jankowski, W., Kalodimos, C. G., Johnsson, K., Stoddard, B. L. & Baker, D. Computational design of ligand-binding proteins with high affinity and selectivity. *Nature* **501**, 212–6 (2013).
18. Roscoe, B. P., Thayer, K. M., Zeldovich, K. B., Fushman, D. & Bolon, D. N. A. Analyses of the effects

of all ubiquitin point mutants on yeast growth rate. *J. Mol. Biol.* **425**, 1363–1377 (2013).

19. Sarkisyan, K. S., Bolotin, D. A., Margarita, V., Mamedov, I. Z., Tawfik, D. S. & Lukyanov, K. A. Local fitness landscape of the green fluorescent protein. *Nature* **533**, 397–401 (2016).
20. Al-Mawsawi, L. Q., Wu, N. C., Olson, C. A., Shi, V. C., Qi, H., Zheng, X., Wu, T.-T. & Sun, R. High-throughput profiling of point mutations across the HIV-1 genome. *Retrovirology* **11**, 124 (2014).
21. Ferguson, A. L., Mann, J. K., Omarjee, S., Ndung, T. & Walker, B. D. Resource Translating HIV Sequences into Quantitative Fitness Landscapes Predicts Viral Vulnerabilities for Rational Immunogen Design. *Immunity* **38**, 606–617 (2012).
22. Lauring, A. S. & Andino, R. Exploring the Fitness Landscape of an RNA Virus by Using a Universal Barcode Microarray. *J. Virol.* **85**, 3780–3791 (2011).
23. Visher, E., Whitefield, S. E., McCrone, J. T., Fitzsimmons, W. & Lauring, A. S. The Mutational Robustness of Influenza A Virus. *PLoS Pathog*. **12**, e1005856 (2016).
24. Betancourt, A. J. Genomewide patterns of substitution in adaptively evolving populations of the RNA bacteriophage MS2. *Genetics* **181**, 1535–1544 (2009).
25. Acevedo, A., Brodsky, L. & Andino, R. Mutational and fitness landscapes of an RNA virus revealed through population sequencing. *Nature* **505**, 686–690 (2013).
26. Butterfield, G. L., Lajoie, M. J., Gustafson, H. H., Sellers, D. L., Nattermann, U., Ellis, D., Bale, J. B., Ke, S., Lenz, G. H., Yehdego, A., Ravichandran, R., Pun, S. H., King, N. P. & Baker, D. Evolution of a designed protein assembly encapsulating its own RNA genome. *Nature* **552**, 415–420 (2017).
27. Qi, H., Wu, N. C., Du, Y., Wu, T. T. & Sun, R. High-resolution genetic profile of viral genomes: why it matters. *Curr. Opin. Virol.* **14**, 62–70 (2015).
28. Wu, W., Hsiao, S. C., Carrico, Z. M. & Francis, M. B. Genome-free viral capsids as multivalent carriers for taxol delivery. *Angew. Chemie - Int. Ed.* **48**, 9493–9497 (2009).
29. Carrico, Z. M., Romanini, D. W., Mehl, R. A. & Francis, M. B. Oxidative coupling of peptides to a virus capsid containing unnatural amino acids. *Chem. Commun. (Camb).* 1205–1207 (2008).
30. Peabody, D. S. A Viral Platform for Chemical Modification and Multivalent Display. *J. Nanobiotechnology* **1**, 5 (2003).
31. Garcea, R. L. & Gissmann, L. Virus-like particles as vaccines and vessels for the delivery of small molecules. *Curr. Opin. Biotechnol.* **15**, 513–517 (2004).
32. Strauss, J. H. & Sinsheimer, R. L. Purification and properties of bacteriophage MS2 and of its ribonucleic acid. *J. Mol. Biol.* **7**, 43–54 (1963).
33. Dai, X., Li, Z., Lai, M., Shu, S., Du, Y., Zhou, Z. H. & Sun, R. *In situ* structures of the genome and genome-delivery apparatus in a single-stranded RNA virus. *Nature* **541**, 112–116 (2016).
34. Ni, C. Z., Syed, R., Kodandapani, R., Wickersham, J., Peabody, D. S. & Ely, K. R. Crystal structure of the MS2 coat protein dimer: implications for RNA binding and virus assembly. *Structure* **3**, 255–63 (1995).
35. Ashley, C. E., Carnes, E. C., Phillips, G. K., Durfee, P. N., Buley, M. D., Lino, C. A., Padilla, D. P., Phillips, B., Carter, M. B., Willman, C. L., Brinker, C. J., Caldeira, J. D. C., Chackerian, B., Wharton, W. & Peabody, D. S. Cell-specific delivery of diverse cargos by bacteriophage MS2 virus-like particles. *ACS Nano* **5**, 5729–5745 (2011).
36. Galaway, F. A. & Stockley, P. G. MS2 viruslike particles: A robust, semisynthetic targeted drug delivery platform. *Mol. Pharm.* **10**, 59–68 (2013).
37. Giessen, T. W. & Silver, P. A. A Catalytic Nanoreactor Based on *in Vivo* Encapsulation of Multiple Enzymes in an Engineered Protein Nanocompartment. *ChemBioChem* **17**, 1931–1935 (2016).

38. Glasgow, J. E., Capehart, S. L., Francis, M. B. & Tullman-Ercek, D. Osmolyte-mediated encapsulation of proteins inside MS2 viral capsids. *ACS Nano* **6**, 8658–8664 (2012).

39. Glasgow, J. & Tullman-Ercek, D. Production and applications of engineered viral capsids. *Appl. Microbiol. Biotechnol.* **98**, 5847–58 (2014).

40. Hooker, J. M., Kovacs, E. W. & Francis, M. B. Interior Surface Modification of Bacteriophage MS2. *J. Am. Chem. Soc.* **126**, 3718–3719 (2004).

41. Caldeira, J. C. & Peabody, D. S. Thermal stability of RNA phage virus-like particles displaying foreign peptides. *J. Nanobiotechnology* **9**, 22 (2011).

42. Farkas, M. E., Aanei, I. L., Behrens, C. R., Tong, G. J., Murphy, S. T., Neil, J. P. O. & Francis, M. B. PET Imaging and Biodistribution of Chemically Modified Bacteriophage MS2. *Mol. Pharm.* **10**, 69–76 (2013).

43. Aanei, I. L., Elsohly, A. M., Farkas, M. E., Netirojjanakul, C., Regan, M., Taylor Murphy, S., O'Neil, J. P., Seo, Y. & Francis, M. B. Biodistribution of antibody-MS2 viral capsid conjugates in breast cancer models. *Mol. Pharm.* **13**, 3764–3772 (2016).

44. Stockley, P. G., Rolfsson, O., Thompson, G. S., Basnak, G., Francese, S., Stonehouse, N. J., Homans, S. W. & Ashcroft, A. E. A Simple, RNA-Mediated Allosteric Switch Controls the Pathway to Formation of a T=3 Viral Capsid. *J. Mol. Biol.* **369**, 541–552 (2007).

45. Zhang, L., Sun, Y., Chang, L., Jia, T., Wang, G., Zhang, R., Zhang, K. & Li, J. A novel method to produce armored double-stranded DNA by encapsulation of MS2 viral capsids. *Appl. Microbiol. Biotechnol.* **99**, 7047–7057 (2015).

46. Glasgow, J. E., Asensio, M. A., Jakobson, C. M., Francis, M. B. & Tullman-Ercek, D. Influence of Electrostatics on Small Molecule Flux through a Protein Nanoreactor. *ACS Synth. Biol.* **4**, 1011–1019 (2015).

47. Capehart, S. L., Coyle, M. P., Glasgow, J. E. & Francis, M. B. Controlled Integration of Gold Nanoparticles and Organic Fluorophores Using Synthetically Modified MS2 Viral Capsids. *J. Am. Chem. Soc.* **135**, 3011–3016 (2013).

48. Hietpas, R. T., Jensen, J. D. & Bolon, D. N. A. Experimental illumination of a fitness landscape. *Proc. Natl. Acad. Sci. U. S. A.* **108**, 7896–7901 (2011).

49. Stewart, J. J., Lee, C. Y., Ibrahim, S., Watts, P., Shlomchik, M., Weigert, M. & Litwin, S. A Shannon entropy analysis of immunoglobulin and T cell receptor. *Mol. Immunol.* **34**, 1067–1082 (1997).

50. Stephanopoulos, N., Tong, G. J., Hsiao, S. C. & Francis, M. B. Dual-surface modified virus capsids for targeted delivery of photodynamic agents to cancer cells. *ACS Nano* **4**, 6014–6020 (2010).

51. Caspar, D. L. & Klug, A. Physical principles in the construction of regular viruses. *Cold Spring Harb. Symp. Quant. Biol.* **27**, 1–24 (1962).

52. Rolfsson, O., Toropova, K., Ranson, N. A. & Stockley, P. G. Mutually-induced conformational switching of RNA and coat protein underpins efficient assembly of a viral capsid. *J. Mol. Biol.* **401**, 309–322 (2010).

53. Lago, H., Fonseca, S. A., Murray, J. B., Stonehouse, N. J. & Stockley, P. G. Dissecting the key recognition features of the MS2 bacteriophage translational repression complex. *Nucleic Acids Res.* **26**, 1337–1344 (1998).

54. LeCuyer, K. A., Behlen, L. S. & Uhlenbeck, O. C. Mutants of the bacteriophage MS2 coat protein that alter its cooperative binding to RNA. *Biochemistry* **34**, 10600–6 (1995).

55. Go, M. & Miyazawa, S. Relationship between mutability, polarity, and exteriority of amino acid residues in protein. *Int. J. Pept. Protein Res.* **15**, 211–224 (1980).

56. Krissinel, E. & Henrick, K. Inference of Macromolecular Assemblies from Crystalline State. *J. Mol. Biol.* **372**, 774–797 (2007).

57. Stockley, P. G., White, S. J., Dykeman, E., Manfield, I., Rolfsson, O., Patel, N., Bingham, R., Barker, A., Wroblewski, E., Chandler-Bostock, R., Weiß, E. U., Ranson, N. A., Tuma, R. & Twarock, R. Bacteriophage MS2 Genomic RNA Encodes an Assembly Instruction Manual for Its Capsid. *Bacteriophage* **6**, e1157666 (2016).

58. Peabody, D. S. The RNA binding site of bacteriophage MS2 coat protein. *EMBO J.* **12**, 595–600 (1993).

59. Peabody, D. S. & Chakerian, A. Asymmetric contributions to RNA binding by the Thr45 residues of the MS2 coat protein dimer. *J. Biol. Chem.* **274**, 25403–25410 (1999).

60. Ashley, C. E., Carnes, E. C., Phillips, G. K., Durfee, P. N., Buley, M. D., Lino, C. A., Padilla, D. P., Phillips, B., Carter, M. B., Willman, C. L., Brinker, C. J., Caldeira, J. D. C., Chackerian, B., Wharton, W. & Peabody, D. S. Cell-specific delivery of diverse cargos by bacteriophage MS2 virus-like particles. *ACS Nano* **5**, 5729–5745 (2011).

61. Engler, C., Gruetzner, R., Kandzia, R. & Marillonnet, S. Golden gate shuffling: a one-pot DNA shuffling method based on type IIs restriction enzymes. *PLoS One* **4**, e5553 (2009).

62. Liu, H. & Naismith, J. H. An efficient one-step site-directed deletion, insertion, single and multiple-site plasmid mutagenesis protocol. *BMC Biotechnol.* **8**, 91 (2008).

63. Don, R. H., Cox, P. T., Wainwright, B. J., Baker, K. & Mattick, J. S. 'Touchdown' PCR to circumvent spurious priming during gene amplification. *Nucleic Acids Res*. **19**, 4008 (1991).

64. Pinto, F., Thapper, A., Sontheim, W. & Lindblad, P. Analysis of current and alternative phenol based RNA extraction methodologies for cyanobacteria. *BMC Mol. Biol.* **10**, 79 (2009).

65. Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: A flexible trimmer for Illumina sequence data. *Bioinformatics* **30**, 2114–2120 (2014).

66. Magoč, T. & Salzberg, S. L. FLASH: Fast length adjustment of short reads to improve genome assemblies. *Bioinformatics* **27**, 2957–2963 (2011).

67. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754–1760 (2009).

68. Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G. & Durbin, R. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).

Chapter 3: Experimental evaluation of coevolution in a self-assembling particle

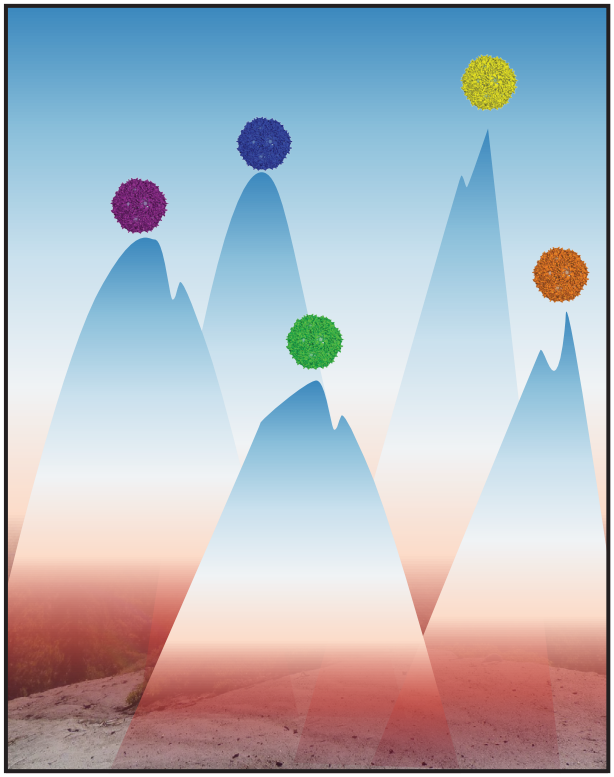
The following is adapted from Hartman, Lobba, Favor, Robinson, Francis, and Tullman-Ercek; Biochemistry, 2019 with permission

Short summary

The outcome of a given mutation changes depending on the genetic background. To study this phenomena, we mapped an additional dimension of the fitness landscape of the MS2 coat protein, generating and characterizing a library of two amino acid mutants.

Abstract

Protein evolution occurs via restricted evolutionary paths that are influenced by both previous and subsequent mutations. This effect, termed epistasis, is critical in population genetics, drug resistance, and immune escape; however, the effect of epistasis on the level of protein fitness is less well characterized. We generated and characterized a 6,615-member library of all two amino acid combinations in a highly mutable loop of a virus-like particle. This particle is a model of protein self-assembly and a promising vehicle for drug delivery and imaging. In addition to characterizing the effect of all double mutants on assembly, thermostability, and acid stability, we observed many instances of epistasis, where combinations of mutations are either more deleterious or more beneficial than expected. These results were used to generate rules governing the effects of multiple mutations on the self-assembly of the virus-like particle.



3.A. Introduction

Protein evolution occurs through complex pathways, often involving nonintuitive leaps between functional variants[1–3]. These paths include local minima and maxima, in which the effect of a given mutation depends entirely on the previous and subsequent mutation[1]. This effect, known as epistasis, is well-studied in population genetics[4–6] and is known to play a central role in drug resistance[7,8] and immune escape[4,6,9]. However, studies that quantify the combinatorial effect of multiple mutations on protein fitness remain relatively rare.

Much effort has focused on characterizing the fitness effect of single mutations on a given protein, producing one-dimensional protein fitness landscapes[10–15] (see **Chapter 2**). While such landscapes are highly useful for describing the effects of one amino acid mutations on a protein, these efforts do not capture the multi-dimensional shape (or ruggedness) of evolutionary landscapes. To date, epistasis has been measured for GFP[16], RNA-binding proteins[17], and several enzymes[2,18,19]. These studies find that instances of negative epistasis—where a secondary mutation is more deleterious than anticipated—are more common than positive epistasis, though both are detectable and play a role in shaping protein fitness landscapes[16,17,20–22].

Complex, multimeric protein scaffolds such as viral capsids, metabolosomes, and other molecular machines are poised to have a significant impact on biotechnology in the coming decades[23,24]. To maximize this potential, it is important to understand how non-native functions can be hindered by unanticipated epistatic effects. To date, our understanding of the design rules governing the self-assembly of these proteins remains limited, complicating the use, predictability, and yield of these particles in non-native contexts[25,26]. In particular, we expect the effects of epistasis to be especially significant in large assemblies with quaternary structure due to many inter- and intra-monomer interactions[27,28].

Here, we characterize the complete pairwise epistasis of a highly mutable loop occurring in a virus-like particle (VLP) (**Figure 3.1**). VLPs are closed-shell protein containers derived from noninfectious viral shell proteins. MS2 bacteriophage is used as a model VLP, as it is well-studied for use in drug delivery[29–31], disease imaging[32–34], vaccine

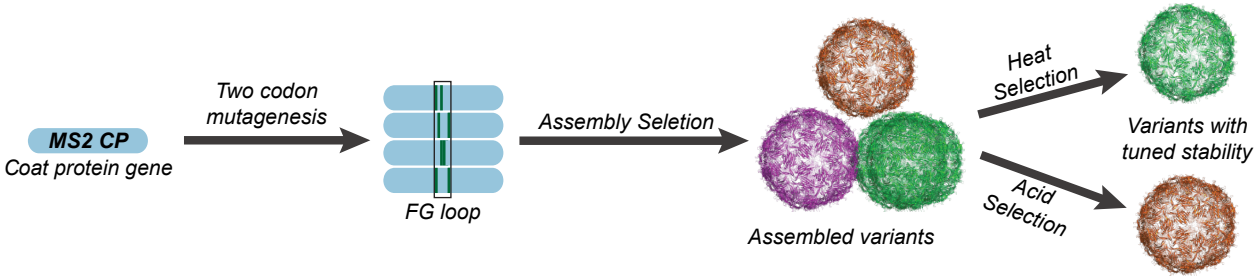


Figure 3.1. FG loop mutagenesis and selection strategy. Two-codon mutagenesis targeted at the FG loop generated a library of 6,615 variants, each with exactly two NNK substitutions in this region. These variants were subjected to an assembly selection, followed by targeted functional selections to identify heat-stable and acid-sensitive variants.

development[35,36], and biomaterials[37,38]. To do this we harness a technique developed in our labs, called SyMAPS[15] (Systematic Mutation and Assembled Particle Selection), to

evaluate how epistasis shapes the assembly, thermostability, and acid stability of the MS2 coat protein (CP). Based on our selection criteria, we find many instances of both positive and negative epistasis for the loop region studied, governed largely by charge and steric bulk. Our studies reveal two residues, distant by sequence but spatially adjacent, that show strong pairwise epistasis. This allows us to describe unexpected design rules governing the mutability of this loop. Moreover, this work establishes a useful experimental protocol that can be used to understand how epistatic effects can be leveraged to obtain new particles with desired physical or chemical features.

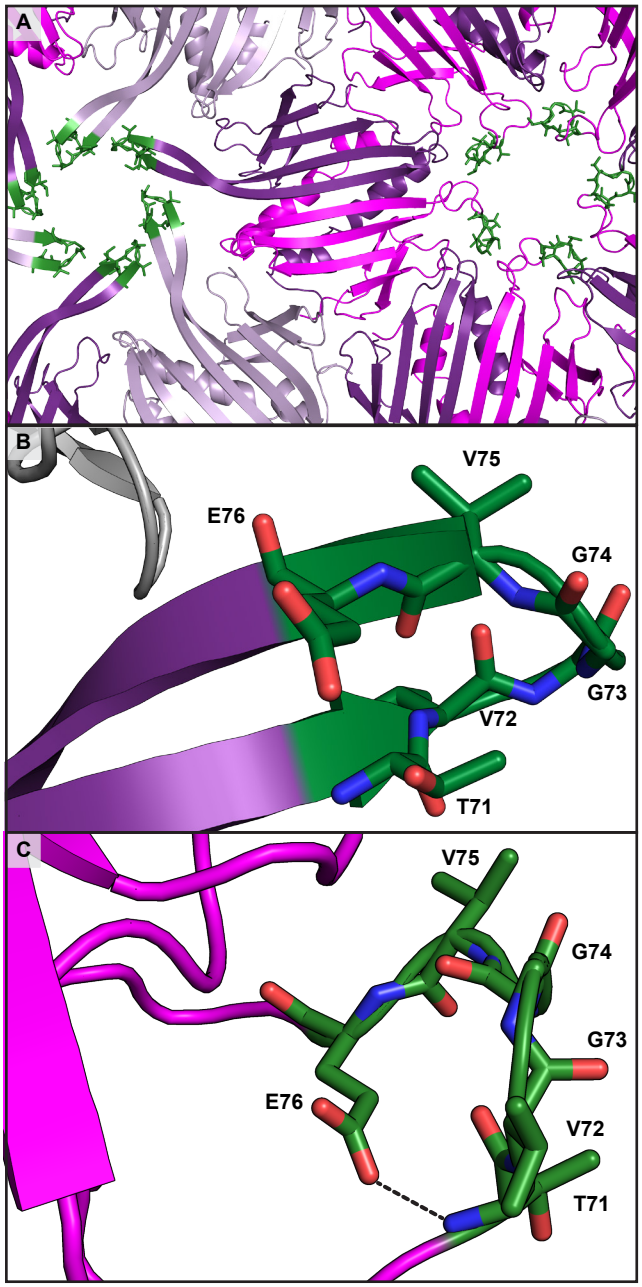


Figure 3.2. Changes in the FG loop of the MS2 CP. A) The view from the interior of the capsid shows how the FG loop, indicated in green, differs between the quasi-six-fold and five-fold axes. *A*-type monomers are in dark purple, *B* monomers are in magenta, and *C* monomers are in light purple. Close-up perspectives of the B) quasi-six-fold and C) five-fold axes highlight the structural changes between the A/C and B-form monomers, respectively. A key hydrogen bond between E76 and the backbone of T71 is indicated with a dashed line in (C).

3.B. Results and Discussion

We recently described the One Dimensional Apparent Fitness Landscape (1D-AFL) of the MS2 CP[15]. In this fitness landscape, we evaluated the mutability at each position across the MS2 CP. While the MS2 CP has been extensively used as an epitope display platform, mutations and peptide insertions are typically performed on an exterior-facing loop between residue 14 and 19[39,40]. In our previous study, we found that a six-amino-acid stretch in a flexible loop connecting two β-sheets, referred to as the FG loop, was highly mutable (**Figure 3.2A**), meaning that many amino acid substitutions assembled into well-formed VLPs. This FG loop undergoes a critical conformational shift during VLP assembly[41–43], which results in two distinct structures near the pore, termed the A/C (**Figure 3.2B**) and B (**Figure 3.2C**) forms. Because a conformational change in this loop is important for VLP assembly[41–43], we were surprised by the mutability of this region.

This loop was used as a model system to study how two amino acid mutations affect protein fitness, thereby characterizing a second dimension of the

MS2 CP protein fitness landscape. To evaluate the fitness of all variants in this library, we used SyMAPS, a technique that generates a quantitative score of assembly-competency across a targeted library of VLP variants. When expressed in *E. coli*, well-formed particles will encapsulate available negative charge within the MS2 CP during assembly[44]. SyMAPS takes advantage of this property, using intrinsic nucleic acid encapsulation as a convenient genotype-to-phenotype link. If a given MS2 CP mutation assembles, then variant mRNA is encapsulated within the VLP and copurifies with it[15]. If assembly is not permitted with a given mutation, then cellular nucleic acids are not recovered.

We generated a 6,615-member library containing all possible two amino acid combinations in the FG loop (T71–E76). This library contained all single and double amino acid mutations from the native MS2 CP sequence. We subjected the library to a selection based on VLP assembly and performed high-throughput sequencing before and after the selection. The library was generated and analyzed in three independent replicates. The percent abundance of each variant before and after the selection was converted into an Apparent Fitness Score (AFS). Library members with positive AFS values correspond to mutations that permit assembly. Conversely, negative AFS values indicate disfavored VLP formation, which could be due to poor expression, inefficient or no assembly, or instability to protein purification. These scores are presented together in a combined Two-Dimensional Apparent Fitness Landscape (2D-AFL), which indicates the effects of all one and two amino acid mutations in this loop on VLP assembly (**Supplementary Fig. 3.1**).

Of the 6,615 possible library members, over 92% of all variants were identified in at least one replicate, and 87% were identified in all three replicates. About 5% of the variants were sequenced in the plasmid library but were absent in the VLP library; these variants are indicated in dark red on the 2D-AFL. Nonsense and silent mutations were used as internal negative and positive controls, respectively. In this study 483 nonsense mutations were measured, and all had an Apparent Fitness Score (AFS) value of –0.19 or lower, correctly identifying each variant as nonassembling. In contrast, 15 silent mutations—or, one per combination—were measured, and all of these AFS values were 0.20 or higher. This indicates that these wild-type VLPs are correctly identified as well-assembled. As expected, these two populations separated into two non-overlapping groups (**Figure 3.3A**), affirming the quality of these data.

3.B.i. Selection for thermal stability enriches wild-type-like variants

Thermostability is desirable in nearly all potential applications, and any variants used as vaccines, biomaterials, or drug delivery vehicles would likely require near-wild-type stability or better. To identify variants that are stable to high temperatures, we subjected library members to a heat challenge of 50 °C for 10 min. We then compared the percent abundance of variants after this selection compared to the starting plasmid library, resulting in a heat-selected 2D-AFL (**Supplementary Fig. 3.2**). Variants that assemble and are stable to 50 °C for 10 min result in positive scores and are indicated in blue.

We compared the distribution of AFS values in the assembly-selected and heat-selected 2D-AFLs. A histogram of all AFS values in the less stringent, assembly-selected case results in a broad distribution centered slightly larger than zero with a long negative tail (**Figure 3.3B**). In contrast, a histogram of the AFS values in the heat-selected case

exhibits bimodality, suggesting that many assembly-competent variants strongly alter the thermostability of the MS2 CP. Variants with a high AFS value in the heat-selected 2D-AFL are likely to be best suited for applications such as drug delivery or imaging due to their uncompromised thermal stability. The bimodality of the heat-selected dataset is also evident in the dark colors and obvious striped patterns in the landscape itself (**Supplementary Fig. 3.2**). For example, with few exceptions, mutations at G74 resulted in negative scores in the heat-selected 2D-AFL, indicating that nearly every combination of mutations at this position resulted in undesirable VLP properties. Similar effects can be seen with V75. At V75, mutation to isoleucine and, in some cases, leucine resulted in thermostable VLPs, but few other mutations were tolerated. Taken together, the stringency of the thermal selection is useful for identifying which variants behave like wild-type VLPs.

We hypothesized that one and two missense mutations may have different average effects on protein fitness. In this dataset, 570 single amino acid mutants were scored, while 5041 two amino acid mutations were scored. When the AFS values of one and two missense mutations were compared, differences between these populations were apparent (**Figure 3.3C**). The histogram comparing these values clearly shows that one amino acid mutations form a bimodal distribution, indicating the VLPs split into well-assembled and poorly-assembled populations. In contrast, AFS values for two amino acid mutations are distributed more evenly and are lower on average, suggesting that an additional mutation is more often detrimental to VLP assembly. These differences were exacerbated by the heat selection, in which two amino acid mutations were clearly less tolerated than one amino acid mutations (**Figure 3.3D**), though both populations exhibit bimodality. These results agree with literature reports that a second mutation—or, an additional step away from wild type in a fitness landscape—

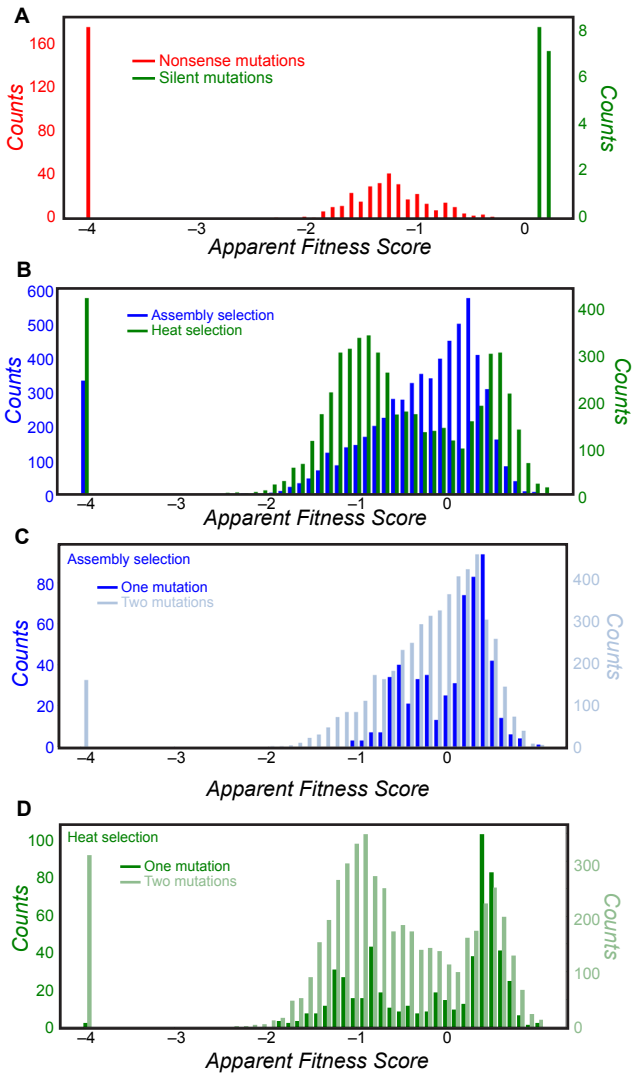


Figure 3.3. Assembly-selected Apparent Fitness Score (AFS) abundances. A) Nonsense and silent mutation AFS values separate into two non-overlapping populations. B) The heat selected 2D-AFL separates into bimodal populations, while the assembly selection 2D-AFL does not. Single amino acid mutations are on average less deleterious than two amino acid mutations in both C) the assembly and D) heat selections.

often has an additive, negative effect on protein fitness[16].

3.B.ii. Selections for increased acid sensitivity with uncompromised thermal stability

In targeted drug delivery, a therapeutic cargo can be protected inside of a container until it is endocytosed into target cells; thus, selective release of drug cargo in the acidic environment of endosomes or lysosomes is potentially advantageous. In addition to its role in VLP assembly, the FG loop is critical for modulating the acid stability of the MS2 CP. Previously, we used a 1D-AFLs to show that mutations CP[T71H] and CP[E76C] increase VLP sensitivity to an acidic environment[15]. However, previous attempts to improve the acid sensitivity of CP[T71H] or CP[E76C] through rational design led to compromised long-term stability, suggesting that the properties of protein stability and acid sensitivity may be intertwined in a non-obvious fashion. As such, identifying variants with inversely-correlated properties—high thermostability with reduced acid stability—is well-suited for protein engineering approaches.

We therefore additionally selected for variants with stability to high temperatures and increased acid sensitivity. These selections were performed on the assembled library, ensuring that variants are competent for assembly. In addition, we specifically sought variants that behaved similarly to the previously-published CP[T71H] variant, which selectively precipitates under acidic conditions[15]. We challenged the assembly-selected library of FG loop variants to pH 5 at 37 ºC for 1 h, mimicking the conditions of the early endosome[45]. High-throughput sequencing was used to identify VLPs that were selectively absent following acidic pressure (**Supplementary Fig. 3.3**).

We screened nine variants that appeared to be depleted after the acid selection but enriched following the thermal selection. In contrast, all instances of silent mutations that encode for CP[WT] showed unchanged abundance following acidic pressure and increased relative abundance following thermal pressure (**Supplementary Fig. 3.4**). We incubated these nine variants overnight in acidic conditions (pH 4.6 and pH 3.6), then quantified the amount of assembled VLP remaining in the supernatant **Supplementary Fig. 3.5**). Several variants appeared to be depleted at pH 4.6, and these were then screened overnight at a wider range of pHs, from 3.9 to 5.3 (**Supplementary Fig. 3.6**). From these assays, three (CP[T71H / E76P], CP[T71H / E76Q], and CP[T71H / E76T]) were determined to exhibit the greatest acid sensitivity compared to the parent CP[T71H] variant (**Figure 3.4A,B**), although a range of pH sensitivities were found. All three of the most acid-sensitive variants contained the parent CP[T71H] mutation[15], combined with a second mutation at residue 76. Gratifyingly, at pH 7.2, all three variants exhibited a melting temperature higher than 50 °C, ranging from 52 °C to 65 °C (**Figure 3.4C**).

CP[T71H/E76P], the most acid sensitive variant, tolerated an additional cysteine mutation in the interior cavity at position N87. This cysteine mutation has previously been used to load fluorophore or drug cargo into the interior of the MS2 CP[46]. The triple mutant CP[T71H/E76P/N87C] formed well-assembled VLPs and was readily modified by AlexaFluor-488 maleimide (**Supplementary Fig. 3.7A**); in contrast, the CP[T71H/E76P] double mutant lacking the introduced cysteine residue remained unmodified. Analysis by HPLC SEC confirmed that the AlexaFluor488 fluorophore coelutes with CP[T71H/E76P/N87C] following modification (**Supplementary Fig. 3.7B,C**), consistent with the behavior

of CP[T71H/N87C] and CP[N87C] (**Supplementary Fig. 3.5D–G**) and indicating that the modified VLPs are assembled.

We next sought to use the internal cysteine to study the behavior of CP[T7H / E76P / N87C]. A better understanding of variant behavior in acidic conditions could provide insight into whether these VLPs would release protein cargo in the acidic environment of the endosome.

We sought to use Förster resonance energy transfer (FRET) to study the behavior of the variant in acid. The interior cysteine in both CP[N87C] and CP[T71H / E76P / N87C] positioned at an ideal distance for FRET, as five cysteines are positioned within 4 nm of one another (**Supplementary Fig. 3.8**). Thus, we labeled the interior cysteine in both CP[N87C] and CP[T71H / E76P / N87C] with AlexaFluor 647 fluorophore and a quencher pair, Tide 5WS. The two compounds were labeled at low concentration of fluorophore and high concentration of quencher. With this experimental setup, we expected that disassociation into dimer would result in a fluorescent burst as the fluorophore and quencher separated from one another. However, no change in fluorescence was detected when variants were subjected to acidic conditions.

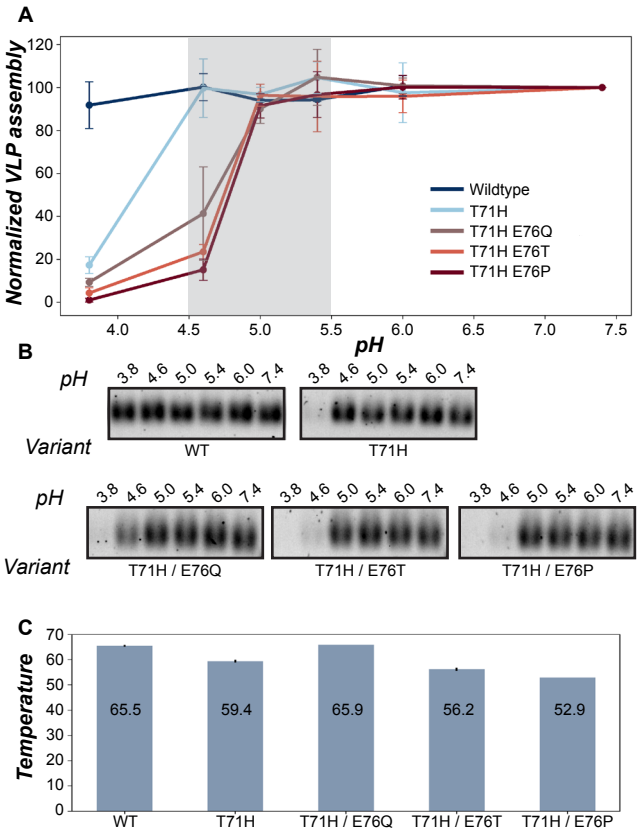


Figure 3.4. Validation of acid-sensitive, heat-stable variants. A,B) Three new variants, CP[T71H/E76Q], CP[T71H/E76T], and CP[T71H/E76P] are more acid sensitive than CP[T71H]. Capsid recoveries after acid challenges were quantified via native gel electrophoresis, followed by ImageJ densitometry analysis. Error bars represent three sample replicates. C) All variants exhibit melting temperatures higher than 50 °C at pH 7.2.

To better understand the lack of fluorescence burst, we designed a set of controls to confirm effective fluorophore quenching and quantify background fluorescence in these experiments. We labeled CP[N87C] and CP[T71H / E76P / N87C] with increasing amounts of fluorophore while keeping the quencher and protein concentration constant (**Supplementary Fig. 3.9A,C**). A no-protein control was also included (**Supplementary Fig. 3.9E**). In this experiment, any fluorescence in the no-protein control is a result of background fluorophore that was not removed by the purification protocol. In contrast, any increase in fluorescence in the VLP cases compared to the no-protein control are a result of incomplete quenching by FRET. We found nearly identical fluorescence between the three experiments (**Supplementary Fig. 3.9A,C,E**), indicating that the labeled fluorophore is completely quenched, and that any detected fluorescence

Table 3.1

Res	M.I.
71	-0.04
72	-0.05
73	-0.01
74	-0.22
75	-0.22
76	-0.09

is likely due to free fluorophore in solution.

We performed the opposite experiment, in which fluorophore and protein concentration were kept constant but quencher concentration during labeling increased from 0 μM to 50 μM. Again, a no-protein control was used as a comparison. We found dramatic differences in the 0 μM quencher case when comparing labeled VLPs to the no-protein control, indicating that the signal from unquenched fluorophores is far above background (**Supplementary Fig. 3.9B,D,F**). Furthermore, fluorophore-modified VLPs were readily quenched with even small amounts of the quencher, as even 5 μM quencher significantly reduced the fluorescence signal. Finally, in the no-protein control, increasing quencher did not affect fluorescence, indicating that free quencher in solution is not disrupting these experiments.

Overall, this set of experiments suggest that the interior fluorophore is effectively quenched via FRET by even small amounts of quencher; however, free fluorophore, likely not fully removed by multiple rounds of spin concentration, is leading to background signal in these experiments. From these results, we hypothesize that the lack of fluorescence burst under acidic conditions suggests that the VLP does not fully dissociate into dimers. However, whether the protein aggregated, dissociated on a time scale that is faster than can be measured using this technique, or dissociated into larger fragments like pentamers, hexamers, or other capsomeres remains to be determined.

With these results, it is difficult to conclude whether the acid sensitive VLPs would be able to release therapeutic cargo in an acidic environment. While acid sensitivity is a highly desired property, cargo release is critical for this property to be useful in drug delivery. Nonetheless, combining highly-targeted libraries with direct functional selections successfully identified variants with a narrow yet desirable combination of properties, a feat that would be challenging with many other protein engineering techniques. Taken together, we conclude that this technique is a promising method for tuning the physical and chemical properties of VLPs.

3.B.iii. Quantifying FG loop pairwise mutability using Shannon entropy

The previously-published 1D-AFL was used to quantify the Mutability Index (MI) of each position in the FG loop (**Table 3.1**). The region contains a range of mutabilities, from poorly mutable (G74, V75) to highly mutable (T71, V72, G73, and E76). In this study, Shannon entropy[47], a calculation that measures diversity at a given position, was used to quantify the Pairwise Mutability at every combination of positions in both the assembly- and heat-selected 2D-AFL. As expected, positions with lower mutability, such as G74, reduced the pairwise mutability, even when combined with positions with high mutability, such as T71 or E76 (**Supplementary Fig. 3.10A**). Residues T71, V72, G73, and E76 are independently highly mutable with similar Mutability Indices, as determined the 1D-AFL. However, pairwise combinations containing E76 were less mutable than pairwise combinations of T71, V72, and G73, suggesting that multiple mutations carry increased penalty at position E76.

Differences in Pairwise Mutability were evaluated following the heat selection (**Supplementary Fig. 3.10B**). In particular, residue V75 pairwise mutability decreased more than other positions, suggesting that combinations of mutations that include V75

may permit assembly but lead to a loss of thermostability. Only the pairwise combination of T71 and E76 inverted from mutable to immutable following thermal pressure, suggesting that mutations at both of these positions may be more deleterious to thermostability.

We also performed a correlation analyses to evaluate reproducibility in this dataset (**Supplementary Fig. 3.11, Supplementary Data 3.1**). In the assembly and acid selections, replicates one and two agreed better than either with replicate three, indicating that some technical error may exist between biological replicates (**Supplementary Fig. 3.11A-F**). Generally, R^2 values for the assembly and acid selections are lower than anticipated (0.25 to 0.53). Interestingly, the heat selections show much better correlation between the replicates, with R^2 values ranging from 0.69 to 0.73 and few off-axis datapoints, which would indicate variants that behave differently between replicates (**Supplementary Fig. 3.11G-I**). Taken together, these data suggest that the assembly and acid selections may be lower stringency selections, resulting in more variation between replicates, particularly with regard to variants that form unstable VLPs. From these results, we can conclude that the heat-selected AFL is best suited for researchers who are looking to select variants for further study, as these are most likely to form reproducibly stable, well-assembled VLPs.

We next sought to predict two-dimensional mutability in this region using a convolutional neural network based on the 1D-AFL (**Supplementary Fig. 3.12A**). The optimized convolutional neural network produced a 2D-AFL in which the average mutability in each combination was comparable to the assembly-selected 2D-AFL; however, individual combinations of amino acids were inconsistently predicted. Upon further analysis, we found that an additive 2D-AFL—which was populated with the summed difference between the AFS value of both mutations and the average AFS_{WT} in the 1D-AFL—outperformed the neural network in predicting the mutability of each combination (**Supplementary Fig. 3.12B**). We hypothesize that this performance discrepancy is due to the limited training data, which consisted of only 2,580 mutants in the parent 1D-AFL library. These efforts underscore the importance of continued experimental work to generate high-quality, multi-dimensional AFLs.

3.B.iv. Epistasis plays a visible role in two-amino-acid mutability across the FG loop

We sought to identify instances of negative and positive sign epistasis in the 2D-AFL. Negative sign epistasis refers to combinations of mutations that do not permit assembly, even though each mutation is permitted individually. Conversely, positive sign epistasis refers to mutations that rescue non-assembling variants, resulting in assembled VLPs.

We identified variants for which the predicted 2D-AFS value (calculated via the simple additive method described above) is a different sign from the measured assembly-selected 2D-AFS value. We quantified epistasis, *E*, as the difference between the predicted and experimental datasets (**Figure 3.5A**). While the majority of these variants have *E* scores closer to zero, a subset exhibits notably positive and negative sign epistasis (**Supplementary Data 3.2**). The median *E* value is negative, consistent with previous studies of sign epistasis, which have shown that deleterious pairwise interactions are more common[48].

Although much of the experimental landscape matches the predicted 2D-AFL, we found that epistasis plays a surprising role in the two-amino-acid mutability across many

positions in this loop. In particular, positions T71 and E76, which are spatially adjacent but distant in sequence (**Figure 3.2B,C**), coevolve with a strong interdependence. Mutations at T71 change which mutations are permitted at E76, and vice versa, with regard to both charge and steric bulk. In addition, combinations of charged mutations are tightly regulated by epistatic effects across the entire loop, in which some evolutionary paths are more available than others.

As examples of this, both positive and negative sign epistasis were observed at combinations of mutations at position T71 and E76, both in the assembly- and heat-selected 2D-AFL (**Figure 3.5B**). While distant in sequence, these positions are structurally close in both the A/C and B conformations. In the A/C conformation (quasi-6-fold axis), T71 and E76 are on adjacent β-sheets, while in the B conformation (5-fold axis), the side chain of E76 hydrogen bonds with the backbone of T71 (**Figure 3.2C**)[49,50].

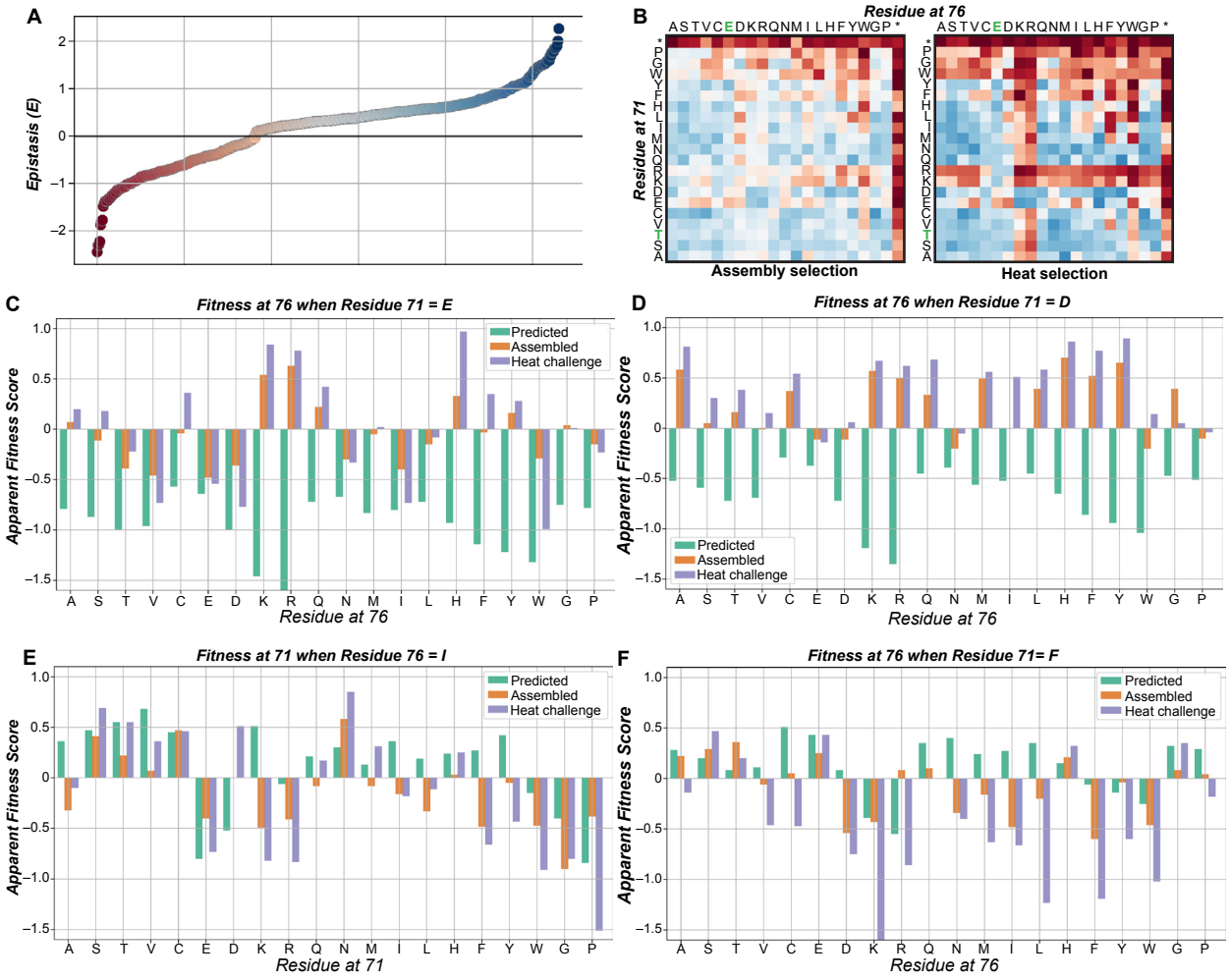


Figure 3.5. Positive and negative epistasis in the FG loop. A) *E*, a measure of epistasis, shows both negative and positive effects, though the overall trend is toward negative epistasis. B) Residues 71 and 76 show significant positive and negative epistasis, as shown for both the assembly and heat selections. Examples of (C, D) predominantly positive epistasis and (E, F) negative epistasis are shown. In these graphs, 2D AFS values predicted from the 1D AFL data are shown in green and compared to the measured assembly-selected (orange) and heat-selected (purple) AFS values. Differences between the predicted and measured scores indicate regions of epistatic interactions.

In the 1D-AFL, a single negative charge at position T71 is not permitted. We previously hypothesized that this is likely due to repulsion with the nearby negative charge at E76. This hypothesis is supported by the 2D-AFL results: T71E can be rescued by a charge inversion at E76, and CP[T71E/E76K], CP[T71E/E76R] and CP[T71E/E76H] are assembly-competent and enriched following the thermal challenge (**Figure 3.5C**). More strikingly, T71D variants were rescued by almost any mutation at E76, with the notable exception of negatively charged residues (D, E) and structurally disruptive residues (G, P) (**Figure 3.5D**).

Visualization in Chimera yielded useful insight into the origin of this pattern. In CP[WT], E76 hydrogen bonds both with Q40 and the backbone of T71 (**Supplementary Fig. 3.13A**) in the B form monomer structure. When E76 is inverted from a hydrogen acceptor to a hydrogen donor (R,K), hydrogen bonding with Q40 could be preserved, as glutamine contains both an acceptor and donor. However, hydrogen bonding between E76 and the backbone of T71 is likely not preserved without significant backbone rearrangement. In addition, in the case of CP[T71E / E76R], the mutated side chains are oriented in opposite directions, again indicating that a new salt bridge is likely not formed without backbone rearrangement in this region (**Supplementary Fig. 3.13B**). Finally, *in silico* mutation to CP[E76R] results in clashes with Q70, further supporting the idea that backbone rearrangement of this flexible loop is likely. Taken together, we anticipate that doubly charged mutants at T71 and E76 are engaging in backbone rearrangement in the B form, thus restoring either a mimic of the native hydrogen bonding pattern or permitting a salt bridge between variant side chains.

Instances of negative sign epistasis were also observed between position T71 and E76. In the 1D-AFL, each position independently permits hydrophobic mutations. For example, CP[E76I] is assembly-competent; however, when coupled with an additional mutation of isoleucine, leucine, phenylalanine, or tyrosine at T71—all mutations that are tolerated individually—the VLP is no longer assembly-competent (**Figure 3.5E**). Similarly, CP[T71F] is permitted in the 1D-AFL, but when combined with additional hydrophobic mutations at position E76, the VLP no longer permits assembly (**Figure 3.5F**).

These trends suggest steric constraint between these two residues, where enough space exists for one but not multiple bulky amino acids. Visualization in Chimera led us to hypothesize that steric constraint is driven by the structure of the A/C form monomers rather than the B form monomers. In the C form, and to a lesser extent in the A form, clashes between the two bulky residues were apparent following *in silico* mutation to CP[T71F/E76I] (**Supplementary Fig. 3.13C**), while the B form allowed mutation without producing clashes. Given the secondary structure in this region, we hypothesize that these clashes are not readily resolved, leading to assembly-incompetent VLPs.

Across the FG loop, combinations of oppositely-charged mutations have varying, and often striking, effects (**Supplementary Fig. 3.14A**). Negative charge at T71 is rescued by positive charge at V72, G73, or E76. Similar, though subtler, effects are seen for negative charge at V72 with corresponding positive charge at T71 or G73, and combinations of oppositely charged mutations at these positions lead to well-assembled, thermostable VLPs. Similarly, charge inversion at E76, which is not tolerated in the 1D-AFL, is rescued in several instances with negative charge elsewhere in the loop.

A different trend is observed with charged mutations at positions V72 in combination with G74 or V75, both poorly mutable positions in general. While assembly of VLPs with charge at V72 is rescued by a second, oppositely-charged mutation at G74 or V75, these mutations lead to thermally unstable VLPs that do not remain assembled through the heat challenge. Similar trends are observed when charge at G74 and V75 is combined with an oppositely-charged mutation at E76. While these trends are challenging to explain in full, we visualized the hydrogen bond network and local environment of the FG loop in both the A/C and B form to gather more insight (**Supplementary Fig. 3.13D,E**). Within a 5 Å region, the A/C form largely interacts with other regions of nearby FG loop, albeit a wider region than was evaluated in this study. In contrast, the FG loop of the B form is spatially near a wide range of residues, including an adjacent loop region. Several key hydrogen bonds appear to maintain the correct geometric shape in this loop. As such, we hypothesize that the strong effect of charge on the FG loop is driven by the hydrogen bonding network in the B form monomer.

Other instances of sign epistasis were less intuitive. We observed nonobvious positive epistasis between residues G73 and V75. A tyrosine mutation at V75 is not tolerated when G73 is wild type. However, surprisingly, positive charge or hydrophobic mutations, including K, R, F, L, M, and Y, rescue assembly and thermostability of V75Y (**Supplementary Fig. 3.13B**). How positive charge rescues a mutation to tyrosine is unclear. One explanation could be that the positive charge forces a new interaction with the negatively-charged E76 residue, contorting the loop enough to allow the bulky V75Y mutation. An alternative possibility is a new cation–pi interaction between V75Y and K or R, or pi-stacking with F and Y.

The only two paired residues that show no evidence of epistasis are G74 and V75, the least mutable residues in this region. Of the 400 possible combinations of mutations at G74 and V75, only CP[G74G/V75V], CP[G74G/V75I], and CP[G74G/V75L] form thermostable VLPs.

Taken together, these detailed analyses work toward defining the design rules for successful VLP formation of the MS2 CP, adding depth to our understanding of self-assembly.

3.B.v. Phylogenic analysis suggests relationship between epistasis and evolutionary path

Epistatic analysis produced several compelling design rules governing the mutability—and, potentially, evolvability—of the FG loop. We sought to compare these rules against sequences of homologous coat proteins from related bacteriophages. A phylogenetic tree from sequenced RNA phages was generated, and these twenty-four bacteriophage coat proteins separated into three distinct clades (**Figure 3.6A**). This tree is similar to previously-published phylogenetic analyses of ssRNA phages[51]. Consistent with the design rules generated by epistatic analysis, most homologues have zero or one negatively-charged residue in this loop. Additional negative residues are frequently accompanied by positive residues, and residues corresponding to 71 and 76 are not both large, hydrophobic amino acids.

Bacteriophages closely related to the MS2 CP show high sequence similarity in the

region corresponding to the FG loop. No charged residues are observed other than E76. Additionally, residues T71, V72, and G73 exhibit the most sequence diversity, consistent with our mutability analyses. In contrast, a clade of structurally-related coat proteins that includes the well-studied Qβ Bacteriophage[26] shows a divergent consensus sequence (**Figure 3.6B**). In this case, the negative charge at E76 is replaced with a cysteine residue. Interestingly, this cysteine is known to participate in inter-subunit disulfide bonding. In addition, exposure to DTT reduces the melting temperature of the Qβ CP by over 40 ºC, indicating that this disulfide bond is critical for thermostability[52]. Within this group, we also see more sequence divergence at residues 71–75, including several charged residues at positions 72 and 74. The third clade, which is most distant from the MS2 CP, shows high divergence from the MS CP sequence, with the exception of the conserved G74 residue. In this group, most sequences have only one negative charge, primarily at a position corresponding to residue 71 in the MS2 CP.

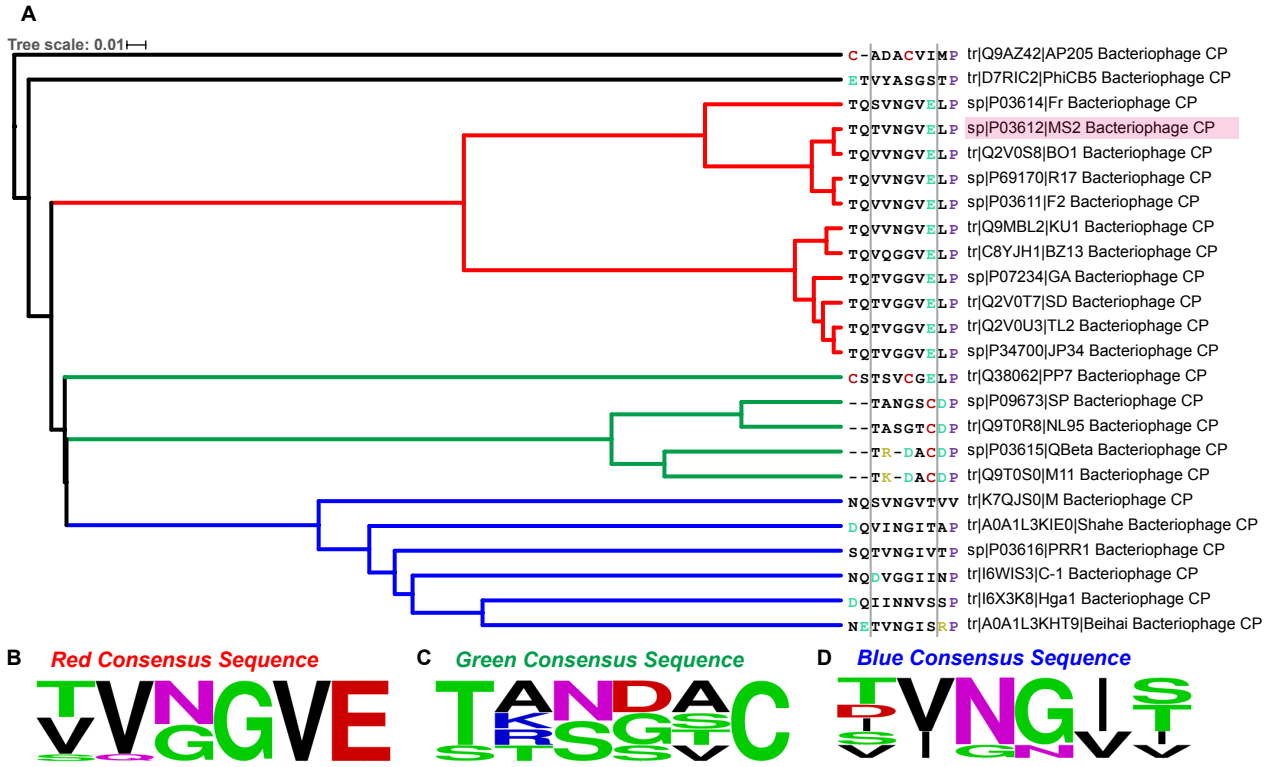


Figure 3.6. Phylogenetic tree of RNA bacteriophage coat proteins. A) Twenty-four coat proteins separated into three distinct clades, indicated in red, blue, and green. The gray lines indicate the regions that are analogous to the FG loop in bacteriophage MS2. B,C,D) These clades exhibit distinct consensus sequences.

From these analyses, we hypothesize that the observed effect of charge on the MS2 CP FG loop arises in the absence of disulfide bonds present in the clade containing Qβ. Without inter-subunit disulfide bonds, the MS2 CP likely relies on the hydrogen bond network in the B form monomer—including the intra-subunit hydrogen bond between residues T71 and E76—to maintain correct geometric shape, and disrupting this interaction with additional charge, charge inversion, or other mutations can result in poorly-formed VLPs. However, compensating with a similar hydrogen bond through an additional mutation, or

balancing additional charge with an opposite charge, can be restorative.

3.C. Conclusion

VLPs are an excellent model system to study the effects of epistasis on protein assembly, due to their genetic simplicity, high yield, and intrinsic genotype-to-phenotype link. We generated and characterized a 6,615-member library of one and two amino acid mutations in the highly mutable FG loop of the MS2 CP. The library was subjected to multiple selections, initially for capsid assembly, followed by thermal stability and acid sensitivity. Negative or positive epistasis were identified and characterized. In particular, two amino acid mutations involving charged residues and steric bulk coevolved in unexpected ways. Our epistatic analysis was used to generate a set of design rules for this loop, which were compared to consensus sequences for related coat protein clades in a phylogenetic analysis. This study represents the first quantitative measure of epistasis in a self-assembling nanoparticle, and design rules generated will inform future efforts to tailor and engineer virus-like particles in a variety of non-native contexts.

3.D. Methods

3.D.i. Strains

MegaX DH10B *E. coli* electrocompetent cells (ThermoFisher Scientific, Cat# C640003) were used for all library experiments, and DH10B chemically competent cells produced in-house were used for expression of individual variants of interest. Overnights from a single colony were grown for 16-20 h at 37 °C shaking at 200 RPM in LB-Miller media (Fisher Scientific, Cat# BP1426-2) with chloramphenicol at 32 mg/L. Expressions were subcultured 1:100 into 2xYT media (Teknova, Cat# Y0210) with 32 mg/L chloramphenicol and allowed to express overnight at 37 °C shaking at 200 RPM.

3.D.ii. Library generation

To generate libraries with two amino acid mutations in the FG loop, we modified a library generation strategy developed by the Bolon lab, known as EMPIRIC cloning[53]. EMPIRIC cloning uses a plasmid with a self-encoded removable fragment (SERF) surrounded by inverted BsaI restriction sites. With this set-up, BsaI digestion simultaneously removes both SERF and BsaI sites. These plasmids are referred to as entry vectors, and the SERF in this study encodes constitutively-expressed GFP to permit green/white screening. We used a previously-described entry vector that replaced a 26-codon segment flanking the FG loop in the MS2 CP with the SERF[15]. Single stranded DNA primers with all fifteen combinations of degenerate codons were purchased, enabling overlap extension PCR to generate double stranded DNA with all possible pairwise combinations in this 6-residue region. These primers were resuspended in water, pooled, and diluted to a final concentration of 50 ng/μL. The reverse strand was filled in by overlap extension PCR with a corresponding forward primer. The amplified, double-stranded DNA was purified using a PCR Clean-up Kit (Promega, Cat# A9282). The purified DNA was diluted to 1-5 ng/μL and cloned into the described entry vector using established Golden gate cloning techniques[54]. The ligated plasmids were desalted on membranes (Millipore Sigma, Cat# VSWP02500) for 20 min, then transformed into MegaX DH10B *E. coli* electrocompetent

cells (ThermoFisher Scientific, Cat# C640003). Following electroporation and recovery, cells were plated onto two large LB-A plates (VWR, Cat# 82050-600) with 32 µg/mL chloramphenicol and allowed to grow at 37 ºC overnight. Colony number varied, but every transformation yielded a number of colonies that was at least 50x the library size. This protocol was repeated in full for three total biological replicates that are fully independent from library generation through selection.

3.D.iii. Size selection

Colonies were scraped from plates into LB-M and allowed to grow for 2 h. Each library was then subcultured 1:100 into 1 L of 2xYT (Teknova, Cat: Y0210) and allowed to grow to an OD600 of 0.6, when they were induced by 0.1 percent arabinose. Libraries of variants were expressed overnight at 37 °C. Cultures were then harvested, resuspended in 10 mM phosphate buffer at pH 7.2 with 2 mM sodium azide, and sonicated. Libraries were subjected to two rounds of 50% w/v ammonium sulfate precipitation, followed by FPLC size exclusion chromatography purification to select for well-formed VLPs.

3.D.iv. FPLC SEC (Assembly selection)

MS2 CP libraries or individual variants were purified on an Akta Pure 25 L Fast Protein Liquid Chromatography (FPLC) system with a HiPrep Sephacryl S-500 HR column (GE Healthcare Life Sciences, Cat# 28935607) Size Exclusion Chromatography (SEC) column via isocratic flow with 10 mM phosphate buffer at pH 7.2, 200 mM sodium chloride, and 2 mM sodium azide. Fractions containing MS2 coat protein were collected for further analysis.

3.D.v. Heat selection

Following the assembly selection (FPLC purification), libraries were incubated at 50 ºC for 10 min. Buffer from FPLC purification (10 mM phosphate buffer at pH 7.2, 200 mM sodium chloride, and 2 mM sodium azide) was used for these studies. Precipitated VLPs were pelleted by centrifugation, and well-formed VLPs were isolated by semi-preparative HPLC SEC. Fractions containing VLP were combined and subjected to RNA extraction, barcoding, and high-throughput sequencing.

3.D.vi. HPLC SEC

MS2 CP variants were analyzed on an Agilent 1290 Infinity HPLC with an Agilent Bio SEC-5 column (5 µm, 2000 Å, 7.8x300 mm) with isocratic flow of 10 mM phosphate buffer, pH 7.2, 200 mM sodium chloride, and 2 mM sodium azide. Fractions were collected at the characteristic elution time for wild-type MS2 (11.2 min) and subjected to RNA extraction and high-throughput sequencing sample preparation.

3.D.vii. Acid selection

Libraries were incubated at pH 5 or 7 for 1 h at 37 ºC, prepared with citric acid and sodium phosphate according to the Sigma Aldrich Buffer Reference Center. Precipitated VLPs were pelleted by centrifugation, and intact VLPs were concentrated using a 100 kDa molecular weight spin cutoff filter (Millipore Sigma, Cat# UFC510024). The supernatant

of the MWCO filter was subjected to RNA extraction, barcoding, and high-throughput sequencing as described above.

3.D.viii. Sample prep for high-throughput sequencing

Plasmid DNA was extracted prior to expressions using a Zyppy Plasmid Miniprep Kit (Zymo, Cat# D4036). RNA was extracted from the MS2 CP library following assembly selections using previously-published protocols[55]. Briefly, TRIzol (Thermo Fisher Cat# 15596026) was used to homogenize samples, followed by chloroform addition. The sample was separated by centrifugation into aqueous, interphase, and organic layers. The aqueous layer, which contained RNA, was isolated, and the RNA was then precipitated with isopropanol and washed with 70% ethanol. RNA was then briefly dried and resuspended in RNAse free water. cDNA was then synthesized using the Superscript III first strand cDNA synthesis kit from Life (cat: 18080051, polyT primer). cDNA and plasmids were both amplified with two rounds of PCR to add barcodes (10 cycles) and the Illumina sequencing handles (8 cycles), respectively, following Illumina 16S Metagenomic Sequencing Library Preparation recommendations (**Supplementary Data 3.3**). Libraries were combined and analyzed by 150 PE MiSeq in collaboration with the UC Davis Sequencing Facilities. Reads in excess of 18 million passed filter, and an overall Q30 > 85%.

3.D.ix. Individual variant cloning

Individual variants were cloned using a variation on the method described above. Briefly, overlap extension PCR (**Supplementary Data 3.3**) yielded a double stranded fragment that spanned the length of the missing 26-codon region in the Entry Vector. Each fragment was cloned into the Entry Vector using standard Golden gate cloning techniques[54]. Variants bearing the CP[N87C] mutation were cloned into a similar Entry Vector bearing the desired mutation at position 87, which was installed via site-directed mutagenesis[56]. Cloned plasmids were transformed into chemically-competent DH10B cells. Individual colonies were sequenced via Sanger sequencing prior to expression.

3.D.x. Individual variant expression

Selected mutants were individually expressed in 5 or 50 mL cultures of 2xYT. These expressions were pelleted, resuspended in 10 mM phosphate buffer at pH 7.2 and 2 mM sodium azide, lysed by sonication, precipitated twice with 50% w/v ammonium sulfate, and evaluated by native gel for VLP formation and acid sensitivity.

3.D.xi. Individual acid screens

Following ammonium sulfate precipitation and resuspension, variants were diluted 1:10 into neutral or acidic buffers ranging from pH 3.9 to 7.4. These buffers contained various concentrations of citric acid and phosphate, prepared according to the Sigma Buffer Reference. Variants were centrifuged at 13,000g for 2 min, then equal volumes were loaded onto a native gel. Densitometry with ImageJ was used to determine VLP formation and sensitivity to acidic conditions.

3.D.xii. Native gel

VLPs were analyzed in a 0.8% agarose gel in 0.5X TBE buffer (45 mM Tris-borate, 1 mM EDTA) with 2X SYBR Safe DNA Gel Stain (ThermoFisher Scientific, Cat# S33102) for 120 min at 40 V. Agarose gels were imaged on a BioRad GelDoc EZ Imager. Densitometry in each condition compared to pH 7.4 was carried out using ImageJ.

3.D.xiii. Sypro Orange Melting Curves

Individual variants were purified by HPLC SEC, then diluted to a final A280 of 1 in 10 mM phosphate buffer at pH 7.2, 200 mM sodium chloride, and 2 mM sodium azide. Purified variants were filtered, and a final concentration of 20x Sypro Orange was added. Melting temperature was determined using a qPCR protocol that measured fluorescence every 0.5 ºC between 30 ºC and 80 ºC using an Applied Biosystems QuantStudio 3. The derivative was taken of the resulting fluorescence curves, and the minimum value was determined to be the melting temperature.

3.D.xiv. Fluorophore and quencher maleimide modification

Variants of interest were purified by HPLC SEC, diluted to 5 µM in 20 mM phosphate buffer at pH 7.2. A solution of AlexaFluor488 maleimide (ThermoFisher Scientific, Cat# A10254) in DMF was added to a final concentration of 20X (relative to capsid monomer) and allowed to react for 1 h. Variants were spin concentrated 3x with a centrifugal filter with a 100 kDa molecular weight cutoff (Millipore Sigma, Cat# UFC510024), then evaluated by HPLC SEC and ESI TOF.

Quencher–fluorophore experiments were conducted as described above with minor modification. Tide Quencher™ 5WS maleimide (AAT Bioquest, Cat# 2079) and AlexaFluor647 maleimide (ThermoFisher Scientific, Cat# A20347) were combined in varied ratios (0 µM, 5 µM, 10 µM, 20 µM and 50 µM quencher with 10 µM fluorophore, as well as switching concentrations of fluorophore and quencher) with 5 µM VLP or a no-protein control. Variants were allowed to react for one hour, then spin concentrated 3x with a centrifugal filter with a 100 kDa molecular weight cutoff (Millipore Sigma, Cat# UFC510024). Fluorescence emission spectra were collected on a Fluoromax-4 Spectrofluorometer (HORIBA Scientific) exciting at 650 nm with a 5 nm window and monitoring emission from 650 to 700 nm.

3.D.xv. Mass spectrometry.

Modified and unmodified proteins were analyzed with an Agilent 1200 series liquid chromatograph (Agilent Technologies, USA) connected in-line with an Agilent 6224 Time-of Flight (TOF) LC/MS system with a Turbospray ion source.

3.D.xvi. High-throughput sequencing data analysis

Data were trimmed and processed as previously described[15] with minor variation. Briefly, data were trimmed with Trimmomatic[57] with a 4-unit sliding quality window of 20 and a minimum length of 30. Reads were merged using FLASH (Fast Length Adjustment of SHort reads)[58] with a maximum overlap of 160 base pairs. Reads were then aligned to the wild-type MS2 CP reference gene with Burrows-Wheeler Aligner (BWA-MEM)[59]. Reads were then sorted and indexed with Samtools[60]. The Picard function CleanSam was

used to filter unmapped reads, and reads longer or shorter than the expected length of the barcoded DNA were removed.

3.D.xvii. AFL calculations

Cleaned and filtered high-throughput sequencing reads were analyzed using Python programs written in-house. Briefly, the mutated region of the MS2 CP was isolated, and the number of mutations per read was calculated. Reads with zero mutations (wild-type reads) or greater than two mutations were both removed. In reads with two mutations, the two non-wild-type codons were identified and counted. In reads with one mutation, the mutated codon was tallied in combination with every wild-type codon. Codons were then translated into amino acids, removing codons that do not end in G or T.

These calculations were repeated for all experiments to generate abundances before and after each selective pressure. Relative percent abundances were calculated as previously described[15]. Briefly, the grand sum, or the sum of all counts at every combination at every position, was calculated. We next divided each matrix by its grand sum, generating a matrix of percent abundances. These calculations were repeated for each biological replicate of VLP, plasmid, heat-selected, or acid-selected libraries, generating twelve Percent Abundance matrices. We calculated relative percent abundances by dividing the percent abundance for the selected library compared by the percent abundance for the plasmid library for each replicate.

We calculated the mean across the three replicates. All Nan (null) values, which indicate variants that were not identified in the plasmid library, were ignored. Scores of zero, which indicate variants that were sequenced in the unselected library but absent in the VLP library, were replaced with an arbitrary score of 0.0001. We calculated the log10 of the Relative Percent Abundance array to calculate the final array for each replicate. Finally, we calculated the average AFS value for each amino acid combination by finding the mean value for every combination, which is displayed in **Supplementary Fig. 3.1**. In addition, all AFS values for assembly and heat selections can be found in **Supplementary Data 3.1**. Correlation analyses were performed between the replicates, excluding any variant with nan or –4 in any replicate. Raw plasmid abundances, AFS values for each replicate, and correlation analyses can be found in **Supplementary Data 3.1**.

3.D.xviii. Shannon entropy calculations

Shannon entropy is defined as:

$$ShannonEntropy = -\sum P \log(P) \qquad \text{(eq. 1)}$$

where P refers to a given probability. We first calculate the probability of a given combination of two amino acids occurring within a single residue. Any zero values are replaced with 0.00001. We then calculated the Shannon entropy at all fifteen combinations as follows:

Shannon entropy values were averaged across three biological replicates for each library. The difference between the Unselected plasmid library and the VLP library, or the unselected plasmid library compared to the heat-selected VLP library, was used to

evaluate how mutability affects thermostability.

3.D.xix. Predicted 2D-AFL: Convolutional Neural Network

All neural network model development was conducted in Python using the Tensorflow library[61]. The neural network design was composed of two sequential convolutional layers—each followed by a pooling operation—that ultimately fed into two fully-connected layers, which output a scalar fitness score prediction. Data were fed in as an array consisting of 12 physical properties listed for each amino acid position along the MS2 backbone[62–66]. Mutations were modeled by swapping the physical properties of one amino acid for another at the relevant site in the MS2 CP. Mutants with fitness scores greater than or equal to –4 were removed, and missing data points were excluded entirely.

The property-columns in our input matrices are separately fed into the function, processed individually, then combined. The first convolutional layer takes small fragments of the input vector (kernel size = 5), corresponding to 5-residue long sequences of amino acids in the backbone of the MS2 CP. This length was chosen to represent small units of sequence that can exhibit characteristic patterns in their physical identities without overburdening the training model. After each pass through the filters, a "pooling" operation reduces the size of the data passed along by taking the maximum value of each 2-unit long subdivision of the filter outputs, and consolidating them to feed into the next layer. The neural net was trained on the full 1D-AFL using mean squared error as our optimization factor, with a max number of iterations set to 50,000.

3.D.xx. Epistasis calculations

Epistasis, *E*, was calculated as described elsewhere[16]. Briefly, we calculated the difference between non-additive effects of one amino acid mutations in log scale. We first calculated the difference between the AFS value of a mutation (AFS_i) and the average AFS_{WT}. $AFS_{predicted}$ was calculated by adding the Δ_i values for each mutation to the average AFS_{WT}. This predicted AFS values for all two amino acid variants are plotted in **Supplementary Fig. 3.12B**. The predicted 2D-AFS value was then subtracted from the measured AFS value to generate *E*, a measure of epistasis.

$$\Delta_i = AFS_i - AFS_{WT} \quad \text{(eq. 2)}$$

$$AFS_{predicted} = (\sum_i \Delta_i) + AFS_{WT} \quad \text{(eq. 3)}$$

$$E = AFS_{measured} - AFS_{predicted} \quad \text{(eq. 4)}$$

Variants where the sign (+, –) between the Predicted 2D-AFL and measured 2D-AFL was inverted were separated for further analysis. *E* values for these variants were plotted in **Figure 3.5A**. Epistasis scores are available in **Supplementary data 3.2**.

3.D.xxi. Phylogeny calculations

Bacteriophage coat proteins related to the MS2 CP (PDB ID: 2MS2) were identified and aligned with UniProt Align, and a phylogenetic tree was calculated using Interactive

Tree of Life (iTOL)[67]. Consensus sequences were generated using Berkeley WebLogo[68].

3.E. References

1. Sailer, Z. R. & Harms, M. J. High-order epistasis shapes evolutionary trajectories. *PLoS Comput. Biol*. **13**, e1005541 (2017).
2. Steinberg, B. & Ostermeier, M. Environmental changes bridge evolutionary valleys. *Sci. Adv.* **2**, e1500921–e1500921 (2016).
3. Miton, C. M. & Tokuriki, N. How mutational epistasis impairs predictability in protein evolution and design. *Protein Sci.* **25**, 1260–1272 (2016).
4. Gong, L. I. & Bloom, J. D. Epistatically Interacting Substitutions Are Enriched during Adaptive Protein Evolution. *PLoS Genet*. **10**, e1004328 (2014).
5. Gupta, A. & Adami, C. Strong Selection Significantly Increases Epistatic Interactions in the Long-Term Evolution of a Protein. *PLOS Genet*. **12**, e1005960 (2016).
6. Wu, N. C., Du, Y., Le, S., Young, A. P., Zhang, T.-H., Wang, Y., Zhou, J., Yoshizawa, J. M., Dong, L., Li, X., Wu, T.-T. & Sun, R. Coupling high-throughput genetics with phylogenetic information reveals an epistatic interaction on the influenza A virus M segment. *BMC Genomics* **17**, 46 (2016).
7. Weinreich, D. M., Delaney, N. F., DePristo, M. A. & Hartl, D. L. Darwinian evolution can follow only very few mutational paths to fitter proteins. *Science* **312**, 111–114 (2006).
8. Tenaillon, O., Rodríguez-Verdugo, A., Gaut, R. L., McDonald, P., Bennett, A. F., Long, A. D. & Gaut, B. S. The Molecular Diverstiy of Adaptive Convergence. *Science* **335**, 457–462 (2012).
9. Kryazhimskiy, S., Dushoff, J., Bazykin, G. A. & Plotkin, J. B. Prevalence of Epistasis in the Evolution of Influenza A Surface Proteins. *PLoS Genet.* **7**, e1001301 (2011).
10. Firnberg, E., Labonte, J. W., Gray, J. J. & Ostermeier, M. A comprehensive, high-resolution map of a Gene's fitness landscape. *Mol. Biol. Evol*. **31**, 1581–1592 (2014).
11. Wrenbeck, E. E., Azouz, L. R. & Whitehead, T. A. Single-mutation fitness landscapes for an enzyme on multiple substrates reveal specificity is globally encoded. *Nat. Commun.* **8**, 1–10 (2017).
12. Fowler, D. M., Araya, C. L., Fleishman, S. J., Kellogg, E. H., Stephany, J. J., Baker, D. & Fields, S. High-resolution mapping of protein sequence-function relationships. *Nat. Methods* **7**, 741–746 (2010).
13. Tinberg, C. E., Khare, S. D., Dou, J., Doyle, L., Nelson, J. W., Schena, A., Jankowski, W., Kalodimos, C. G., Johnsson, K., Stoddard, B. L. & Baker, D. Computational design of ligand-binding proteins with high affinity and selectivity. *Nature* **501**, 212–6 (2013).
14. Roscoe, B. P., Thayer, K. M., Zeldovich, K. B., Fushman, D. & Bolon, D. N. A. Analyses of the effects of all ubiquitin point mutants on yeast growth rate. *J. Mol. Biol*. **425**, 1363–1377 (2013).
15. Hartman, E. C., Jakobson, C. M., Favor, A. H., Lobba, M. J., Álvarez-Benedicto, E., Francis, M. B. & Tullman-Ercek, D. Quantitative characterization of all single amino acid variants of a viral capsid-based drug delivery vehicle. *Nat. Commun.* **9**, 1385 (2018).
16. Sarkisyan, K. S., Bolotin, D. A., Meer, M. V., Usmanova, D. R., Mishin, A. S., Sharonov, G. V., Ivankov, D. N., Bozhanova, N. G., Baranov, M. S., Soylemez, O., Bogatyreva, N. S., Vlasov, P. K., Egorov, E. S., Logacheva, M. D., Kondrashov, A. S., Chudakov, D. M., Putintseva, E. V., Mamedov, I. Z., Tawfik, D. S., *et al*. Local fitness landscape of the green fluorescent protein. *Nature* **533**, 397–401 (2016).
17. Melamed, D., Young, D. L., Gamble, C. E., Miller, C. R. & Fields, S. Deep mutational scanning of an RRM domain of the Saccharomyces cerevisiae poly(A)-binding protein. *RNA* **19**, 1537–1551 (2013).
18. Jacquier, H., Birgy, A., Le Nagard, H., Mechulam, Y., Schmitt, E., Glodt, J., Bercot, B., Petit, E., Poulain, J., Barnaud, G., Gros, P.-A. & Tenaillon, O. Capturing the mutational landscape of the beta-lactamase

TEM-1. *Proc. Natl. Acad. Sci.* **110**, 13067–13072 (2013).

19. Bershtein, S., Segal, M., Bekerman, R., Tokuriki, N. & Tawfik, D. S. Robustness–epistasis link shapes the fitness landscape of a randomly drifting protein. *Nature* **444**, 929–932 (2006).

20. Bank, C., Hietpas, R. T., Jensen, J. D. & Bolon, D. N. A. A Systematic Survey of an Intragenic Epistatic Landscape. *Mol. Biol. Evo* **32**, 229–238 (2015).

21. Olson, C. A., Wu, N. C. & Sun, R. A comprehensive biophysical description of pairwise epistasis throughout an entire protein domain. *Curr. Biol*. **24**, 2643–2651 (2014).

22. Li, C., Qian, W., Maclean, C. J. & Zhang, J. The fitness landscape of a tRNA gene. *Science* **352**, 837–840 (2016).

23. Glasgow, J. & Tullman-Ercek, D. Production and applications of engineered viral capsids. *Appl. Microbiol. Biotechnol*. **98**, 5847–58 (2014).

24. Slininger Lee, M. Practical considerations for the encapsulation of multi-enzyme cargos within the bacterial microcompartment for metabolic engineering. *Curr. Opin. Syst. Biol.* **5**, 16–22 (2017).

25. Kanaan, N. M., Sellnow, R. C., Boye, S. L., Coberly, B., Bennett, A., Agbandje-McKenna, M., Sortwell, C. E., Hauswirth, W. W., Boye, S. E. & Manfredsson, F. P. Rationally Engineered AAV Capsids Improve Transduction and Volumetric Spread in the CNS. *Mol. Ther. - Nucleic Acids* **8**, 184–197 (2017).

26. Fiedler, J. D., Higginson, C., Hovlid, M. L., Kislukhin, A. A., Castillejos, A., Manzenrieder, F., Campbell, M. G., Voss, N. R., Potter, C. S., Carragher, B. & Finn, M. G. Engineered mutations change the structure and stability of a virus-like particle. *Biomacromolecules* **13**, 2339–48 (2012).

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

28. Sutter, M., Greber, B., Aussignargues, C. & Kerfeld, C. A. Assembly principles and structure of a 6.5-MDa bacterial microcompartment shell. *Science* **356**, 1293–1297 (2017).

29. Ashley, C. E., Carnes, E. C., Phillips, G. K., Durfee, P. N., Buley, M. D., Lino, C. A., Padilla, D. P., Phillips, B., Carter, M. B., Willman, C. L., Brinker, C. J., Caldeira, J. D. C., Chackerian, B., Wharton, W. & Peabody, D. S. Cell-specific delivery of diverse cargos by bacteriophage MS2 virus-like particles. *ACS Nano* **5**, 5729–5745 (2011).

30. Galaway, F. A. & Stockley, P. G. MS2 viruslike particles: A robust, semisynthetic targeted drug delivery platform. *Mol. Pharm*. **10**, 59–68 (2013).

31. ElSohly, A. M., Netirojjanakul, C., Aanei, I. L., Jager, A., Bendall, S. C., Farkas, M. E., Nolan, G. P. & Francis, M. B. Synthetically Modified Viral Capsids as Versatile Carriers for Use in Antibody-Based Cell Targeting. *Bioconjug. Chem.* **26**, 1590–1596 (2015).

32. Aanei, I. L., Elsohly, A. M., Farkas, M. E., Netirojjanakul, C., Regan, M., Taylor Murphy, S., O'Neil, J. P., Seo, Y. & Francis, M. B. Biodistribution of antibody-MS2 viral capsid conjugates in breast cancer models. *Mol. Pharm.* **13**, 3764–3772 (2016).

33. Farkas, M. E., Aanei, I. L., Behrens, C. R., Tong, G. J., Murphy, S. T., Neil, J. P. O. & Francis, M. B. PET Imaging and Biodistribution of Chemically Modified Bacteriophage MS2. *Mol. Pharm*. **10**, 69–76 (2013).

34. Jeong, K., Netirojjanakul, C., Munch, H. K., Sun, J., Finbloom, J. A., Wemmer, D. E., Pines, A. & Francis, M. B. Targeted Molecular Imaging of Cancer Cells Using MS2-Based 129Xe NMR. *Bioconjug. Chem.* **27**, 1796–1801 (2016).

35. Peabody, D. S., Manifold-Wheeler, B., Medford, A., Jordan, S. K., do Carmo Caldeira, J. & Chackerian, B. Immunogenic Display of Diverse Peptides on Virus-like Particles of RNA Phage MS2. *J. Mol. Biol.* **380**, 252–263 (2008).

36. Tumban, E., Peabody, J., Tyler, M., Peabody, D. S. & Chackerian, B. VLPs Displaying a Single L2 Epitope Induce Broadly Cross-Neutralizing Antibodies against Human Papillomavirus. *PLoS One* **7**, 1–10 (2012).

37. Capehart, S. L., ElSohly, A. M., Obermeyer, A. C. & Francis, M. B. Bioconjugation of Gold Nanoparticles through the Oxidative Coupling of ortho -Aminophenols and Anilines. *Bioconjug. Chem.* **25**, 1888–1892 (2014).

38. Capehart, S. L., Coyle, M. P., Glasgow, J. E. & Francis, M. B. Controlled Integration of Gold Nanoparticles and Organic Fluorophores Using Synthetically Modified MS2 Viral Capsids. *J. Am. Chem. Soc.* **135**, 3011–3016 (2013).

39. Peabody, D. S. Subunit fusion confers tolerance to peptide insertions in a virus coat protein. *Arch. Biochem. Biophys.* **347**, 85–92 (1997).

40. Caldeira, J. C. & Peabody, D. S. Thermal stability of RNA phage virus-like particles displaying foreign peptides. *J. Nanobiotechnology* **9**, 22 (2011).

41. Caspar, D. L. & Klug, A. Physical principles in the construction of regular viruses. *Cold Spring Harb. Symp. Quant. Biol.* **27**, 1–24 (1962).

42. Rolfsson, O., Toropova, K., Ranson, N. A. & Stockley, P. G. Mutually-induced conformational switching of RNA and coat protein underpins efficient assembly of a viral capsid. *J. Mol. Biol.* **401**, 309–322 (2010).

43. Ni, C. Z., Syed, R., Kodandapani, R., Wickersham, J., Peabody, D. S. & Ely, K. R. Crystal structure of the MS2 coat protein dimer: implications for RNA binding and virus assembly. *Structure* **3**, 255–63 (1995).

44. Glasgow, J. E., Capehart, S. L., Francis, M. B. & Tullman-Ercek, D. Osmolyte-mediated encapsulation of proteins inside MS2 viral capsids. *ACS Nano* **6**, 8658–8664 (2012).

45. Meng, F., Cheng, R., Deng, C. & Zhong, Z. Intracellular drug release nanosystems. *Materials Today* **15**, (2012).

46. Stephanopoulos, N., Tong, G. J., Hsiao, S. C. & Francis, M. B. Dual-surface modified virus capsids for targeted delivery of photodynamic agents to cancer cells. *ACS Nano* **4**, 6014–6020 (2010).

47. Stewart, J. J., Lee, C. Y., Ibrahim, S., Watts, P., Shlomchik, M., Weigert, M. & Litwin, S. A Shannon entropy analysis of immunoglobulin and T cell receptor. *Mol. Immunol.* **34**, 1067–1082 (1997).

48. Starr, T. N. & Thornton, J. W. Epistasis in protein evolution. *Protein Sci.* **25**, 1204–1218 (2016).

49. Golmohammadi, R., Valegård, K., Fridborg, K. & Liljas, L. The Refined Structure of Bacteriophage MS2 at 2·8 Å Resolution. *J. Mol. Biol.* **234**, 620–639 (1993).

50. Dai, X., Li, Z., Lai, M., Shu, S., Du, Y., Zhou, Z. H. & Sun, R. *In situ* structures of the genome and genome-delivery apparatus in a single-stranded RNA virus. *Nature* **541**, 112–116 (2016).

51. Kannoly, S., Shao, Y. & Wang, I.-N. Rethinking the evolution of single-stranded RNA (ssRNA) bacteriophages based on genomic sequences and characterizations of two R-plasmid-dependent ssRNA phages, C-1 and Hgal1. *J. Bacteriol.* **194**, 5073–9 (2012).

52. Ashcroft, A. E., Lago, H., Macedo, J. M. B., Horn, W. T., Stonehouse, N. J. & Stockley, P. G. Engineering thermal stability in RNA phage capsids via disulphide bonds. *J. Nanosci. Nanotechnol.* **5**, 2034–41 (2005).

53. Hietpas, R. T., Jensen, J. D. & Bolon, D. N. A. Experimental illumination of a fitness landscape. *Proc. Natl. Acad. Sci. U. S. A.* **108**, 7896–7901 (2011).

54. Engler, C., Gruetzner, R., Kandzia, R. & Marillonnet, S. Golden gate shuffling: a one-pot DNA shuffling method based on type IIs restriction enzymes. *PLoS One* **4**, e5553 (2009).

55. Pinto, F., Thapper, A., Sontheim, W. & Lindblad, P. Analysis of current and alternative phenol based RNA extraction methodologies for cyanobacteria. *BMC Mol. Biol.* **10**, 79 (2009).

56. Wang, W. & Malcolm, B. A. Two-Stage PCR Protocol Allowing Introduction of Multiple Mutations, Deletions and Insertions Using QuikChangeTM Site-Directed Mutagenesis. *Biotechniques* **26**, 680–682 (1999).

57. Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: A flexible trimmer for Illumina sequence data. *Bioinformatics* **30**, 2114–2120 (2014).

58. Magoč, T. & Salzberg, S. L. FLASH: Fast length adjustment of short reads to improve genome assemblies. *Bioinformatics* **27**, 2957–2963 (2011).

59. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754–1760 (2009).

60. Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G. & Durbin, R. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).

61. Shin, H.-C., Roth, H. R., Gao, M., Lu, L., Xu, Z., Nogues, I., Yao, J., Mollura, D. & Summers, R. M. Deep Convolutional Neural Networks for Computer-Aided Detection: CNN Architectures, Dataset Characteristics and Transfer Learning. *IEEE Trans. Med. Imaging* **35**, 1285–1298 (2016).

62. Pontius, J., Richelle, J. & Wodak, S. J. Deviations from standard atomic volumes as a quality measure for protein crystal structures. *J. Mol. Biol.* **264**, 121–136 (1996).

63. Sandberg, M., Eriksson, L., Jonsson, J., Sjöström, M. & Wold, S. New chemical descriptors relevant for the design of biologically active peptides. A multivariate characterization of 87 amino acids. *J. Med. Chem.* **41**, 2481–2491 (1998).

64. Krigbaum, W. R. & Komoriya, A. Local interactions as a structure determinant for protein molecules II. *Biochim. Biophys. Acta* **576**, 204–228 (1979).

65. Fauchère, J.-L., Charton, M., Kier, L. B., Verloop, A. & Pliska, V. Amino acid side chain parameters for correlation studies in biology and pharmacology. *Int. J. Pept. Protein Res.* **32**, 269–278 (2009).

66. Fasman, G. D. *Practical handbook of biochemistry and molecular biology.* (1989).

67. Letunic, I. & Bork, P. Interactive tree of life (iTOL) v3: an online tool for the display and annotation of phylogenetic and other trees. *Nucleic Acids Res.* **44**, W242–W245 (2016).

68. Crooks, G. E., Hon, G., Chandonia, J. M. & Brenner, S. E. WebLogo: A sequence logo generator. *Genome Res*. **14**, 1188–1190 (2004).

Chapter 4: Systematic Engineering of a Protein Nanocage for High-yield, Site-specific Modification

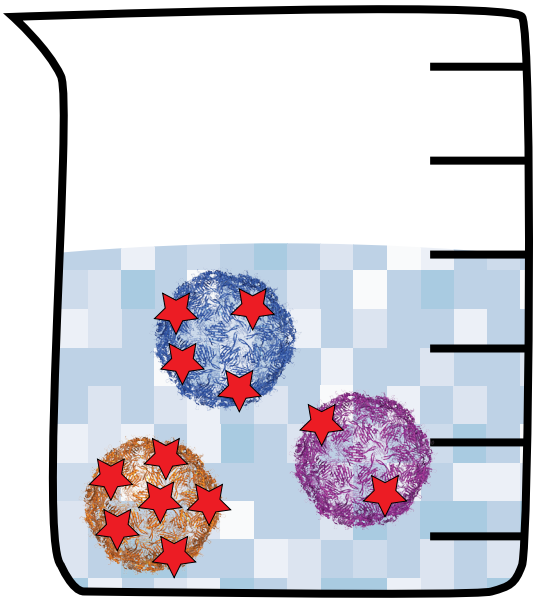
The following is adapted from Brauer, Hartman*, Bader, Merz, Tullman-Ercek, and Francis; J. Am. Chem. Soc., 2019 with permission*

Short summary

The MS2 N terminus is involved in secondary structure and is not chemically reactive. We generated all possible three amino acid extensions of the MS2 N terminus, then evaluated how those variants assembled and tolerated both thermal and chemical stress. We found a subset of extensions, called HiPerX variants, were both stable and chemically reactive, enabling a new method to modify MS2.

Abstract

Site-specific protein modification is a widely-used strategy to attach drugs, imaging agents, or other useful small molecules to protein carriers. N-terminal modification is particularly useful as a high-yielding, site-selective modification strategy that can be compatible with a wide array of proteins. However, this modification strategy is incompatible with proteins with buried or sterically-hindered N termini, such as virus-like particles composed of the well-studied MS2 bacteriophage coat protein. To assess VLPs with improved compatibility with these techniques, we generated a targeted library based on the MS2-derived protein cage with N-terminal proline residues followed by three variable positions. We subjected the library to assembly, heat, and chemical selections, and we identified variants that were modified in high yield with no reduction in thermostability. Positive charge adjacent to the native N terminus is surprisingly beneficial for successful extension, and over 50% of the highest performing variants contained positive charge at this position. Taken together, these studies described nonintuitive design rules governing N-terminal extensions and identified successful extensions with high modification potential.



4.A. Introduction

Site-specific bioconjugation techniques are widely used to produce useful conjugate biomaterials. Many recently-developed N-terminal modification strategies are of particular interest, as these reactions are high-yielding, can proceed under mild reaction conditions, and have the capacity to be site-selective[1–9]. Because nearly all proteins contain a single instance of an N terminus, these reactions are useful in a wide variety of contexts[10], including the loading of cargo onto protein carriers[11] or the development of new biomaterials[12,13]. However, such reactions require free N-terminal residues that are uninvolved in secondary structure, limiting their usefulness on proteins with sterically-hindered N termini. One such case is the MS2 bacteriophage, a well-studied protein nanocage that is being actively explored for applications in drug delivery[14–16], disease imaging[17], vaccines[18,19], and biomaterials[20–22]. Limited genetic manipulations can be made to the MS2 coat protein (CP) without disrupting the assembly state[23], and many inter- and intra-subunit contacts make mutability challenging to predict[24]. Additionally, the native N terminus is sterically hindered, and efforts to extend the N terminus have had limited success[25]. As such, currently developed N-terminal modification strategies are not compatible with the MS2 CP. Instead, the attachment of targeting groups to the exterior of the MS2 CP either relies on nonspecific chemistry, such as lysine modification, or requires the incorporation of nonstandard amino acids, lowering expression yields and complicating protocols[26,27]. The usefulness of the MS2 scaffold would be expanded substantially by enabling N-terminal modification of the CP in a manner that yields stable, easy-to-produce, and modifiable virus-like particles (VLPs).

Here, we combine a systematically generated library with direct functional selections to identify N-terminally extended variants of the MS2 CP that are well-assembled, thermostable, and amenable to chemical modification (**Figure 4.1**). In addition to identifying highly useful extensions that can be modified to >99% by oxidative couplings between the

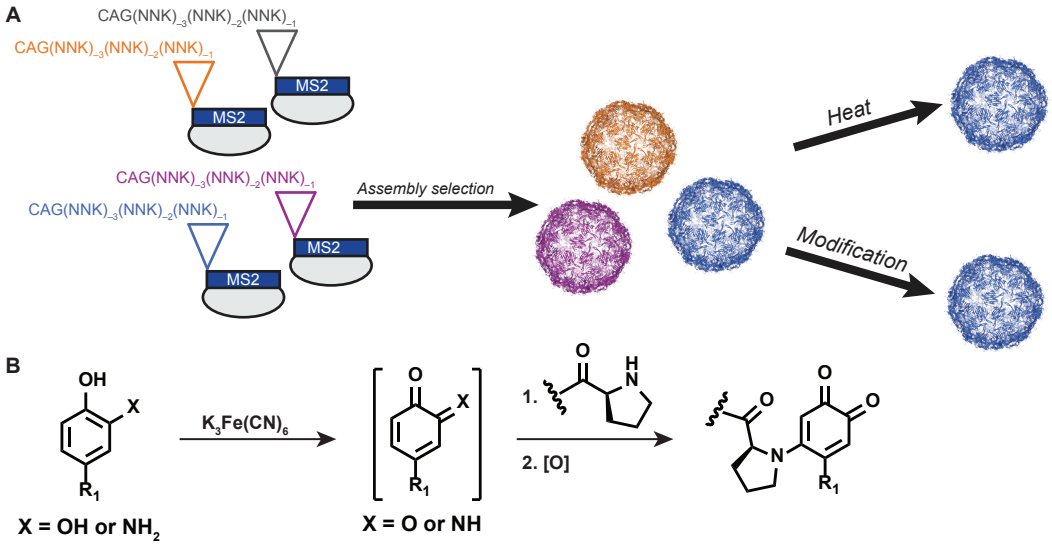


Figure 4.1. Scheme to isolate N-terminally extended VLPs with desired properties. A) Three-codon NNK extensions at the N terminus generated a library of 8,000 variants. Assembly, thermostability, and chemical modification challenges were used to identify HiPerX variants, or high-performing N-terminal extensions with desirable properties, indicated in blue. B) Oxidative coupling reactions can be used to modify N-terminal proline residues.

N terminus and oxidized catechols, we also uncovered surprising design rules governing which extensions are compatible with particle assembly. Of 8,000 possible combinations of N-terminal extensions, merely 3% of the library remained assembled through stringent chemical and thermal selections. In addition to identifying useful VLP variants for biomedical applications, this study represents the first time that chemical modification conditions have been used as a selection for protein fitness. This approach could be adapted to study the modification efficiency for other reactions or protein substrates, and could provide rich information about the effects of amino acid sequence on reactivity.

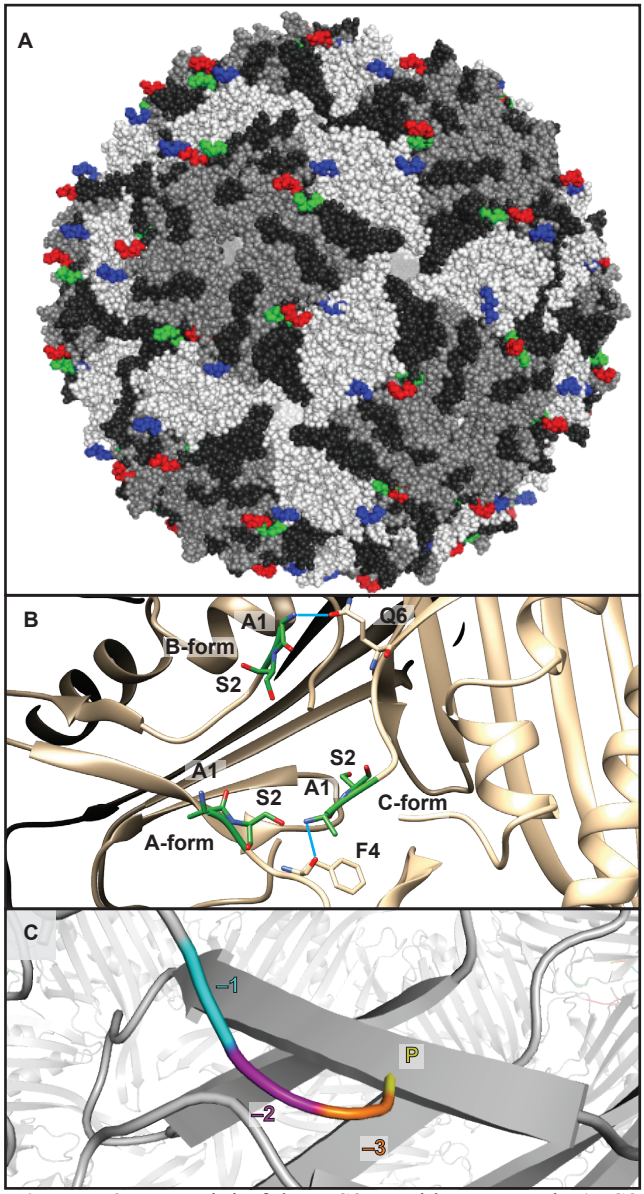


Figure 4.2. N termini of the MS2 capsid coat protein (MS2 CP) monomers. A) Each quasi-equivalent form and N terminus is indicated with a shade of gray or color, respectively. The N terminus of the A form (dark gray) is shown in red; the N terminus of the B form (white) is shown in blue; and the N terminus of the C form (gray) is shown in green. B) Hydrogen bonding interactions are shown for the native N terminus. Hydrogen bonds are shown in blue. C) The –1 (cyan), –2 (purple), –3 (orange), and –4 (proline, yellow) positions are indicated in relation to the native N terminus (alanine) of the MS2 CP.

4.B. Results and Discussion

4.B.i. Characterization of a comprehensive N-terminally extended MS2 bacteriophage library

The MS2 VLP is a 27 nm icosahedral particle that is composed of 180 copies of a protein monomer. Three N termini of these quasi-equivalent proteins are clustered together, forming a triangle with lengths of 11.7 Å, 12.8 Å, and 7.9 Å (**Figure 4.2A**)[28]. This sterically confined local environment suggests that few N-terminal extensions would be compatible with particle assembly. As such, we sought to use Systematic Mutagenesis and Assembled Particle Selection (SyMAPS), a technique developed previously in our labs[23], to evaluate all possible proline-terminated extensions of the MS2 CP with the pattern P-X-X-X-MS2, where X represents all amino acids. When expressed in *E. coli*, assembly-competent variants of the MS2 CP encapsulate available negative charge, including mRNA. SyMAPS capitalizes on this property, using the encapsulated nucleic acid as a convenient genotype-to-phenotype link. Well-assembled VLPs copurify with a snapshot of cellular nucleic

acid, including variant mRNA, while mRNA from poorly-assembled VLPs is lost.

As shown in **Figure 4.1A**, an NNK-based strategy was used to encode all variants while minimizing biases due to genetic code redundancies[23]. Following expression, the N-terminal methionine of wild-type MS2 CP is cleaved, yielding an alanine in position one. In the library, extensions were appended directly before alanine 1, starting with a –1 position. With this numbering, the N-terminal proline is located at the –4 position (**Figure 4.2B**). Proline also is compatible with efficient methionine cleavage, leading to a library with four total extended residues[29]. The invariant N-terminal proline was chosen because these residues were shown to modify to high conversion via an oxidative coupling bioconjugation reaction (**Figure 4.1B**)[1,30]. While this modification strategy is mild, fast, and efficient, the wild-type MS2 CP was observed to modify in less than 5% yield.

Using SyMAPS, we characterized the assembly competency of each variant in the P-X-X-X-MS2 library, generating an Apparent Fitness Landscape (AFL). We generated a quantitative assembly score for every mutant in the targeted library by comparing the relative log% abundance of each variant before and after an assembly selection with size exclusion chromatography (SEC), identifying the subset of P-X-X-X-MS2 extensions that were competent for VLP assembly (**Figure 4.3**, **Supplementary Fig. 4.1**). In addition, we generated a non-proline terminated library, X-X-X-MS2, to distinguish which assembly trends were general and which were specific to a proline at position –4 (**Supplementary Fig. 4.2**).

Of the 8,000 variants, around 92% were observed in the starting plasmid library, consistent with coverage of previous SyMAPS libraries[23,24]. Of these, 48% were absent in the VLP library after the assembly selection, indicating that these extensions likely did not permit assembly. These low-scoring variants could be a result of mutations that are assembly-incompetent, poorly expressed, or unstable to protein expression[23]. Around 24% of the variants scored Apparent Fitness Score (AFS) values greater than 0.2, indicating that assembly occurred readily[23]. Variants with a nonsense mutation had an average AFS value of –3.0 with a standard deviation of 1.5, indicating that these sequences were depleted from the population of selected VLPs by 1000-fold.

We observed striking trends in the AFL when the data were grouped by the identity of the –1 position (or the position nearest to the native N-terminus) (**Figure 4.3**). We evaluated the number of variants with P-X-X-Z-MS2 that were compatible with assembly (**Figure 4.4A**), in which Z is the amino acid at the –1 position. Positive charge was particularly well-tolerated at this position and enabled a wide variety of extensions with the pattern P-X-X-[R/K]-MS2 (**Figure 4.3**, **Figure 4.4A**). This was surprising given the sterically hindered environment of the N terminus in the MS2 CP. Nearly 80% of extensions with the pattern P-X-X-R-MS2 assembled (AFS value > 0.2), and over 60% of extensions with P-X-X-K-MS2 assembled, compared to merely 12% of P-X-X-D-MS2 and 8% of P-X-X-E-MS2. These results suggest that the beneficial effect is due specifically to positive charges rather than any charge at all.

Glycine and alanine, both common choices for rational N-terminal extensions, performed worse than expected compared to other amino acids, with 23% and 18% extensions permitted, respectively (**Figure 4.4A**). More intuitively, bulky residues such as tryptophan, phenylalanine, or leucine were poorly tolerated at the –1 position. In contrast,

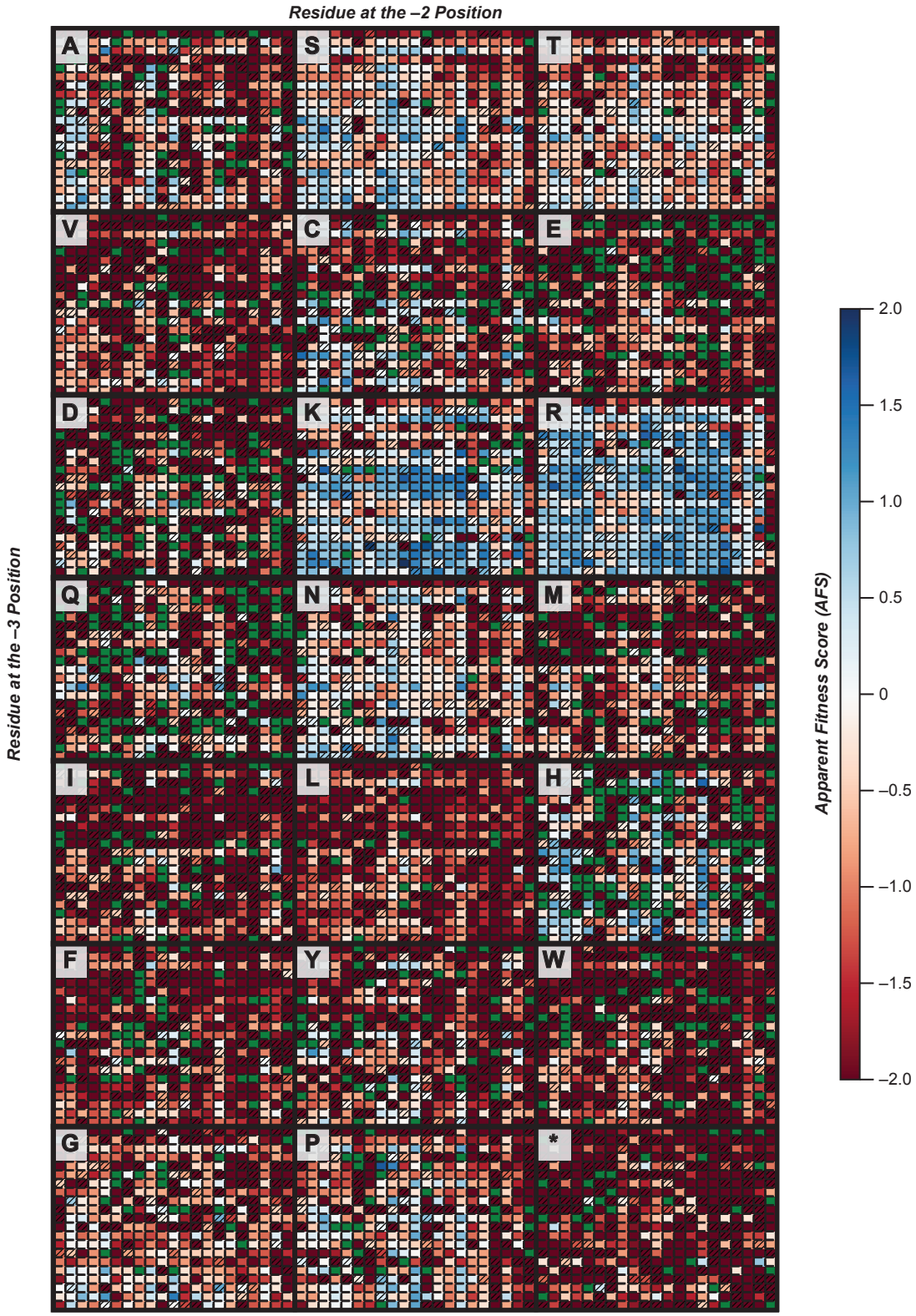


Figure 4.3. Apparent Fitness Landscape of P-X-X-X-MS2 N-terminal extensions. Extensions are labeled as the distance from the native N terminus (alanine), and the –1 position is indicated in the upper left corner of each quadrant. Color scheme is as follows: blue (enriched variants), red (unenriched variants), dark red (variants present in the plasmid library but absent in the VLP library), green (missing values). The unaveraged AFS is reported for variants with a missing value in a single replicate and is indicated by hatching. Nonsense mutations are marked with asterisks.

polar residues that can act as hydrogen donor and acceptors, such as serine, threonine, or asparagine, performed well. Asparagine was better tolerated than glutamine, indicating that side chain length may contribute to mutability of this position. Interestingly, histidine was also relatively well-tolerated and was the fifth most permitted amino acid at this position; however, only 40% of extensions with the pattern P-X-X-H-MS2 assembled, which is far lower than either arginine or lysine.

To visualize this effect, we plotted a histogram of Apparent Fitness Score (AFS) values with arginine or lysine at the –1 position compared to all other AFS values (**Figure 4.4B**). These residues in this position shift the average AFS values to be more positive, indicating that a higher percent of variants was compatible with self-assembly. Additionally, a histogram of arginine or lysine in the –1 position was compared with arginine or lysine at the –2 or –3 position to evaluate whether this effect was location specific. In this case, a notable shift to more positive AFS values was found with arginine or lysine only at the –1 position, suggesting the charge effect is indeed specific to this location (**Figure 4.4C**). A larger version of the data for arginine in position –1 appears in **Figure 4.4D**.

Finally, we confirmed that these trends were similar to N-terminal extensions in the absence of proline in the –4 position (**Supplementary Fig. 4.2**). In this library, X-X-[K/R]-MS2 also resulted in a disproportionally high number of assembled particles compared to other amino acids at the –1 position, indicating that this trend is likely general for N-terminal extensions of the MS2 CP rather than specific to those with starting with proline.

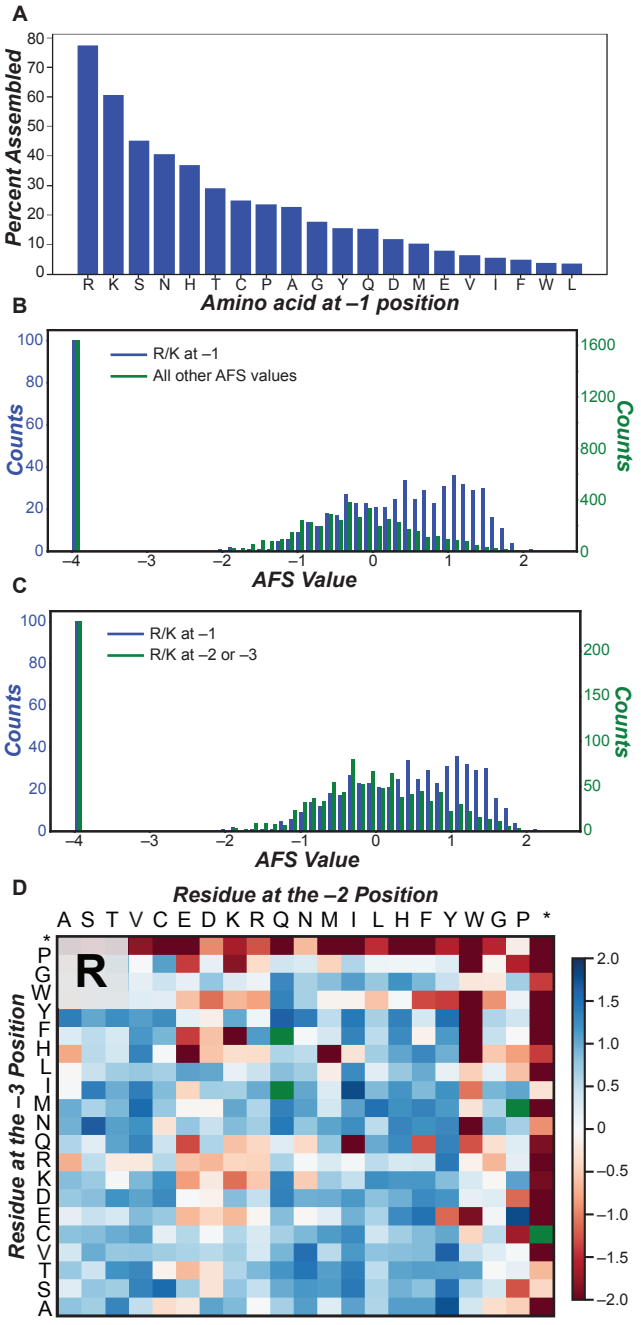


Figure 4.4. Effect of positive charge at the –1 position in N-terminal extensions. A) The amino acid identity at the –1 position alters how many extensions are assembly-competent. Arginine and lysine permit 77% and 61% of possible extensions with the pattern P-X-X-[R/K]-MS2. Arginine and lysine at position –1 result in a higher percent of positive AFS values (B) compared with all other AFS values or (C) compared to positive charge at –2 or –3. D) The assembly scores of all P-X-X-R-MS2 extensions are shown as an example, in which arginine is at the –1 position.

In order to evaluate the potential interactions responsible for this favorable effect, we performed a conformational search of a hexameric unit of P-A-A-R-MS2 (**Supplementary Fig. 4.3A**). Most notably, a new salt bridge is formed in the *in silico* study between the N-terminal proline of B chain monomer and Asp17 of the C chain (**Supplementary Fig. 4.3B**). In addition, hydrogen bonding is observed between arginine at the –1 position and Gln6 of the C chain in the minimized structure. We hypothesize that these hydrogen bonds are beneficial for assembly, as many extension combinations with a hydrogen bond donor residue at the –1 position are permitted.

In a conformational search of P-A-A-A-MS2 and P-A-R-A-MS2, neither the hydrogen bond nor the salt bridge were observed in either extension. These variants have lower AFL scores and lack a hydrogen bond donor side chain at the –1 position.

Finally, we analyzed P-A-R-R-MS2, a relatively poorly performing variant, via structure minimization. We found that while the –1 arginine formed the presumably beneficial salt bridge, multiple van der Waals clashing interactions were also found between arginine residues of the A and C chains (**Supplementary Fig. 4.3C**). The stringent positional specificity of these interactions highlights the remarkable level of detail offered by a comprehensive mutational strategy such as SyMAPS.

4.B.ii. Interpreting the Apparent Fitness Landscape

We evaluated the consistency of the data by plotting the two biological replicates of the P-X-X-X-MS2 dataset as a scatterplot (**Supplementary Fig. 4.4A, Supplementary Data 4.1**). In addition, we plotted the three biological replicates of the X-X-X-MS2 dataset (**Supplementary Fig 4.4B-D**). In general, we find that the datasets do correlate, though the R^2 values are relatively low (0.42–0.59). This variability may arise from a number of sources, including technical differences between assembly selections: for example, bacterial growth rates or expression levels are both variables that are not controlled that may affect the selections beyond assembly competency. Correlations between the two chemical modification selections are even lower (0.37, **Supplementary Fig 4.4E**), indicating that significant variability may exist between replicates of the same challenge.

Interestingly, correlations within a replicate are somewhat higher, even when comparing chemical modification and assembly selections. While several extensions are positive in the assembly selection and negative in the chemical modification selection (as is to be expected for additional selective pressure), very few of the opposite are seen. Correlations for these are 0.52 and 0.67 for replicates 1 and 2, respectively (**Supplementary Fig. 4.4F,G**). We find that the heat selection correlates well with the chemical selection for replicate one, yielding an R^2 of 0.75 and few off-axis datapoints (**Supplementary Fig. 4.4H**). From these analyses, we see that replicate variability likely arises from growth or expression rather than the selections themselves.

Finally, we evaluated whether low abundances in the plasmid library contributed to the low correlation. Requiring at least two reads in both replicates for the P-X-X-X-MS2 dataset did increase the correlation of the replicates to 0.52 from 0.42 (**Supplementary Fig. 4.4I**), and further requiring at least 10 reads in both replicates increased the R^2 to 0.67. While increased stringency does improve correlation, it appears that factors beyond read abundances, such as biological noise, contribute to differences between the replicates.

To reduce the impact of stochastic variation on our dataset and highlight variants with higher certainty in their determined fitness scores, we developed two additional methods for data processing. The first method filters out variants with low plasmid read counts (<4), as these are more prone to error. Though this reduces the coverage of our P-X-X-X library, we were pleased to find that the high-level trends remain apparent in the filtered heatmaps (**Supplementary Fig. 4.5-4.7**). For example, the stark favorable effect of positive charge at the –1 position on assembly competency was retained, as over half (54%) of all assembling variants bear a lysine or arginine at this position.

The second method of data processing aimed to remove all variants with ambiguous assembly competency and simplify the output to a binary 'assembling' and 'non-assembling' value. All variants with an AFS near 0 or an AFS that changed sign between replicants were removed from analysis. Variants with consistent high or low fitness scores were marked as 'assembling' or 'non-assembling' mutants, respectively (**Supplementary Fig. 4.8-4.10**). This method of processing precludes detailed comparison of variant scores but allows for rapid selection of N-terminally extended MS2 variants with a clear assembly phenotype. While many of the trends discussed above are replicated in all methods of data analysis, we recommend using either of the high stringency methods to select individual extensions for further experiment. These supplementary processing methods serve as complementary approaches to interpreting SyMAPS datasets.

4.B.iii. Direct functional selections for HiPerX variants

The chemical modification of VLPs imposes a number of challenges to self-assembly, and any useful variant must tolerate reaction conditions as well as strain introduced by the covalent attachment of new functionality. As such, we designed a selection for tolerance to chemical modification conditions to identify variants that are well-suited for use as protein scaffolds. We used an N-terminal oxidative coupling reaction for this challenge[1]. The oxidative coupling uses a mild metal oxidant to convert methoxyphenols[31], aminophenols[1], and catechols[32] to *ortho*-quinone and *ortho*-iminoquinone intermediates that react selectively with anilines[27], reduced cysteines[33], and N-terminal amines of proteins or peptides[1]. In this study, we used aminophenols and catechols as *ortho*-quinone precursors, as both can be rapidly oxidized via $K_3Fe(CN)_6$ (**Figure 4.1B**).

The library was chemically coupled to DNA oligomers bearing *o*-aminophenol handles, simultaneously exposing the library to chemical modification conditions and to the strain of coupling large biomolecules to the VLP surfaces (**Figure 4.5A**). Variants that remained assembled under these conditions were enriched through HPLC SEC and sequenced. As with the assembly-selected AFL, we compared percent abundance of the library after the selection to the plasmid library to generate a quantitative score of chemical modification compatibility. As a complement, we also evaluated the thermostability of all variants, subjecting the library to 50 ºC for 10 min to differentiate between wild-type-like variants and those with reduced thermostability. As a comparison, the wild-type VLPs are stable up to 65 ºC. Variants that remained assembled after this challenge were also purified by HPLC SEC, sequenced, and processed to generate a heat-selected AFL.

Surprisingly, the chemical modification selection was more stringent than the thermal selection: only 16% of the mutants were assembled following exposure to chemical

modification conditions, while 22% of the mutants tolerated 50 ºC for 10 min. In addition, chemical modification and thermostability scores showed stark differences in trends when compared to assembly-selected AFS values. While variants with multiple positive charges expressed and assembled far better than the library average (61% compared to 24%), these VLPs were almost universally sensitive to chemical and thermal challenges, suggesting that these types of extensions are unstable and therefore undesirable. Histidine behaved similarly in these challenges, and histidine at the –1 position when combined with positive charge at the –2 or –3 position was sensitive to thermal or chemical challenges. AFLs following thermal (**Supplementary Fig. 4.11**) or chemical modification (**Supplementary Fig. 4.12**) challenges present this phenomenon as distinct red bands within plots in which lysine and arginine are grouped by the –1 position. These data exhibit why functional challenges to variant libraries are crucial to disentangle subtle changes to VLP properties.

We next sought to generate insight into the variants performing well across all selections, which included assembly, thermal stability, and oxidative coupling selections. We generated an aggregate AFL that incorporated the results of each enrichment, in which

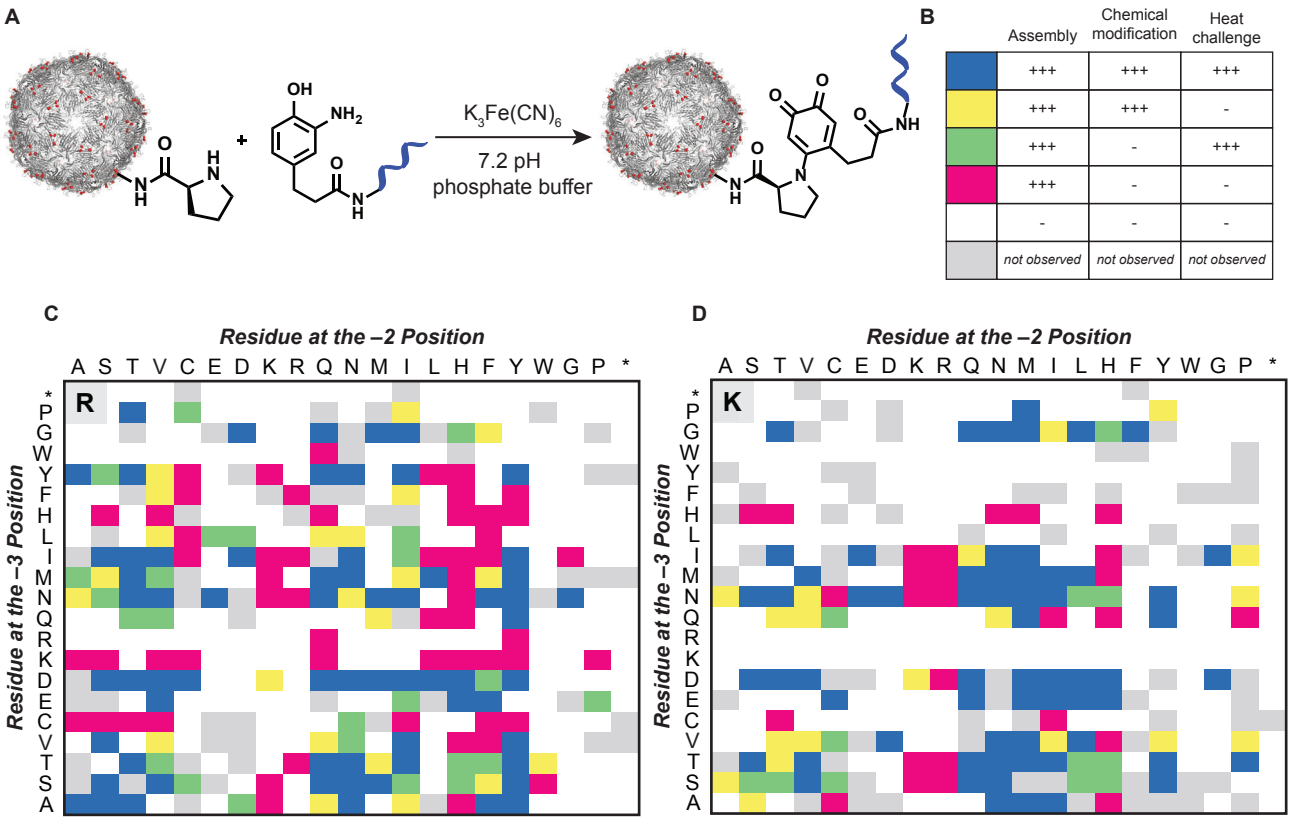


Figure 4.5. Combined Fitness Landscape of the P-X-X-X-MS2 N-terminal extensions. A) The chemical modification-based selection of the variant library employs bioconjugation to a 25 bp DNA strand. B) A color key is provided for the combined AFL data. (+++) indicates a score greater than 1.0 in the selection, and HiPerX, or high performing extensions, are indicated in blue. (–) indicates a score less than 1.0 in the selection. Combined AFLs are displayed for extensions (C) P-X-X-R-MS2 and (D) P-X-X-K-MS2. The full combined fitness landscape can be found in **Supplemental Fig. 4.13**.

a stringent threshold score for each parameter was used to isolate the most promising and useful variants. This aggregate AFL identified 238 thermally-stable, chemically-modifiable N-terminal extensions of the MS2 CP, indicated in blue (**Figure 4.5B**, **Supplementary Fig. 4.13**), and termed High Performing eXtended (HiPerX) variants. Consistent with the findings above, 129 of these 238 variants possessed lysine or arginine at the –1 position, accounting for 54% of the HiPerX variants (**Figure 4.5C,D**). With a stringent score of 10-fold enrichment in all selections, most amino acids at the –1 position resulted in few or no HiPerX variants. Interestingly, unsuccessful sequences included glycine, which is commonly used in rational design to engineer extensions or linkers between protein domains[34,35]. Branched amino acids were also poorly tolerated at the –1 position: a comparison between serine and threonine at the –1 position revealed that threonine performed far worse than serine. Proline was better tolerated and outperformed glycine, even though these extensions have at least two proline residues in the first four amino acids.

We also found many nonintuitive results that diverged from common protein engineering assumptions. For example, tyrosine at the –2 or –3 position, when combined with arginine at the –1 position, was observed in many HiPerX variants. Combinations with multiple charges (P-D-H-R-MS2) or multiple large amino acids (P-S-Y-R-MS2) are also assembly-competent, thermostable, and highly modifiable extensions. In particular, P-D-X-R-MS2 folded well across a broad range of X identities, such as when X was a small residue like serine, a hydrophobic residue such as isoleucine, or a polar, bulky residue like tyrosine. These results underscore the importance of experimental efforts to describe the mutability of large protein assemblies.

Bulky residues were tolerated at the –2 and –3 position in combination with arginine at the –1 position; however, by this metric, multiple positive charges were still detrimental to VLP stability. Even negative charge could not rescue stability in nearly all of these cases. The only extensions with multiple positive charges with any increased stability are P-D-K-[K/R]-MS2, which are thermally stable but do not tolerate chemical modification. Additionally, glycine was only tolerated at the –3 position, and even then, only when there is a positively-charged residue at the –1 position.

These trends, where multiply-charged or bulky combinations of residues are permitted, are difficult to reconcile with the structure of the N terminus of the MS2 CP. For example, the close proximity of the monomer N termini means that an extension like P-S-Y-R-MS2 positions multiple large and/or charged residues within 9 to 12 Å. These results also contrast with most rational N-terminal extensions, which rely on small residues such as serine or glycine to disrupt the local protein folding environment minimally[34,35].

We hypothesize that many of these mutations may enhance the critical charge interactions that make lysine and arginine desirable variants. For example, hydrophobic residues at the –2 position could create a more hydrophobic environment, reducing the local dielectric constant[36,37]. This in turn could strengthen the interactions involved in the proposed salt bridges. Alternatively, nonpolar residues in the –2 or –3 position could interact through hydrophobic effects. Regardless of the cause, in the absence of a systematic library approach and direct functional selections, these many nonintuitive yet critical findings would almost certainly have been missed. Ultimately, only 3% of the 8000

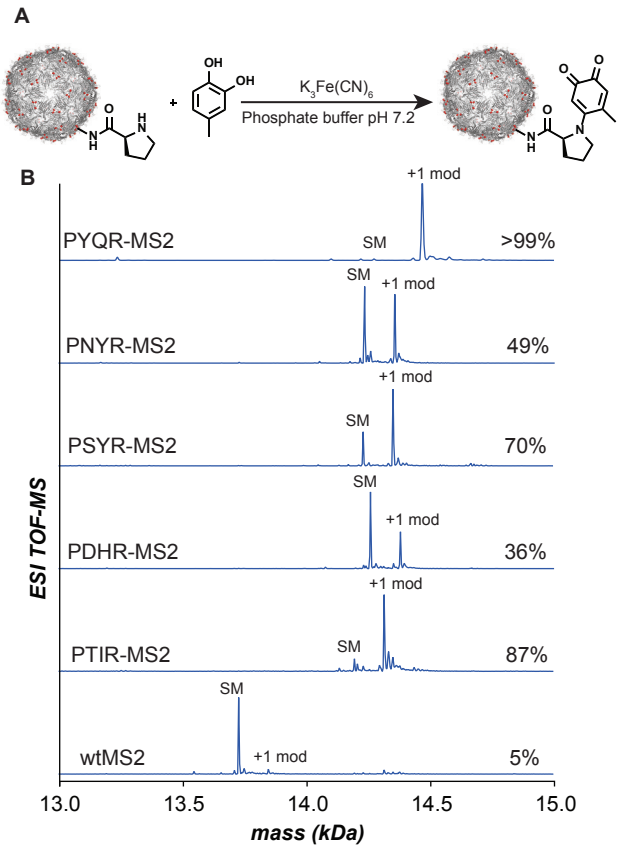


Figure 4.6. Chemical modification of HiPerX MS2 variants. A) An oxidative coupling reaction was evaluated for proline-terminated MS2 variants. B) Mass spectra of chemically-modified HiPerX variants of the MS2 CP are shown. Percent modification is determined by integration of the unmodified (SM) vs modified (+1 mod) peaks.

possible P-X-X-X-MS2 extensions were identified as HiPerX variants, enriched in assembly, thermal stability, and chemical modification.

4.B.iv. Characterization and modification of HiPerX variants

Based on the stringent selection conditions, HiPerX variants were expected to have increased tolerance to chemical reaction conditions; however, it was not known whether the N termini of these variants would be modified at higher rates than CP[WT]. As such, we sought to validate trends identified in high-throughput sequencing and to characterize the usefulness of HiPerX variants as protein scaffolds. To do so, five randomly-selected HiPerX variants with P-X-X-R-MS2 extensions were cloned and evaluated individually. These variants were selected because this population showed the largest enrichment of across all challenges. All five variants expressed in high yield, formed assembled VLPs, and tolerated the thermal challenge of 50 ºC for 10 min, supporting the quality of the AFLs (**Supplementary Fig. 4.14**).

We next evaluated whether the engineered extensions enhanced reactivity to the N-terminal oxidative coupling reaction. We performed a reactivity test to modify the VLP N termini with a small molecule catechol derivative (**Figure 4.6A**). Gratifyingly, all tested variants showed a significant enhancement in reaction conversion compared to the wild-type MS2 CP. HiPerX variants showed 36-87% modification, compared to <5% modification in wild-type VLPs (**Figure 4.6B**). The dramatic increase in conversion under these conditions is notable in and of itself; additionally, as there are 180 MS2 CPs per VLP, these N-terminally extended variants are capable of displaying up to 65-160 copies of the new functionality per VLP, representing a substantial increase in targeting or drug carrying capabilities. HPLC SEC of modified samples confirmed that all of these VLPs remained assembled after modification (**Supplementary Fig. 4.15**). This result shows, that for the first time, SyMAPS can be combined with a chemical modification enrichment to identify highly modifiable variants. In addition, given that all five randomly-selected variants modified at higher rates than CP[WT]—and because we expect the N-terminal prolines to be solvent accessible—we anticipate that many other HiPerX variants will also be amenable to modification.

We next sought to evaluate whether these extensions were compatible with other bioconjugation strategies (**Figure 4.7A-C**). One such N-terminal modification strategy using 2-pyridinecarboxaldhyde (2PCA) modifies most N terminal residues to high-yield in a single step through a mechanism that is distinct from oxidative coupling reactions (**Figure 4.7B**)[2]. These two chemistries do not share common intermediates and proceed under different reaction conditions. To evaluate HiPerX variant performance with 2PCA, extended variants and CP[WT] were incubated with excess reagent overnight at room temperature, according to the published protocol[2]. We observed that HiPerX variants resulted in around 80% modification with 2PCA, compared to around 30% modification with CP[WT], even though these extensions were optimized for the $K_3Fe(CN)_6$ oxidative coupling reaction (**Figure 4.7D**). While the fold improvement was lower for this reaction, an increase from 30% modification (CP[WT]) to 80% (CP[HiPerX]) modification represents a useful increase in the number of functional groups installed on the exterior, from 50 modifications to 140 modifications (**Figure 4.7E**).

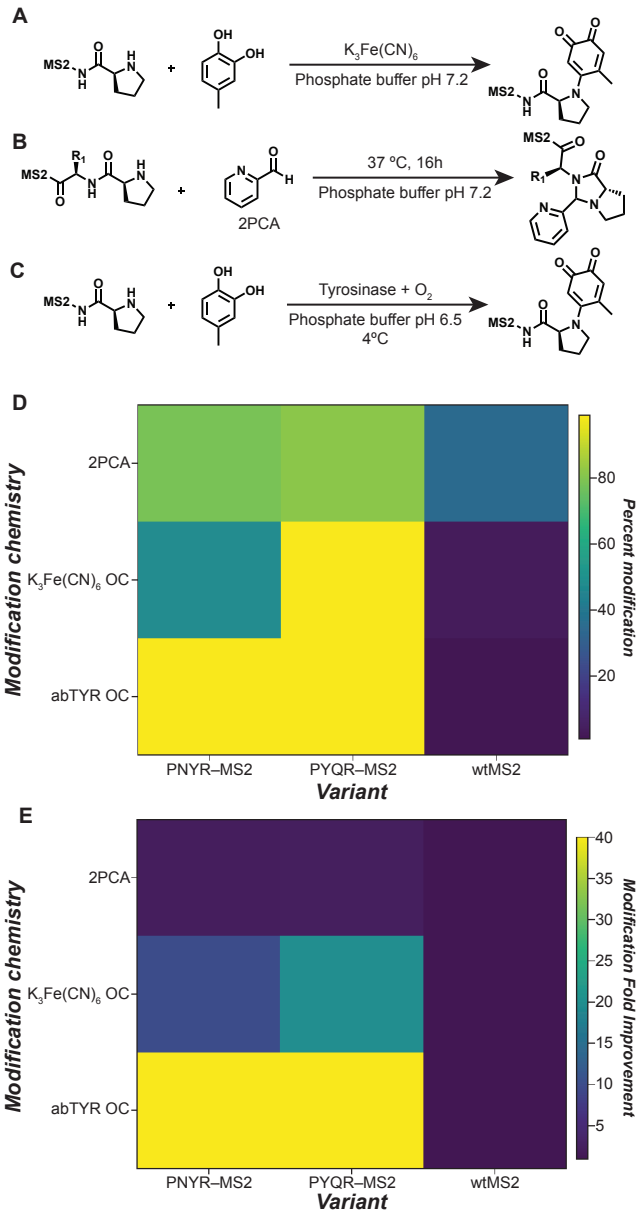


Figure 4.7. Conversion and fold improvement of N-terminal modification strategies of HiPerX MS2 variants. Reaction schemes are shown for (A) potassium ferricyanide-mediated oxidative coupling, (B) 2-pyridinecarboxaldehyde (2PCA) modification, and (C) tyrosinase-mediated (abTYR) oxidative coupling reactions. D) Modification of two HiPerX MS2 variants is shown in contrast to wild-type MS2 across these modification strategies. E) Fold improvement compared to wild-type MS2 is shown.

We also investigated a new tyrosinase-mediated variant of the oxidative coupling reaction (abTYR) that proceeds through a similar mechanism to $K_3Fe(CN)_6$ oxidative coupling after the ortho-quinone intermediate is produced (**Figure 4.7C**)[30]. The enzymatic oxidation is compatible with phenols as well as catechols; thus, compatibility with abTYR would widen the scope of potential small molecule partners to include many shelf-stable phenols. We found that modification yields with catechols increased from good (36-87%) to near-quantitative (>99%) in all cases (**Figure 4.7E**). In addition, CP[PQYR] was found to be compatible with installation

and modification of a reactive cysteine in the interior cavity[38]. Interior labeling was performed with an AlexaFluor-488 maleimide dye, and modification efficiency with this strategy was high (>99%), as previously reported[16,38–40]. More importantly, subsequent exterior modification via abTYR-mediated oxidative coupling also proceeded to over 99% conversion, resulting in doubly modified VLPs with 180 copies of both functionalities (**Supplementary Fig. 4.16**). All together, these extensions are thermally stable, highly modifiable, and can carry cargo, making them promising carriers with highly desirable properties for a number of biomedical applications.

Finally, we determined whether the N terminus is compatible with enzymatic glycosylation. Glycosyltransferases can modify proteins at short peptide sequences, and fitness landscape studies have mapped how sequence identity alters glycosylation efficiency for several glycosyltransferases[41]. We compared the Apparent Fitness Landscape for the MS2 N-terminal extensions to a rigorous analysis of permitted glycosylation sequence for ApNGT, a promiscuous glycosyltransferase[41]. Eight glycosylation sequences were designed to be compatible with VLP assembly and enzymatic glycosylation. Seven of these eight extensions formed well-assembled particles, though longer extensions had somewhat reduced yield (**Supplementary Fig. 4.17A**).

Glycosylation was evaluated in collaboration with the Jewett lab at Northwestern University. A highly active enzyme, ApNGT Q469A, was incubated with VLPs bearing all seven sequences (**Supplementary Fig. 4.17B**)[41]. These variants were then evaluated for glucose modification by mass spectrometry. We found that glycosylation was not permitted with any of the extensions, despite sequence compatibility, perhaps indicating that steric hinderance in this region prevents enzyme access to the glycosylation sequence (**Supplementary Fig. 4.17C**). Future libraries and functional selections can expand on these efforts to yield variants that are more easily accessible for enzymatic modification.

4.B.v. Extensions are well-assembled and modified in combination with CP[S37P]

HiPerX variants were evaluated for compatibility with other VLP structures. Qβ is a well-studied bacteriophage with promise as a drug delivery vehicle and vaccine candidate[42,43]. The VLP is structurally related to the MS2 CP; however, its coat protein is only 20% identical by sequence to the CP of MS2. We appended three HiPerX variants to the N terminus of Qβ. An assembly assay showed that the extensions were not compatible with VLP assembly, indicating that the extensions may be specific to the MS2 CP (**Supplementary Fig. 4.18**).

We next evaluated whether HiPerX variants were compatible with an alternative structure of the MS2 CP. Previous work in our lab identified a variant of the MS2 VLP with altered quaternary geometry[44]. This CP[S37P] mutation alters the global structure from a 27 nm wild-type-sized VLP to a smaller, 17 nm VLP (**Figure 4.8A**). This smaller-sized variant retains similar thermostability and is a useful tool to probe the effect of carrier size directly in applications such as drug delivery or imaging[45]. However, the N terminus of the CP[S37P] is distinct from CP[WT], both in minor structural differences and spatial positioning. To date, the exterior of CP[S37P] has not been modified, and its N terminus is sterically unavailable, similar to the parent CP[WT].

We sought to determine whether HiPerX sequences could be appended to the

CP[S37P] structure, enabling facile modification without repeating the library generation and functional selections. Despite the differences in geometry and secondary structure, all three N-terminally extended CP[HiPerX–S37P] variants assembled into well-formed VLPs. Each variant retained the T=1 geometry and smaller size, as confirmed by dynamic light scattering (**Supplementary Fig. 4.19A**). Additionally, variants tolerated 50 ºC for 10 min, indicating that thermostability was preserved in the new genetic background (**Supplementary Fig. 4.19B**).

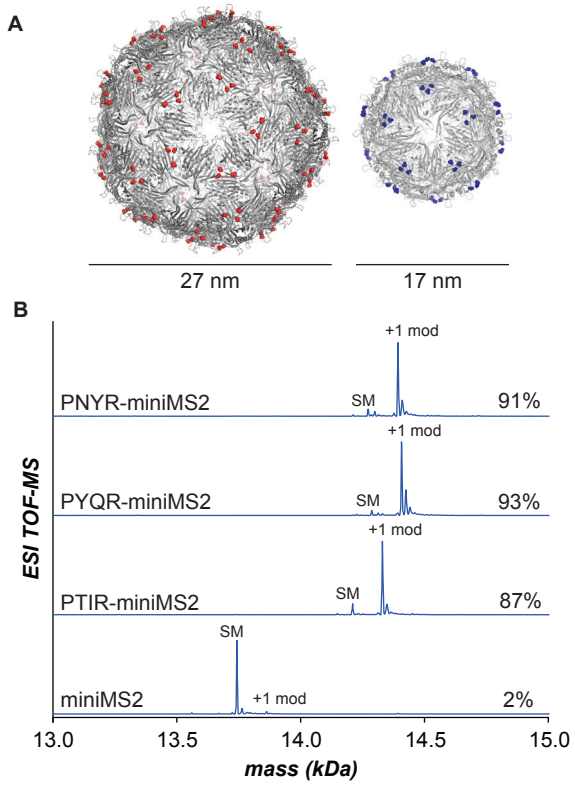


Figure 4.8. Chemical modification of HiPerX miniMS2 variants (CP[HiPerX–S37P]. A) Crystal structures of CP[WT] and CP[S37P] are shown with N termini highlighted in red and blue, respectively. B) Mass spectra of chemically-modified HiPerX variants of the MS2 CP are shown.

We next modified the exterior of the N-terminally extended CP[S37P] variants with the $K_3Fe(CN)_6$ oxidative coupling reaction, appending a catechol small molecule to the N-terminus. We found that CP[HiPerX–S37P] variants modified equally as well as the parent HiPerX variants, achieving >85% modification in all cases (**Figure 4.8B**). As a comparison, CP[S37P] modified <5%, indicating that the extensions are critical to achieve high modification rates. Despite changes to surface curvature and quaternary structure geometry, the selected HiPerX variants performed remarkably well as useful N-terminal extensions with CP[S37P]. Furthermore, this presents the first successful exterior modification of MS2 CP[S37P], enabling future study of 17 nm VLP variant as a targeted protein scaffold.

4.C. Conclusion

The site-specific modification of proteins is of fundamental importance for many applications, including drug delivery, vaccines, and protein biomaterials. Here, we combined a systematically-generated library with a functional selection under chemical modification conditions to identify variants of the MS2 CP that are highly compatible with N-terminal modification. The fact that only 3% of the library were enriched after the full set of challenges underscores the fact that the introduction of non-native amino acids into proteins remains a nonintuitive process *a priori*. This is particularly true in the case of self-assembling proteins, as single point mutations lead to amplified effects when propagated throughout the quaternary structure. In this study, an unexpected charge interaction was uncovered that counters these effects, and in some cases, was bolstered by additional hydrophobic interactions. The selection procedure for bioconjugation conditions could be

used with many future libraries to identify new reactive sequences. Finally, the MS2 CP variants identified in this study can be doubly modified to >99% yield on both the interior and exterior surfaces, providing homogeneous carrier materials in two different sizes for a variety of drug delivery applications.

4.D. Methods

4.D.i. Library generation

To generate libraries of N-terminal extensions, we modified a library generation and selection strategy used previously in our lab, called SyMAPS[23]. SyMAPS uses the EMPIRIC cloning developed in the Bolon lab[46]. In EMPIRIC cloning, a plasmid contains a self-encoded removable fragment (SERF) flanked by inverted BsaI recognition sites. Thus, BsaI digestion simultaneously removes the SERF and BsaI sites. This plasmid, referred to as an entry vector, was one of several previously used to generate all one amino acid mutations of the MS2 CP. Two single-stranded DNA primers with $(NNK)_3$ extensions, either with or without an N-terminal proline, were purchased, resuspended, and diluted to final concentrations of 50 ng/µL. The reverse strand was completed by overlap extension PCR with a corresponding reverse primer. Subsequently, the double-stranded DNA was purified by PCR clean-up (Promega, Cat# A9282), diluted to 1-5 ng/µL, then cloned into the entry vector using Golden Gate cloning. Ligated plasmids were incubated on desalting membranes (Millipore Sigma, Cat# VSWP02500) for 20 min, followed by transformation into store-bought electrocompetent DH10B *E. coli* cells (Invitrogen, Cat# 18290015) for the $P(NNK)_3$ extension, or homemade electrocompetent DH10B *E. coli* cells for the $(NNK)_3$ extension. Following electroporation and recovery, cells were plated onto two LB-A plates (VWR, Cat# 82050-600) containing 32 µg/mL chloramphenicol. Cells were grown overnight at 37 ºC. Colony number varied, but every transformation produced at least 3x the library size. The protocol was repeated in full for three (for $(NNK)_3$ extensions) or two (for $P(NNK)_3$ extensions) total biological replicates that are entirely independent from library generation through selection.

4.D.ii. Assembly selection

Colonies were harvested by scraping plates into LB-M and growing the mixture for 2 h. Each library was subcultured 1:100 into 1 L of 2xYT (Teknova, Cat# Y0210) and grown to an OD of 0.6, then induced with 0.1% arabinose. Variant libraries were expressed overnight at 37 ºC. Cells were harvested by centrifugation and lysed by sonication. Libraries were subjected to two rounds of ammonium sulfate precipitation (50% saturation), followed by FPLC size exclusion chromatography purification to select for well-assembled VLPs.

4.D.iii. Heat selection

Assembly-selected libraries were subjected to 50 ºC for 10 min. Precipitated VLPs were pelleted via centrifugation, and assembled VLPs were isolated via semi-preparative HPLC size exclusion chromatography. Fractions containing assembled VLPs (at the characteristic elution time of 11.2 min) were combined, subjected to RNA extraction, barcoded, and identified via high-throughput sequencing.

4.D.iv. Chemical modification selection

P(NNK)$_3$ libraries (final concentration 200 µM) were added to a solution of aminophenol-DNA (final concentration 1 mM) in 10 mM phosphate buffer, pH 7.2. $K_3Fe(CN)_6$ in Milli-Q water was added (final concentration 5 mM) and the solution was incubated at room temperature for 30 min. The reaction was quenched with tris(2-carboxyethyl)phosphine (TCEP, final concentration 50 mM) and excess oxidant and DNA were removed using a 100 kDa MWCO filter spin filter (EMD Millipore, Burlington, MA). Well-formed VLPs were isolated by semi-preparative HPLC SEC, sample preparation, and high-throughput sequencing as described.

4.D.v. Sample prep for high-throughput sequencing

Plasmid DNA was isolated prior to expressions using Zyppy Plasmid Miniprep Kit (Zymo, Cat# D4036). RNA was extracted from the assembly-selected libraries using previously-published protocols[23,47]. Briefly, homogenization was carried out with TRIzol (Thermo Fisher Cat# 15596026), followed by chloroform addition. The sample was centrifuged to separate into aqueous, interphase, and organic layers. The aqueous layer, containing RNA, was isolated, and the RNA was precipitated with isopropanol and washed with 70% ethanol. RNA was dried and resuspended in RNAse free water. cDNA was synthesized with the Superscript III first strand cDNA synthesis kit from Life (cat: 18080051, polyT primer). PolyT primers have historically been used with success for SyMAPS projects[23,24], likely because a small percent of the *E. coli* transcriptome is known to be polyadenylated in small amounts[48,49]. A head-to-head comparison of random hexamer primers with polyT primers shows that both successfully produce cDNA that can be used for downstream high-throughput sequencing steps (**Supplementary Fig. 4.20**). Though unusual, this low level of polyadenylation, coupled with high expression levels achieved in the library generation stage, has led to the successful use of polyT primers. Future work will further compare polyT vs. random hexamer primers in this system.

cDNA and plasmids were both barcoded for high-throughput sequencing. Both types of samples were amplified with two rounds of PCR (10 cycles, followed by 8 cycles) to add barcodes and Illumina sequencing handles, following Illumina 16S Metagenomic Sequencing Library Preparation recommendations (**Supplementary Data 4.2**). Libraries were quantified with Qubit and combined in equal molar ratio. Libraries were analyzed by 150 PE MiSeq in collaboration with the UC Davis Sequencing Facilities. A total of 18 million reads passed filter, and had an overall Q30 > 85%.

4.D.vi. Individual variant cloning

Individual variants were cloned through a variation of the methods described above. Briefly, overlap extension PCR was used to produce double-stranded fragment that spanned the length of the missing 26-codon region in the entry vector (**Supplementary Data 4.2**). Each fragment was cloned into the entry vector using standard Golden Gate cloning techniques[50]. Plasmids were then transformed into chemically-competent DH10B cells. Plasmid identities were confirmed by Sanger sequencing prior to expression. Plasmids with multiple mutations (i.e. CP[HiPerX–S37P] and CP[HiPerX–N87C]) were cloned into a similar Entry Vector bearing the desired mutation at position 37 or 87, which

was installed via site-directed mutagenesis[51]. Glycosylation sequences were cloned as described here, and Qβ HiPerX variants were cloned with overlap extension PCR. All primer sequences can be found in **Supplementary Data 4.2**.

4.D.vii. Individual variant expression

Selected variants were expressed individually in 5 or 50 mL cultures of 2xYT. Expressions were lysed by sonication and precipitated twice with 50% ammonium sulfate. Variants were evaluated for assembly by HPLC SEC and thermostability by native gel.

4.D.viii. Oxidative coupling of HiPerX MS2 variants

To a solution of HiPerX MS2 (final concentration 10 μM, 1 eq.) in 10 mM sodium phosphate buffer at pH 7.2 was added 4-methylcatechol in DMF (final concentration 100 μM, 10 eq.). To initiate oxidation, $K_3Fe(CN)_6$ in Milli-Q water was added (final concentration 1 mM, 100 eq.) and the solution was incubated at room temperature for 30 min. The reaction was quenched with tris(2-carboxyethyl)phosphine (TCEP, final concentration 10 mM) and excess oxidant and DNA were removed using a 100 kDa MWCO filter spin filter. Percent conversion was analyzed by ESI-TOF-LC/MS analysis and VLP integrity was confirmed by HPLC SEC.

4.D.ix. Enzyme-catalyzed oxidative coupling of HiPerX MS2 variants

A solution of HiPerX MS2 (final concentration 10 μM, 1 eq.) in 50 mM sodium phosphate buffer at pH 6.5 was supplied with 4-methylcatechol in DMF (final concentration 100 μM, 10 eq.). To initiate the oxidative coupling reaction, tyrosinase enzyme (final concentration 0.5 μM) was added to the reaction mixture. After 2 h of incubation at room temperature, the reaction was quenched with a solution of tropolone and TCEP (final concentration 2.1 mM tropolone and 2.1 mM TCEP). Excess 4-methylcatechol was removed using a 100 kDa MWCO filter spin filter. The percent conversion was determined using ESI-TOF-LC/MS analysis and VLP integrity was confirmed by HPLC SEC.

4.D.x. 2PCA modification of HiPerX MS2 variants

A 10 mM solution of 2-pyridiencarboxaldehyde (2PCA) in water was added to a solution of HiPerX MS2 (final concentrations: 50 μM MS2 CP, 12.5 mM 2PCA) in 10 mM sodium phosphate buffer, pH 7.2 with 100 mM NaCl. The reaction was allowed to proceed for 18 h at room temperature. Excess 2-pyridinecarboxaldehyde was removed using a 100 kDa MWCO filter spin filter. Percent conversion was analyzed by ESI-TOF-LC/MS, and VLP assembly was assessed by HPLC SEC.

4.D.xi. MS2-fluorophore labeling

P-X-X-R/N87C MS2 (final concentration 100 μM) was mixed with Alexa Fluor 488-Maleimide (stock solution 10 mM in DMSO, final concentration 100 μM, Invitrogen, Cat# A10254) in 50 mM phosphate buffer, pH 7.2. Solution was briefly vortexed then incubated at room temperature for 1 h. Excess dye was removed by Nap-5 size exclusion column (GE Healthcare, Cat# 17-0853-01) equilibrated with 10 mM phosphate buffer, pH 7.2. Modification of assembled VLPs was verified by HPLC SEC and quantified by ESI-

MS.

4.D.xii. FPLC SEC

MS2 CP libraries and select individual variants were purified on an Akta Pure 25 L Fast Protein Liquid Chromatography (FPLC) system with a HiPrep Sephacryl S-500 HR column (GE Healthcare Life Sciences, Cat# 28935607) Size Exclusion Chromatography (SEC) column via isocratic flow with 10 mM phosphate buffer at pH 7.2 with 200 mM sodium chloride and 2 mM sodium azide.

4.D.xiii. HPLC SEC

Variants or libraries were analyzed on an Agilent 1290 Infinity HPLC with an Agilent Bio SEC-5 column (5 um, 2000Å, 7.8x300mm) with isocratic flow of 10 mM phosphate buffer at pH 7.2 with 200 mM sodium chloride and 2 mM sodium azide. Fractions were harvested at 11.2 min, or the characteristic elution time for wild-type MS2. Harvested VLPs were then subjected to RNA extraction and high-throughput sequencing sample preparation.

4.D.xiv. Native gel

Variants were analyzed in a 0.8% agarose gel in 0.5X TBE buffer (45 mM Tris-borate, 1 mM EDTA) and 2X SYBR Safe DNA Gel Stain (ThermoFisher Scientific, Cat# S33102) for 120 min at 40 V. Agarose gels were imaged on a BioRad GelDoc EZ Imager. Densitometry with ImageJ was carried out to compare experimental conditions to pH 7.4 in each case.

4.D.xv. ESI-TOF

Modified and unmodified variants were analyzed with an Agilent 1200 series liquid chromatograph (Agilent Technologies, USA) connected in-line with an Agilent 6224 Time-of Flight (TOF) LC/MS system with a Turbospray ion source.

4.D.xvi. Strains

All library experiments were conducted with MegaX DH10B *E. coli* electrocompetent cells (ThermoFisher Scientific, Cat# C640003). Chemically-competent DH10B cells were used for expression of individual variants of interest. Overnight growths from a single colony were incubated for 16-20 h at 37 °C shaking at 200 RPM in LB-Miller media (Fisher Scientific, Cat# BP1426-2) with chloramphenicol at 32 mg/L. Expressions were subcultured 1:100 into 2xYT media (Teknova, Cat# Y0210) with 32 mg/L chloramphenicol, allowed to grow to an OD600 of 0.6, then induced with 0.1% arabinose. Expressions continued overnight at 37 °C shaking at 200 RPM.

4.D.xvii. High-throughput sequencing data analysis

Data were trimmed and processed described previously,[1] with minor variation. Briefly, data were trimmed with Trimmomatic[52] with a 4-unit sliding quality window of 20 and a minimum length of 32. Reads were merged with FLASh (fast length adjustment of short reads)[53] with a maximum overlap of 150 base pairs. Reads were sorted and indexed with

SAMtools[54] and unmapped reads were filtered with the Picard function CleanSam. Reads shorter than 106 base pairs were removed. Reads were processed with in-house code to produce various AFLs.

4.D.xvii. AFL calculations

Cleaned and filtered high-throughput sequencing reads were analyzed using Python programs written in-house. Briefly, the N-terminal region of the MS2 CP was located, and the three codons following the proline at the new N terminus were recorded. Codons were then translated into amino acids to generate counts for each tripeptide combination.

These calculations were repeated for all experiments to generate abundances before and after each selective pressure. Relative abundances were calculated similarly to the previously described protocol[23]. Briefly, the relative percent abundance of a selection was calculated by dividing the percent abundance generated from an assembled VLP, heat-selected, or chemically-selected library by the plasmid percent abundance from its respective biological replicate. Two biological replicates were assessed for each challenge except for the heat-selected library, where one biological replicate was analyzed. All Nan (null) values, which indicate variants that were not identified in the plasmid library were ignored. Scores of zero, which indicate variants that were sequenced in the unselected library but absent in the VLP library, were replaced with an arbitrary score of 0.0001. We calculated the mean relative abundances across replicates. We then calculated the log10 of the Relative Abundance array to calculate the final array for each replicate. The log10 relative abundance value for an individual variant in a particular challenge is referred to as its Apparent Fitness Score (AFS). Arrays are displayed in **Figure 4.3** and **Supplementary Fig. 4.1, 4.2, 4.11, and 4.12**. Heatmaps for each of five individual biological replicates (two replicates of P-X-X-X-MS2 and three replicates of X-X-X-MS2) are shown in **Supplementary Fig. 4.21-25**, and the correlation between all replicates are shown in **Supplementary Fig. 4.4**, with plasmid abundance cutoffs indicated where relevant. For the correlation analyses, any extension with a –4 or nan score in either replicate was not included. All values from each individual replicate and the mean AFL values are included in **Supplementary Data 4.1**.

4.D.xvii. High-stringency heatmap calculation

Two additional data processing methods were applied to the P-X-X-X libraries in order reduce the impact of stochastic variation on the generated heatmaps show in **Supplementary Fig. 4.5-4.10**. The first set of heatmaps were produced by removing low read count variants from the dataset. Variants with three or fewer reads in the plasmid library of any individual replicate were treated as null values in the average heatmap and are shown in green (**Supplementary Fig. 4.5-4.7**).

The second processing method removed ambiguous fitness scores and simplified scoring to a positive (blue) or negative (red) score. In this filter, any variant with a fitness score between –0.2 and 0.2 in any individual replicate was considered a null value and colored in grey. Variants with scores of opposite charge between replicates were also removed. The remaining variants were marked blue if their fitness scores were >0.2 in all replicates or marked red if their fitness scores were <–0.2 in all replicates (**Supplementary**

Fig. 4.8-4.10).

4.D.xviii. Making the aggregate AFL

The aggregate AFL was produced using in-house Python code. In brief, variants were sorted by their AFS values in assembly, heat, and chemical selection. Variants with an AFS value greater than or equal to 1.0 in all three criteria were designated Highest Performing Extensions (HiPerX). Variants with AFS values of less that 1.0 in one or more challenge(s) were sorted as shown in **Figure 4.4B**. If a variant returned a null value for any of the three challenges it was sorted into the not observed category.

4.D.xix. Random selection of 5 HiPerX variants

All HiPerX variants of the form P-X-X-R-MS2 were assigned a unique value from 1-65. Five random integers between 1-65 were produced using Python's *randint*() function and the corresponding variants were subsequently cloned and expressed.

4.D.xx. Computational modeling of extended MS2 coat proteins

A hexameric subunit of the MS2 capsid was prepared in based off of the crystal structure of wildtype MS2 (PDB ID:2MS2). The structure was imported into Schrodinger's Maestro suite and various N-terminal extensions were constructed with the build tool. Preprocessing of the extended structures was performed with Maestro's protein preparation wizard. In brief, hydrogen bonds were assigned with the H-bond optimization tool at a PROPKA pH of 7. Subsequently, a restrained minimization of the structure using an OPLS3e forcefield was performed. MacroModel was then used to carry out a Large-scale Low Mode conformational search of the minimized structures. All residues within 10 Å of the N-terminal extensions were restrained with a force constant of 200 kJ/mol. Atoms beyond this subshell were frozen in place. Sampling used 1000 maximum steps with 100 steps per rotatable bond.

4.E. References

1. Obermeyer, A. C., Jarman, J. B. & Francis, M. B. N-Terminal Modification of Proteins with *o*-Aminophenols. *J. Am. Chem. Soc.* **136**, 9572–9579 (2014).
2. MacDonald, J. I., Munch, H. K., Moore, T. & Francis, M. B. One-step site-specific modification of native proteins with 2-pyridinecarboxyaldehydes. *Nat. Chem. Biol.* **11**, 326–331 (2015).
3. Witus, L. S., Netirojjanakul, C., Palla, K. S., Muehl, E. M., Weng, C. H., Iavarone, A. T. & Francis, M. B. Site-specific protein transamination using N-methylpyridinium-4-carboxaldehyde. *J. Am. Chem. Soc.* **135**, 17223–17229 (2013).
4. Sur, S., Qiao, Y., Fries, A., O'Meally, R. N., Cole, R. N., Kinzler, K. W., Vogelstein, B. & Zhou, S. A . Protein Bioconjugation Method with Exquisite N-Terminal Specificity. *Sci. Rep.* **5**, 1–7 (2015).
5. Spicer, C. D., Pashuck, E. T. & Stevens, M. M. Achieving Controlled Biomolecule-Biomaterial Conjugation. *Chem. Rev.* **118**, 7702–7743 (2018).
6. Li, X., Zhang, L., Hall, S. E. & Tam, J. P. A new ligation method for N-terminal tryptophan-containing peptides using the Pictet-Spengler reaction. Pergamon *Tetrahedron Lett.* **41**, (2000).
7. Geoghegan, K. F. & Stroh, J. G. Site-Directed Conjugation of Nonpeptide Groups to Peptides and Proteins via Periodate Oxidation of a 2-Amino Alcohol. Application to Modification at N-Terminal

Serine. *Bioconjugate Chem* **3**, 138–146 (1992).

8. Casi, G., Huguenin-Dezot, N., Zuberbühler, K., Scheuermann, J. & Neri, D. Site-specific traceless coupling of potent cytotoxic drugs to recombinant antibodies for pharmacodelivery. *J. Am. Chem. Soc.* **134**, 5887–5892 (2012).
9. Palla, K. S., Witus, L. S., Mackenzie, K. J., Netirojjanakul, C. & Francis, M. B. Optimization and Expansion of a Site-Selective N -Methylpyridinium-4-carboxaldehyde-Mediated Transamination for Bacterially Expressed Proteins. *J. Am. Chem. Soc.* **137**, 1123–1129 (2015).
10. Rosen, C. B. & Francis, M. B. Targeting the N terminus for site-selective protein modification. *Nat. Chem. Biol.* **13**, 697–705 (2017).
11. Li, D., Han, B., Wei, R., Yao, G., Chen, Z., Liu, J., Poon, T. C. W., Su, W., Zhu, Z., Dimitrov, D. S. & Zhao, Q. N-terminal α-amino group modification of antibodies using a site-selective click chemistry method. *MAbs* **10**, 712–719 (2018).
12. Lee, J. P., Kassianidou, E., Macdonald, J. I., Francis, M. B. & Kumar, S. N-terminal specific conjugation of extracellular matrix proteins to 2-pyridinecarboxaldehyde functionalized polyacrylamide hydrogels. *Biomaterials* **102**, 268–276 (2016).
13. Esser-Kahn, A. P., Iavarone, A. T. & Francis, M. B. Metallothionein-cross-linked hydrogels for the selective removal of heavy metals from water. *J. Am. Chem. Soc.* **130**, 15820–15822 (2008).
14. ElSohly, A. M., Netirojjanakul, C., Aanei, I. L., Jager, A., Bendall, S. C., Farkas, M. E., Nolan, G. P. & Francis, M. B. Synthetically Modified Viral Capsids as Versatile Carriers for Use in Antibody-Based Cell Targeting. *Bioconjug. Chem.* **26**, 1590–1596 (2015).
15. Ashley, C. E., Carnes, E. C., Phillips, G. K., Durfee, P. N., Buley, M. D., Lino, C. A., Padilla, D. P., Phillips, B., Carter, M. B., Willman, C. L., Brinker, C. J., Caldeira, J. D. C., Chackerian, B., Wharton, W. & Peabody, D. S. Cell-specific delivery of diverse cargos by bacteriophage MS2 virus-like particles. *ACS Nano* **5**, 5729–5745 (2011).
16. Aanei, I. L., Huynh, T., Seo, Y. & Francis, M. B. Vascular Cell Adhesion Molecule-Targeted MS2 Viral Capsids for the Detection of Early-Stage Atherosclerotic Plaques. *Bioconjug. Chem*. **29**, 2526–2530 (2018).
17. Farkas, M. E., Aanei, I. L., Behrens, C. R., Tong, G. J., Murphy, S. T., Neil, J. P. O. & Francis, M. B. PET Imaging and Biodistribution of Chemically Modified Bacteriophage MS2. *Mol. Pharm.* **10**, 69–76 (2013).
18. Crossey, E., Frietze, K., Narum, D. L., Peabody, D. S. & Chackerian, B. Identification of an immunogenic mimic of a conserved epitope on the plasmodium falciparum blood stage antigen ama1 using virus-like particle (VLP) peptide display. *PLoS One* **10**, 1–19 (2015).
19. Zhai, L., Peabody, J., Pang, Y. Y. S., Schiller, J., Chackerian, B. & Tumban, E. A novel candidate HPV vaccine: MS2 phage VLP displaying a tandem HPV L2 peptide offers similar protection in mice to Gardasil-9. *Antiviral Res*. **147**, 116–123 (2017).
20. Capehart, S. L., Coyle, M. P., Glasgow, J. E. & Francis, M. B. Controlled Integration of Gold Nanoparticles and Organic Fluorophores Using Synthetically Modified MS2 Viral Capsids. *J. Am. Chem. Soc*. **135**, 3011–3016 (2013).
21. Glasgow, J. E., Asensio, M. A., Jakobson, C. M., Francis, M. B. & Tullman-Ercek, D. Influence of Electrostatics on Small Molecule Flux through a Protein Nanoreactor. *ACS Synth. Biol*. **4**, 1011–1019 (2015).
22. Glasgow, J. E., Capehart, S. L., Francis, M. B. & Tullman-Ercek, D. Osmolyte-mediated encapsulation of proteins inside MS2 viral capsids. *ACS Nano* **6**, 8658–8664 (2012).
23. Hartman, E. C., Jakobson, C. M., Favor, A. H., Lobba, M. J., Álvarez-Benedicto, E., Francis, M. B.

& Tullman-Ercek, D. Quantitative characterization of all single amino acid variants of a viral capsid-based drug delivery vehicle. *Nat. Commun.* **9**, 1385 (2018).

24. Hartman, E. C., Lobba, M. J., Favor, A. H., Robinson, S. A., Francis, M. B. & Tullman-Ercek, D. Experimental Evaluation of Coevolution in a Self-Assembling Particle. *Biochemistry* (2018). *ASAP*.

25. Peabody, D. S. Subunit fusion confers tolerance to peptide insertions in a virus coat protein. *Arch. Biochem. Biophys.* **347**, 85–92 (1997).

26. Hooker, J. M., Esser-Kahn, A. P. & Francis, M. B. Modification of aniline containing proteins using an oxidative coupling strategy. *J. Am. Chem. Soc.* **128**, 15558–15559 (2006).

27. Behrens, C. R., Hooker, J. M., Obermeyer, A. C., Romanini, D. W., Katz, E. M. & Francis, M. B. Rapid chemoselective bioconjugation through oxidative coupling of anilines and aminophenols. *J. Am. Chem. Soc.* **133**, 16398–16401 (2011).

28. Ni, C. Z., Syed, R., Kodandapani, R., Wickersham, J., Peabody, D. S. & Ely, K. R. Crystal structure of the MS2 coat protein dimer: implications for RNA binding and virus assembly. *Structure* **3**, 255–63 (1995).

29. Hirel, P.-H., Schmitter, J.-M., Dessen, P., Fayat, G. & Blanquet, S. Extent of N-terminal methionine excision from *Escherichia coli* proteins is governed by the side-chain length of the penultimate amino acid. *Proc. Natl. Acad. Sci USA* **86**, 8247–8251 (1989).

30. Maza, J., Bader, D., L. V., Xiao, L., Marmelstein, A. M., Brauer, D. D., ElSohly, A. M., Smith, M. J., Francis, M. B. Enzymatic Modification of N-Terminal Proline Residues Using Simple Phenol Derivatives. *ChemRxiv Preprint*.

31. ElSohly, A. M., MacDonald, J. I., Hentzen, N. B., Aanei, I. L., El Muslemany, K. M. & Francis, M. B. *ortho*-Methoxyphenols as Convenient Oxidative Bioconjugation Reagents with Application to Site-Selective Heterobifunctional Cross-Linkers. *J. Am. Chem. Soc.* **139**, 3767–3773 (2017).

32. Furst, A. L., Smith, M. J., Lee, M. C. & Francis, M. B. DNA Hybridization To Interface Current-Producing Cells with Electrode Surfaces. *ACS Cent. Sci* **4**, 884 (2018).

33. Obermeyer, A. C., Jarman, J. B., Netirojjanakul, C., El Muslemany, K. & Francis, M. B. Mild bioconjugation through the oxidative coupling of *ortho*-aminophenols and anilines with ferricyanide. *Angew. Chem. Int. Ed.* **53**, 1057–61 (2014).

34. Reddy Chichili, V. P., Kumar, V. & Sivaraman, J. Linkers in the structural biology of protein-protein interactions. *Protein Sci*. **22**, 153–167 (2013).

35. Klein, J. S., Jiang, S., Galimidi, R. P., Keeffe, J. R. & Bjorkman, P. J. Design and characterization of structured protein linkers with differing flexibilities. *Protein Eng. Des. Sel.* **27**, 325–330 (2014).

36. Isom, D. G., Castaneda, C. A., Cannon, B. R., Velu, P. D. & Garcia-Moreno E., B. Charges in the hydrophobic interior of proteins. *Proc. Natl. Acad. Sci.* **107**, 16096–16100 (2010).

37. Dwyer, J. J., Gittis, A. G., Karp, D. A., Lattman, E. E., Spencer, D. S., Stites, W. E. & García-Moreno, E. B. High apparent dielectric constants in the interior of a protein reflect water penetration. *Biophys. J.* **79**, 1610–1620 (2000).

38. Tong, G. J., Hsiao, S. C., Carrico, Z. M. & Francis, M. B. Viral capsid DNA aptamer conjugates as multivalent cell-targeting vehicles. *J. Am. Chem. Soc.* **131**, 11174–11178 (2009).

39. Stephanopoulos, N., Tong, G. J., Hsiao, S. C. & Francis, M. B. Dual-surface modified virus capsids for targeted delivery of photodynamic agents to cancer cells. *ACS Nano* **4**, 6014–6020 (2010).

40. Aanei, I. L., Francis, M. B. Dual Surface Modification of Genome-Free MS2 Capsids for Delivery Applications. *In Virus-Derived Nanoparticles for Advanced Technologies*; Wege, C., Lomonossoff, G. P., Eds.; *Springer New York*: New York, NY, 2018; Vol. 1776, pp 629–642.

41. Kightlinger, W., Lin, L., Rosztoczy, M., Li, W., DeLisa, M. P., Mrksich, M. & Jewett, M. C. Design of glycosylation sites by rapid synthesis and analysis of glycosyltransferases. *Nat. Chem. Biol.* **14**, 627–635 (2018).

42. Yin, Z., Wu, X., Kaczanowska, K., Sungsuwan, S., Comellas Aragones, M., Pett, C., Yu, J., Baniel, C., Westerlind, U., Finn, M. G. & Huang, X. Antitumor Humoral and T Cell Responses by Mucin-1 Conjugates of Bacteriophage Qβ in Wild-type Mice. *ACS Chem. Biol.* **13**, 1668–1676 (2018).

43. Brito, C. R. N., McKay, C. S., Azevedo, M. A., Santos, L. C. B., Venuto, A. P., Nunes, D. F., D'Ávila, D. A., Rodrigues da Cunha, G. M., Almeida, I. C., Gazzinelli, R. T., Galvão, L. M. C., Chiari, E., Sanhueza, C. A., Finn, M. G. & Marques, A. F. Virus-like Particle Display of the α-Gal Epitope for the Diagnostic Assessment of Chagas Disease. *ACS Infect. Dis.* **2**, 917–922 (2016).

44. Asensio, M. A., Morella, N. M., Jakobson, C. M., Hartman, E. C., Glasgow, J. E., Sankaran, B., Zwart, P. H. & Tullman-Ercek, D. A Selection for Assembly Reveals That a Single Amino Acid Mutant of the Bacteriophage MS2 Coat Protein Forms a Smaller Virus-like Particle. *Nano Lett.* **16**, 5944–5950 (2016).

45. Gaumet, M., Vargas, A., Gurny, R. & Delie, F. Nanoparticles for drug delivery: The need for precision in reporting particle size parameters. *Eur. J. Pharm. Biopharm.* **69**, 1–9 (2008).

46. Hietpas, R. T.; Jensen, J. D.; Bolon, D. N. A. Experimental Illumination of a Fitness Landscape. *Proc. Natl. Acad. Sci. U. S. A.* **108** (19), 7896–7901 (2011).

47. Pinto, F.; Thapper, A.; Sontheim, W.; Lindblad, P. Analysis of Current and Alternative Phenol Based RNA Extraction Methodologies for Cyanobacteria. *BMC Mol. Biol.* **10** (1), 79 (2009).

48. Raynal, L. C. & Carpousis, A. J. Poly(A) polymerase I of *Escherichia coli*: Characterization of the catalytic domain, an RNA binding site and regions for the interaction with proteins involved in mRNA degradation. *Mol. Microbiol.* **32**, 765–775 (1999).

49. Kushner, S. R. Polyadenylation in *E. coli*: A 20 year odyssey. *RNA* **21**, 673–674 (2015).

50. Engler, C., Gruetzner, R., Kandzia, R. & Marillonnet, S. Golden gate shuffling: a one-pot DNA shuffling method based on type IIs restriction enzymes. *PLoS One* **4**, e5553 (2009).

51. Wang, W. & Malcolm, B. A. Two-Stage PCR Protocol Allowing Introduction of Multiple Mutations, Deletions and Insertions Using QuikChangeTM Site-Directed Mutagenesis. *Biotechniques* **26**, 680–682 (1999).

52. Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: A flexible trimmer for Illumina sequence data. *Bioinformatics* **30**, 2114–2120 (2014).

53. Magoč, T. & Salzberg, S. L. FLASH: Fast length adjustment of short reads to improve genome assemblies. *Bioinformatics* **27**, 2957–2963 (2011)

54. Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G. & Durbin, R. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).

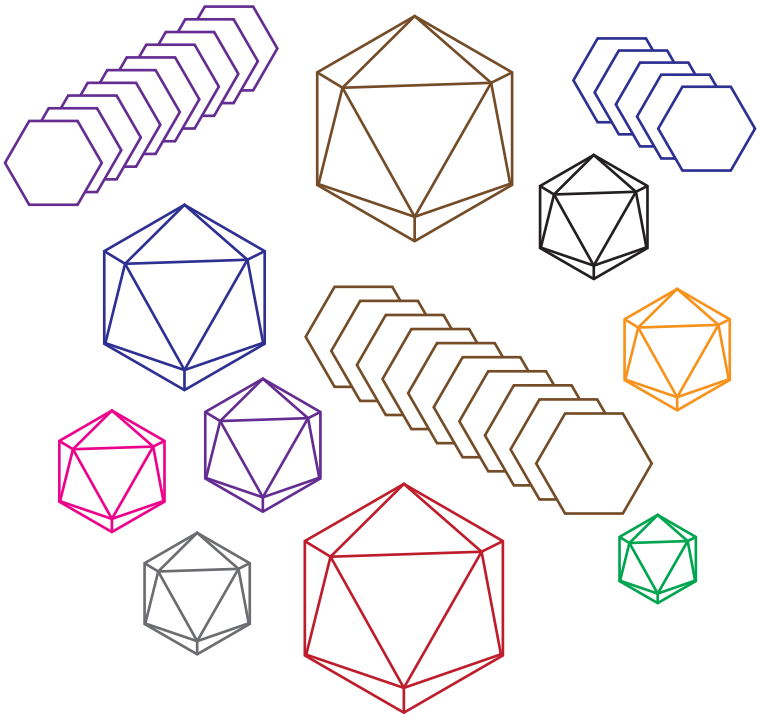
Chapter 5: Design Rules for Altering the Quaternary Size and Shape of the MS2 CP

Short summary

A point mutation to the MS2 bacteriophage confers a geometric shift from a 27 nm to a 17 nm particle. We explored how a similar mutation affected related bacteriophages, as well as how other mutations alter the geometry of the MS2 coat protein.

Abstract:

Virus-like particles derived from the MS2 bacteriophage are promising nanocontainers for drug delivery. However, recent studies suggest that extravasation may be limiting the effectiveness of these carriers in mouse models. Size and shape are known to impact efficiency of extravasation, and recently, a homogeneous size shift to a 17 nm particle was described for MS2. Here, we sought to understand this size shift using two complementary methods. We evaluated the effect of this mutation in related bacteriophages with sequence similarity ranging from 25% to 85%. We also generated a comprehensive library in the MS2 CP to determine how other mutations affect size and shape. We found that genetic background alters the behavior of the size shift mutation. We also identified a new mutation that confers a stables shift to long, thin, apparently hollow fibrils. Taken together, these efforts work to better describe the rules that govern size and shape in a virus-like particle.



5.A. Introduction

Virus-like particles (VLPs) derived from the MS2 bacteriophage coat protein (CP) are promising vehicles for drug delivery and imaging[1–6]. These nanocontainers can protect cargo during delivery[7], carry a large cargo load[8-11], and display targeting agents[3,8,12]. However, despite much interest in repurposing these structures as cancer therapeutics, MS2 VLPs have yet to have a significant clinical impact, and no drug using this or other bacteriophage scaffold has been approved by the FDA.

Studies in mouse models suggest that further optimization of scaffold properties is needed to maximize the impact of these carriers[4,6,13]. In 2016, our lab evaluated the therapeutic potential of MS2 VLPs in a mouse model of breast cancer[4]. These VLPs were loaded with Cu-64 bound to a chelating agent, coated with an anti-EGFR antibody, and administered to mice with breast cancer tumors. Despite a long circulation time (>24h), MS2 did not preferentially accumulate in the tumor environment. In contrast, VLPs targeting atherosclerotic plaque material—a target found in the bloodstream—did preferentially accumulate in or around the atherosclerotic cells when coated with a targeting antibody or peptide[6]. Together, these studies suggest that MS2 VLPs may not be effectively exiting the bloodstream, which hinders accumulation in the tumor microenvironment.

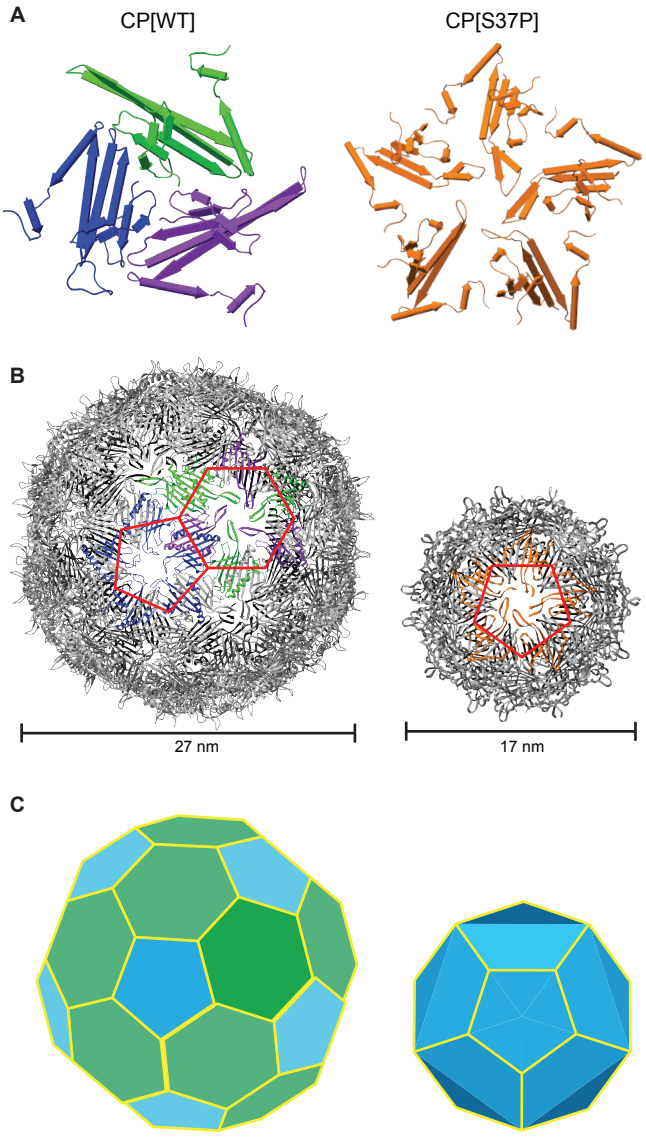


Figure 5.1. Altered quaternary geometry from a one amino acid mutation of the MS2 CP. The (A) Asymetric unit, (B) VLP structure, and (C) icosahedral symmetry are shown for CP[WT] and CP[S37P].

While extravasation is impacted by many variables, the size and shape of the carrier are known to be critical factors[14,15]. However, isolating size or shape as an independent variable is challenging, and most studies either compare heterogeneous materials[14,16] or change the identity of the scaffold to achieve size and shape change[13]. Recently, our lab identified a single amino acid variant, MS2 CP[S37P], that causes a homogeneous size shift from a 27 nm to a 17 nm particle (**Figure 5.1**)[18]. This variant maintains nearly identical protein sequence and remarkably similar secondary and tertiary structure to the parent coat

protein; however, the quaternary structure changes dramatically. This set of variants enables researchers, for the first time, to isolate the impact of size on carrier efficiency and tumor accumulation.

To date, MS2 is the only structure with a variant known to confer a uniform change in the icosahedral symmetry of a VLP, and much is still unknown about how mutations impact quaternary structure of self-assembled particles. Here, we used the MS2 VLP as a model to better understand how mutations affect the size and shape of a VLP. We pursued two complementary approaches. First, we determined if rational incorporation of an analogous mutation in related bacteriophages would confer the same homogeneous shift in VLP structure. In addition, we evaluated how other mutations at position 36 and 37 in the MS2 CP affected quaternary structure. Through this effort, we found that hydrophobic mutations at position 37 lead to fibril-like structures that appear to be hollow, identifying a new and potentially useful quaternary geometric structure of the MS2 CP.

5.B. Results and Discussion

5.B.i. Mutation in evolutionarily related bacteriophages do not confer similar geometric shift

Many other closely-related bacteriophages can form VLPs by overexpressing a coat protein in *E. coli*, and a number of these are structurally similar to the MS2 VLP. We sought to better understand the importance of the DE loop, which contains residue 37, in structurally similar bacteriophages by rationally introducing a proline mutation in the same loop in other VLPs. We hypothesized that a proline mutation in closely-related bacteriophages could lead to similar changes in quaternary geometry.

We selected four bacteriophages that satisfied the following criteria: at least 20% coat protein sequence similarity to MS2; VLP formation following overexpression in *E. coli*; and solved capsid structures. From this set of criteria, we selected VLPs derived from fr (87% identical to MS2), GA (60% identical), Qβ (20% identical), and Prr1 (27% identical) bacteriophages for further study.

To determine the most similar place to insert a proline mutation, the structure and sequence of these coat proteins were evaluated via sequence alignment (**Supplementary Fig. 5.1A**) and structural analyses (**Supplementary Fig. 5.1B**). Positions analogous to position 37 in MS2 were identified (**Table 5.1**). First, we evaluated whether CP[WT] for all bacteriophages could be expressed to high yield in *E. coli*. We found that wild-type VLPs for all four virus-like particles were expressed to relatively high yield in *E. coli*, though fr bacteriophage yield was lower than the other three. Following purification, virus-like particle formation was confirmed by transmission electron microscopy (TEM) (**Figure 5.2A-E**).

We next evaluated whether proline mutations altered the quaternary geometry of these VLPs. Site-directed mutagenesis was used to incorporate proline mutations at each specified position, and the quaternary structure of each mutation was then evaluated. As with the wild-type virus-like particles, the mutants were expressed and purified, and the size and shape of virus-like particles were determined via TEM. Surprisingly, the proline mutations resulted in a variety of phenotypes across the four different bacteriophages. As described previously, in MS2, CP[S37P] results in a homogeneous shift to particle

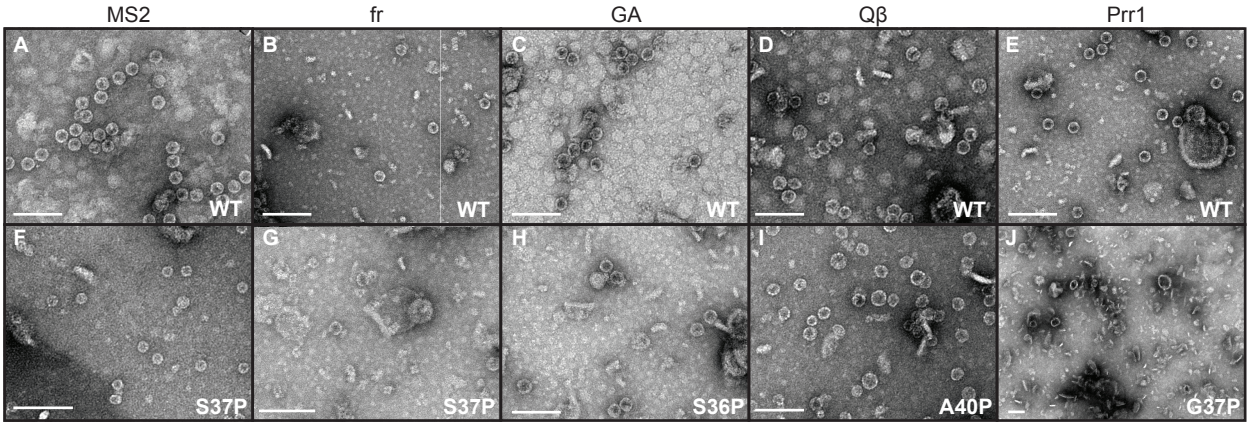


Figure 5.2. Effect of DE loop mutations on VLPs formed from bacteriophage coat proteins. TEM images of (A) MS2 WT, (B) fr WT, (C) GA WT, (D) Qβ WT, and (E) Prr1 WT. These can be compared to (F) MS2 S37P, (G) fr S37P, (H) GA S36P, (I) Qβ A40P, and (J) Prr1 G37P.

that is 17 nm in diameter (**Figure 5.2F**). In fr (**Figure 5.2G**) and GA (**Figure 5.2H**), we see no size shift, and the particles retain the wild-type-sized 27 nm diameter. In Prr1, no particles are visible, likely indicating that the mutation disrupts capsid formation (**Figure 5.2J**). However, in the Qβ—which is most distant in sequence identity to the MS2 CP—the CP[A40P] results in a distribution of sizes by TEM (**Figure 5.2I**). In general, these surprising results suggest that sequence similarity does not provide any predictive power for how these VLPs would behave following mutation to proline in the DE loop.

We sought to quantify the distribution of VLP sizes in Qβ CP[A40P] using several different techniques. First, TEM images were used to quantify VLP diameters of MS2[WT], MS2[S37P], GA[WT], GA[S36P], Qβ[WT], AND Qβ[A40P] in triplicate. While the average size of Qβ[A40P] and Qβ[WT] is approximately the same, the variance in Qβ[A40P] is higher, while the variance for GA[S36P] is not higher than GA[WT] (**Supplementary Fig. 5.2A,B**). HPLC SEC was used to separate the VLP peak into seven discrete fractions, labeled as H4 (earliest fraction) through H10 (latest fraction) (**Supplementary Fig. 5.2C**). These fractions, when re-evaluated by HPLC SEC, appeared to separate into populations enriched for large and small species, respectively. This suggests that the size shift is stable, and populations, once fractionated, do not quickly reform the entire size distribution. Taken together, these data suggest that discrete and stable changes in size are detected in Qβ[A40P].

Table 5.1	
MS2	S37
Fr	S37
GA	S36
Prr1	G37
Qβ	A40

We next determined the likely geometry that gives rise to VLPs in this size and molecular weight regime. We used surface area and volume calculations, assuming that the icosahedral particles are approximately spherical in shape. The surface area of a single monomer can be calculated as follows:

$4\pi r^2$ = Total S.A.

$4\pi(14\text{ nm})^2$ = Total S.A.

2462 nm^2 = Total S.A.

180 monomers in a T = 3 capsid

(2462 nm^2)/(180 monomers) = 13.7 nm^2 S.A. per monomer

We can then calculate the approximate diameter of a T = 1, T = 4, and T = 7 VLP[19]:

#	monomers	Expected S.A.	Expected diameter
T = 1	60	3,300 nm^2	16 nm
T = 3	180	9,900 nm^2	28 nm
T = 4	240	13,200 nm^2	32 nm
T = 7	420	23,100 nm^2	43 nm

Thus, the diameters of VLPs visualized by TEM and CD-MS likely corresponds to VLPs ranging from a T = 1 to T = 4 geometry, but do not achieve a T = 7 geometry. More study will be required to understand the geometric or biochemical changes that give rise to this mixture of VLP sizes, and how the genetic background of Qβ gives rise to this behavior.

In summary, we showed that making an analogous mutation in evolutionarily-related bacteriophages do not confer the same geometric shift as it does in the MS2 CP. Even the fr bacteriophage, which is 85% identical to the MS2 CP, behaved differently from the MS2 CP, indicating that the genetic background is critical for the stable geometric shift to a T=1 particle. The only coat protein with any effect on size is Qβ, which, surprisingly, has the lowest sequence identity to MS2 among the four VLPs. These results highlight the challenges in predicting how mutations affect quaternary structure, emphasizing that further experimental work is necessary to understand the design rules of VLP assembly.

5.B.ii. SyMAPS identifies new mutants yielding unique quaternary structures

We next sought a better understanding of how mutations in the DE loop affect the quaternary structure of the MS2 CP using SyMAPS (see **Chapter 2**). SyMAPS combines systematic libraries and direct functional selections rather than relying on rational design. Using this technique, we replaced residues 36 and 37 in the MS2 CP with two degenerate NNK codons, which encode for all amino acids and one stop codon, yielding a targeted 400-member library with two mutations in each gene (**Figure 5.3**).

We used size exclusion chromatography to select for well-formed virus-like particles, using methods described and evaluated in **Chapter 2**[20]. To probe changes to the quaternary structure, we separated VLPs that eluted early (fractions 12-13) or late (fractions 21-22) by size exclusion chromatography, which we hypothesized would be enriched for large

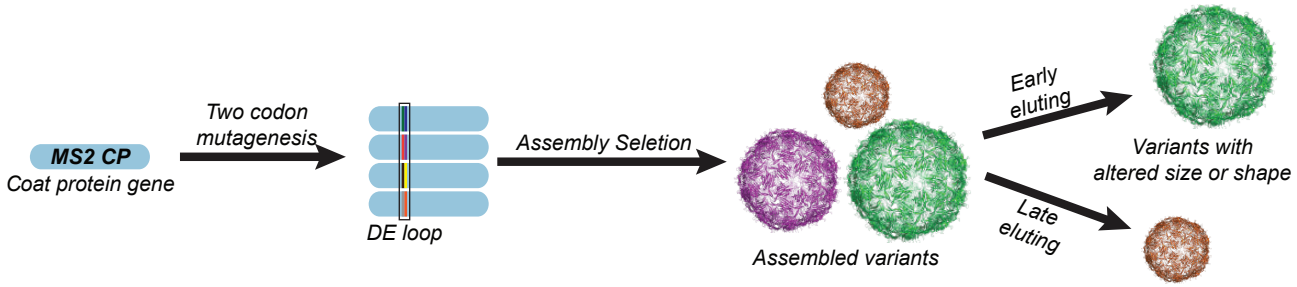


Figure 5.3. SyMAPS strategy to study quaternary structure in the MS2 CP. Two NNK codons were substituted in the DE loop, and variants were separated by size. Early eluting and late eluting fractions were separately analyzed to identify variants with altered size and shape.

and small structures, respectively. These enriched fractions were separately barcoded and analyzed by high-throughput sequencing. For each of the 400 mutants, we calculated the size score, which is the relative abundance in the late-eluting fractions compared to the early-eluting fractions (**Figure 5.4**). We anticipate positive size scores for stable T=1 structures and negative size scores for structure that is larger than wild type VLPs (3.5 MDa). Wild-type, size mixtures, and nonassembling variants are all expected to have scores of zero or N/A.

To validate this technique, we evaluated the size scores of variants for which the quaternary structure is known. Wild-type MS2 and other 27 nm sized particles (CP[N36N / S37C], CP[N36N / S37G]), nonassembling particles (CP[N36N / S37Q]), and mixed sized particles (CP[N36A / S37P]) all have a size score of near zero, as expected, as these variants should not be enriched in early or late fractions (**Figure 5.4A**). Critically, the CP[S37P] variant was enriched in the late-eluting fraction (**Figure 5.4A,B**), confirming that the small size phenotype results in positive size scores. As such, we can conclude that the size score does not differentiate among wild-type, nonassembling, and mixed size phenotypes; however, we do expect that a stable geometric shift would be detectable.

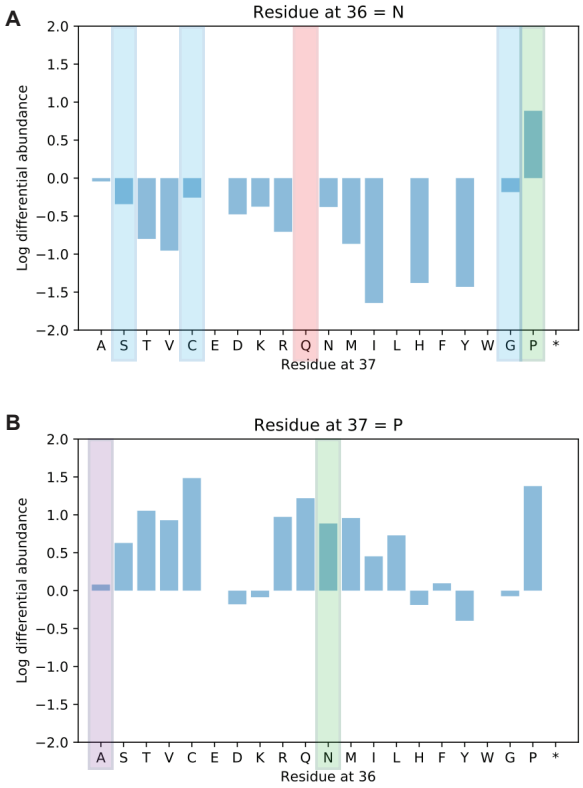


Figure 5.4. Effect of amino acid mutations on the size or shape of the MS2 CP. Log differential abuandance of late-eluting (small) variants compared to early-eluting (large) variants in which (A) N36 is wild-type, or (B) S37 is mutated to proline. Large scores indicate variants that are enriched in the late-eluting fractions from the size selection, while low scores indicate variants that are enriched in early-eluting fractions. Green highlights variants that are known to be 17 nm in diameter (CP[N36N / S37P]); blue variants are 27 nm in diameter (CP[N36N / S37S]; CP[N36N / S37C]; CP[N36N / S37G]); orange do not assemble into VLPs (CP[N36N / S37Q]); and purple variants assemble into a mixture of sizes (CP[N36A / S37P]).

We evaluated which variants bearing the S37P mutation maintained the smaller size phenotype. Variants with charged mutations at position 36 appeared to ablate the 17 nm size phenotype: CP[N36E / S37P], CP[N36D / S37P] and CP[N36K / S37P] all have scores of N/A or close to zero. This indicates that the VLPs are likely size mixtures, wild-type sized, or nonassembling. CP[N36R / S37P] is a surprising exception, perhaps indicating that side chain hydrogen bonding is preserved with this mutation.

Previous studies showed that CP[N36A / S37P] produces a mixed size phenotype[13]. Because of this, we hypothesized that the alanine mutation is unable to form a hydrogen bond with position 98, which we proposed was critical to pinning the monomer into the

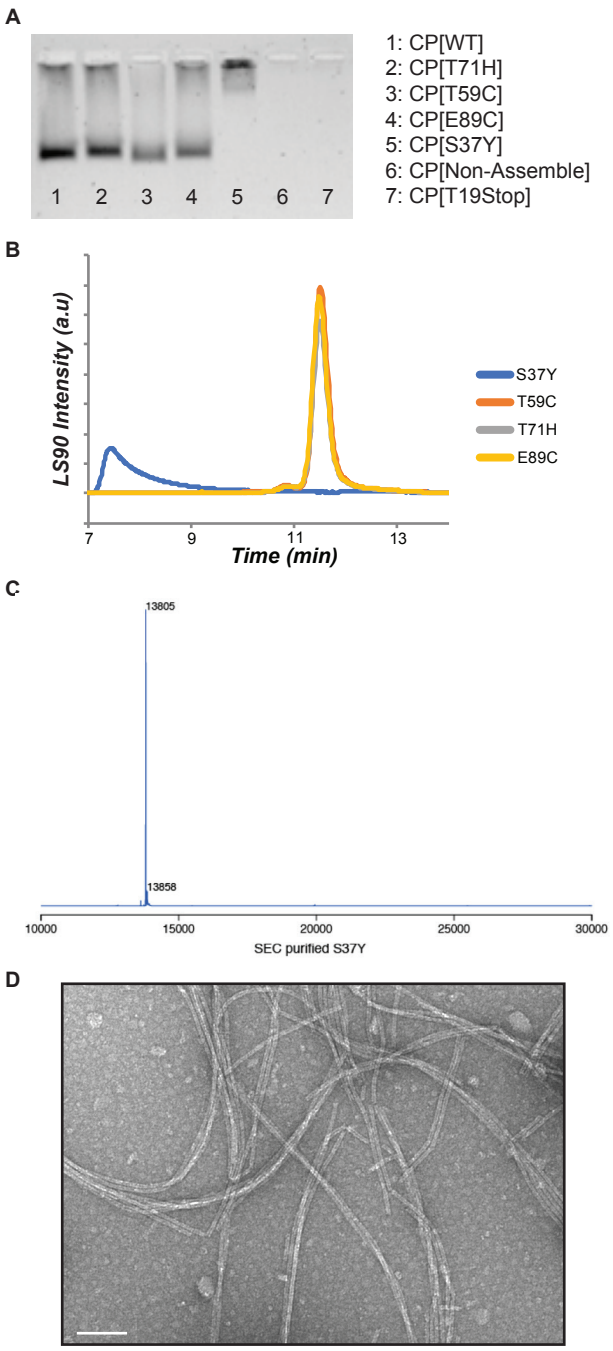


Figure 5.5. Analysis of CP[S37Y] variant. A) Native gel shows CP[S37Y] moves less than wild-type sized variants. B) HPLC SEC confirms a large-eluting structure with (C) the molecular weight of CP[S37Y] when collected and evaluated by mass spectrometry. D) The CP[S37Y] variant makes stable fibrils. Scale indicates 100 nm.

5-fold conformation. However, in this study, mutation at position 36 to valine, isoleucine, and leucine all seem to preserve the 17 nm size, potentially contradicting the hypothesis that side chain hydrogen bonding is required to pin the five-fold axis into place. Large side chains, such as tyrosine and phenylalanine, appear to break the particle or restore the wild-type (or mixed) phenotype. CP[N36P / S37P] appears to maintain a small size phenotype, even though Pro-Pro dipeptides can be conformationally restricted[21].

Perhaps more surprisingly, a number of mutations at position 37 result in an apparent shift to particles that are even larger than wild-type MS2 when position 36 is wild type (asparagine) (**Figure 5.4A**). Single amino acid mutations at position 37 to hydrophobic amino acids such as tyrosine, isoleucine, methionine, or even histidine have negative size scores, indicating they were enriched in the late-eluting fractions and are larger than the wild-type-sized VLP structure.

To further evaluate these variants, we used site-directed mutagenesis to generate CP[S37Y]. The variant was expressed in *E. coli* and precipitated via ammonium sulfate. By native gel, the particles appeared to be different from the wild-type sized controls (CP[T59C], CP[T71H], and CP[E89C]) and two nonassembling variants (**Figure 5.5A**). By HPLC SEC, a peak with light scattering properties (indicating large size) eluted in the void volume (**Figure 5.5B**), suggesting the species was larger than MS2 VLPs. That peak was collected and analyzed by mass spectrometry, confirming that the particle was indeed composed of the CP[S37Y] variant (**Figure 5.5C**).

Finally, we visualized the CP[S37Y] variant by TEM. The variant appeared to be forming long, thin, apparently hollow fibrils, often stretching over microns in length (**Figure 5.5D**). This is similar to a phenotype that has been observed elsewhere, when an overabundance of TR-DNA was added to a disassembly–reassembly reaction of MS2 VLPs[1]. We hypothesize that these fibrils arise because hydrophobic mutations at position 37 prevent the monomer from adopting the confirmation needed to fold into the 5-fold axis. With hexamers only, the CP[S37Y] mutation would be unable to make a closed shell. This phenotype has been observed for other self-assembling hexameric structures such as microcompartent shell proteins[22]. This in effect produces the opposite effect of the S37P mutation, which preferentially forms 5-fold axis over the 6-fold axis[18].

We evaluated whether useful properties of the MS2 CP, such as tolerance to disassembly–reassembly[23], acid stability, and heat stability, were preserved in the fibrils (**Supplementary Fig. 5.3**). First, we used TEM to visualize stability to various conditions, as this was the most reliable way to determine if fibrils (rather than aggregates) were present. For these experiments, CP[S37I] was used, another variant that forms fibril structures. Fibrils were visible up to 70 ºC (**Supplementary Fig. 5.3A**) and at pH 3 (**Supplementary Fig. 5.3B**), similar to the wild-type MS2 CP, though some morphological changes are apparent by TEM. In addition, 80 ºC and pH 1 were both tested, but while fibrils were not visible, it is possible that TEM grids may have been damaged by the harsh conditions. Unlike VLPs, at very protein high concentration, fibrils formed a hydrogel-like mesh that could be disrupted with solubilizing agents such as urea. However, we were unable to tune the disassembly–reassembly conditions to stimulate fibril reformation or wild-type particle formation, including when the fibril monomers were combined with wild type or S37P monomers, with and without DNA to simulate reassembly (**Supplementary Fig. 5.3C**).

Together, these results show that fibrils are tolerant to a wide range of conditions, though further work is needed to fully characterize these structures. Fibril proteins—particularly those that form hydrogels—are potential of significant interest in medical applications or as biodegradable scaffolds. As one example, proteinaceous hydrogels can mimic fibrin networks found in extracellular matrices[24]. Thus, sensitive cell culture experiments, such as stem cell differentiation, have great potential use for new and tunable biological scaffolds for improved differentiation control[24].

These studies have led to two additional projects that further probe how mutations affect quaternary structure. First, we are carrying out SyMAPS-based studies on the four related bacteriophages described here, with 400-member libraries in the DE loop. This study seeks to identify mutations that confer stable geometric changes to these related VLPs, evaluating more comprehensively whether any mutations at these positions lead to altered quaternary structure.

In addition, in the course of another project, we found that expression of unrelated fibril-forming proteins resulted in a useful and characteristic defect in cell division, which can be seen by microscopy. When these same techniques were applied to cells expressing CP[S37Y], the cell division defect was also observed (**Figure 5.6A**). Upon further evaluation, we found that flow cytometry can effectively separate fibril-containing *E. coli* from cells expressing nonassembling MS2 variants, indicating that the cell division

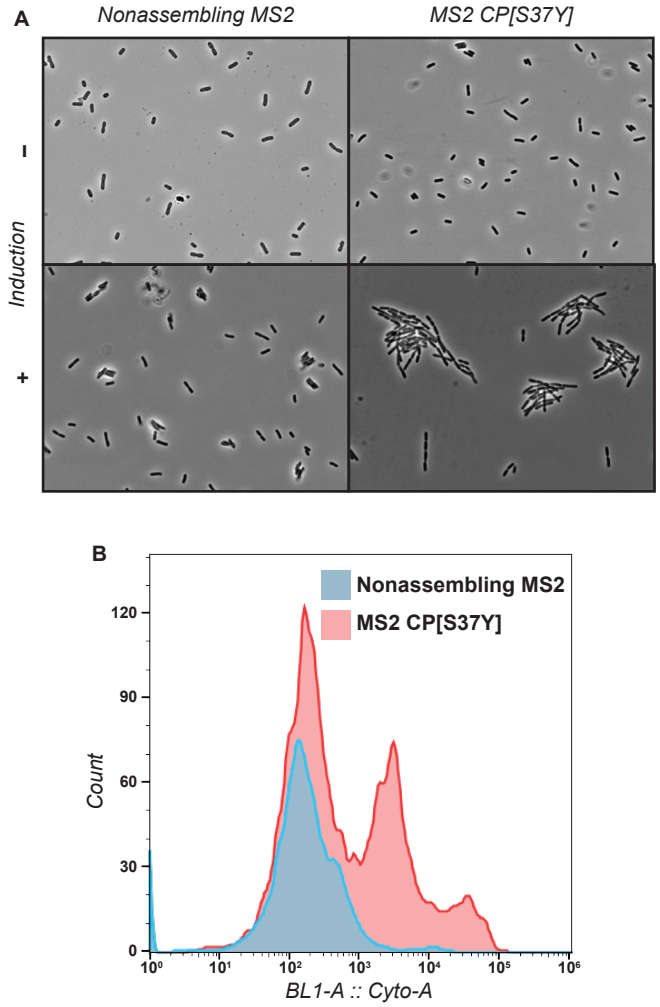


Figure 5.6. Cell division defect caused by CP[S37Y]. A) Induction of CP[S37Y] results in a reproducible cell division defect that can be seen by microscopy. B) These linked cells can be separated from cells expressing a nonassembling MS2 variant by flow cytometry and FACS.

important for controlling quaternary structure in understanding of stable shifts in size and shape.

phenotype is specific to fibril-forming VLP variants (**Figure 5.6B**). This cell division defect allows us to screen many variants simultaneously into fibril-forming and non-fibril-forming populations. Moving forward, we are using SyMAPS to mutagenize the CP[S37Y] gene, then sort the resulting library using flow cytometry. This study will allow us to generate an AFL that is specific to the 6-fold axis of the MS2 CP, allowing useful comparisons to the existing wild-type VLP apparent fitness landscape (AFL) described in **Chapter 2**.

5.C. Conclusions

In this study, we sought to understand how mutations to a self-assembled protein affect quaternary structure. We learned that we were generally unable to predict the behavior of mutations to other bacteriophages, even those that are closely related by sequence and structure to the MS2 CP. We identified a variant of Qβ that exhibits diverse sizes, ranging from 20 to 34 nm, as well as hydrophobic mutations in MS2 VLPs that results in fibril structures. These results support the hypothesis that the DE loop is the MS2 CP and add nuance to our

5.D. Methods

5.D.i. Variant cloning and expression

Single amino acid variants were cloned via site-directed mutagenesis using primers in **Supplementary Data 5.1**[25]. Cloned plasmids were transformed into chemically-competent DH10B cells, and individual colonies were sequenced via Sanger sequencing prior to expression. Selected mutants were individually expressed in 1L cultures of 2xYT. These expressions were pelleted, then purified via FPLC Anion Exchange. Briefly, pellets were lysed by sonication, precipitated with 50% ammonium sulfate, then resuspended in 20 mM taurine buffer, pH 9. Variants were then purified an Akta Pure 25L with a handpacked DEAE Sepharose Fast Flow column (GE Healthcare Life Sciences, Cat# 17070901).

Variants were eluted with 20 mM taurine ph 9.

5.D.ii. Transmission Electron Microscopy (TEM).

Samples were diluted to an A280 of approximately 1, then applied to carbon-coated copper grids for 2 min. Grids were triple rinsed with dd-H_2O. Grids were coated with a 1.6% aqueous solution of uranyl acetate for 1 min to apply a negative stain. Images were acquired at the Berkeley Electron Microscope Lab using a FEI Tecnai 12 TEM with 120 kV accelerating voltage.

5.D.iii. Strains

DH10B chemically competent cells produced in-house were used for all experiments, including variants of interest and libraries. Overnights from a single colony were grown for 16-20 h at 37 °C shaking at 200 RPM in LB-Miller media (Fisher Scientific, Cat# BP1426-2) with chloramphenicol at 32 mg/L. Expressions were subcultured 1:100 into 2xYT media (Teknova, Cat# Y0210) with chloramphenicol and expressed overnight at 37 °C shaking at 200 RPM. Related bacteriophages were identified via Uniprot and compared by ClustalW sequence alignment[26].

5.D.iv. Library generation

To generate libraries with two amino acid mutations in the DE loop, we used a modified library generation strategy based on EMPIRIC cloning[27]. We used Golden Gate compatible entry vectors, the construction of which is described in **Chapter 2**. One of these previously-described entry vector replaced a 26-codon segment flanking the DE loop in the MS2 CP a self-encoding replacement fragemnt[20]. Briefly, a self-encoded removable fragment (SERF) is a genetic sequence with inverted BsaI restriction sites. BsaI digestion can simultaneously removes both SERF and BsaI sites, resulting in scarless cloning. A single stranded DNA primer with two degenerate codons at position 36 and 37 was purchased, and overlap extension PCR was used to generate double stranded DNA with a corresponding reverse primer. The gene fragment was then purified using a PCR Clean-up Kit (Promega, Cat# A9282), diluted to 1-5 ng/µL, then cloned into the entry vector using Golden gate cloning techniques[28]. The ligated plasmids were transformed into chemically competent DH10B cells, allowed to recover with SOC for one hour, then plated on large LB-A plates. Colony count varied but was at least 3x the size of the library. This protocol was repeated independently from library generation for two total biological replicates.

5.D.v. Size selection

Colonies containing library members were scraped from plates into LB-M and were grown for 2 h. Each library was subcultured such that 30 OD units were added to the 1L flask, then allowed to grow to an OD600 of 0.6. Libraries were then induced by 0.1% arabinose, expressed overnight at 37 ºC, then harvested. Harvested libraries were resuspended in 10 mM phosphate buffer at pH 7.2 with 2 mM sodium azide , lysed by sonication, clarified, then precipitated twice with 50% ammonium sulfate. FPLC size exclusion chromatography was used to select for well-formed VLPs, as well as VLPs with

altered geometries.

5.D.vi. FPLC SEC (Assembly selection)

Libraries were purified on an Akta Pure 25 L Fast Protein Liquid Chromatography (FPLC) system via Size Exclusion Chromatography. The column used was a HiPrep Sephacryl S-500 HR (GE Healthcare Life Sciences, Cat# 28935607) column, and isocratic flow with 10 mM phosphate buffer at pH 7.2, 200 mM sodium chloride, and 2 mM sodium azide was used for elution. Fractions containing MS2 coat protein were collected for further analysis (fractions 12-22). To identify variants with altered geometry, fractions 12 and 13 were pooled and separately analyzed, as were fractions 21 and 22.

5.D.vii. HPLC SEC

Variants were evaluated on an Agilent 1290 Infinity HPLC with an Agilent Bio SEC-5 column (5 µm, 2000 Å, 7.8x300 mm). Isocratic flow was used with 10 mM phosphate buffer, pH 7.2, 200 mM sodium chloride, and 2 mM sodium azide. Fractions were collected where a large light scattering peak indicated particles such as fibrils or VLPs.

5.D.viii. Sample prep for high-throughput sequencing

Plasmid DNA was purified prior to library expressions using Zyppy Plasmid Miniprep Kit (Zymo, Cat# D4036). RNA was extracted from MS2 CP library after the assembly selections using previously-published protocols[29]. Briefly, we used TRIzol (Thermo Fisher Cat# 15596026) to homogenize the samples. Chloroform was added, and the samples were separated by centrifugation into aqueous, interphase, and organic layers. The aqueous layer, containing RNA, was isolated, and RNA was precipitated with isopropanol and washed with 70% ethanol. RNA was dried and resuspended in RNAse free water. cDNA was synthesized using the Superscript III first strand cDNA synthesis kit from Life (cat: 18080051, polyT primer). To prepare libraries for high-throughput sequencing, cDNA and plasmids were both amplified with two rounds of PCR (10 cycles and 8 cycles, respectively) to add barcodes and Illumina sequencing handles, following Illumina 16S Metagenomic Sequencing Library Preparation recommendations (Dataset S1). Libraries were combined and analyzed by 300 PE MiSeq in collaboration with the UC Davis Sequencing Facilities. 20.6 million reads passed filter, and an overall Q30 > 79%.

5.D.ix. High-throughput sequencing data analysis

Data were trimmed as previously described[20] with minor variation. Briefly, Trimmomatic[30] was used to trim data with a 4-unit sliding quality window of 20 and a minimum length of 30. FLASH (Fast Length Adjustment of SHort reads)[31] was used to merge reads with a maximum overlap of 160 base pairs. Alignment to the wild-type MS2 CP reference gene was performed with Burrows-Wheeler Aligner (BWA-MEM)[32], then reads were sorted and indexed with Samtools[33]. The CleanSam function from Picard filtered unmapped reads, and any reads that were longer or shorter than the expected length of the barcoded DNA were removed.

5.D.x. Size Score calculations

Cleaned and filtered reads were analyzed using Python programs written in-house. Briefly, the mutated region of the MS2 CP was isolated, and combinations of mutations position were tallied. Codons were then translated into amino acids, removing codons that do not end in G or T. These calculations were repeated for both experiments to calculate abundances before and after the assembly selection. Relative percent abundances were calculated as previously described[20]. Briefly, the grand sum was found for each matrix, and every position was divided by that sum, generating a matrix of percent abundances. Size scores were generated by comparing the percent abundance in the small fractions compared to the large fractions for each replicate. We then calculated the mean across the two replicates. We calculated the log10 of the Relative Percent Abundance array to calculate the final size score for each replicate.

5.D.xi. Native gel

Variants were analyzed by native gel in a 0.8% agarose gel with 0.5X TBE buffer (45 mM Tris-borate, 1 mM EDTA) with 2X SYBR Safe DNA Gel Stain (ThermoFisher Scientific, Cat# S33102). Gels were run for 120 min at 40 V, then imaged on a BioRad GelDoc EZ Imager.

5.D.xii. Mass spectrometry.

Proteins were analyzed with an Agilent 1200 series liquid chromatograph (Agilent Technologies, USA) connected in-line with an Agilent 6224 Time-of Flight (TOF) LC/MS system with a Turbospray ion source.

5.D.xiii. Fibril expression and purification

CP[S37Y] or CP[S37I] were both expressed individually in 50 mL of 2xYT as described above. A single round of ammonium sulfate precipitation enriched for fibrils, allowing for analysis of thermal and acid stability.

5.D.xiv. Fibril analysis (heat, acid, and disassembly–reassembly)

Fibrils were incubated at the indicated temperature for 10 min, then directly applied to a TEM grid for analysis. Acid challenges proceeded in a similar manner, in which fibrils were incubated with buffer of various pH, then directly evaluated by TEM. Disassembly–reassembly experiments were conducted as described previously[23]. CP[S37Y], CP[S37P], and CP[WT] were added to acetic acid to a final concentration of 66% acetic acid, then were allowed to sit on ice for 30 min. The aggregates were pelleted in a microcentrifuge at top speed, then desalted into 100 mM acetic acid using a Nap5 desalting column (GE Healthcare, Cat# 17-0853-01). Disassembled variants were combined in different ratios with and without TR-DNA to stimulate reassembly, then evaluated by HPLC SEC for reassembly success.

5.E. References

1. Galaway, F. A. & Stockley, P. G. MS2 viruslike particles: A robust, semisynthetic targeted drug delivery platform. *Mol. Pharm.* **10**, 59–68 (2013).
2. Ashley, C. E., Carnes, E. C., Phillips, G. K., Durfee, P. N., Buley, M. D., Lino, C. A., Padilla, D. P.,

Phillips, B., Carter, M. B., Willman, C. L., Brinker, C. J., Caldeira, J. D. C., Chackerian, B., Wharton, W. & Peabody, D. S. Cell-specific delivery of diverse cargos by bacteriophage MS2 virus-like particles. *ACS Nano* **5**, 5729–5745 (2011).

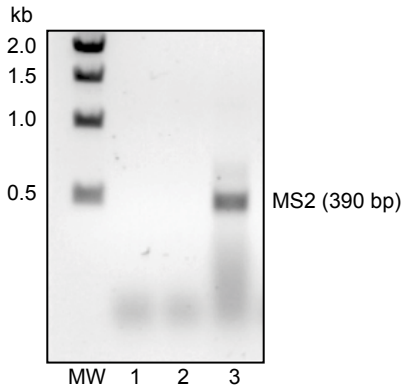
3. ElSohly, A. M., Netirojjanakul, C., Aanei, I. L., Jager, A., Bendall, S. C., Farkas, M. E., Nolan, G. P. & Francis, M. B. Synthetically Modified Viral Capsids as Versatile Carriers for Use in Antibody-Based Cell Targeting. *Bioconjug. Chem*. **26**, 1590–1596 (2015).
4. Aanei, I. L., Elsohly, A. M., Farkas, M. E., Netirojjanakul, C., Regan, M., Taylor Murphy, S., O'Neil, J. P., Seo, Y. & Francis, M. B. Biodistribution of antibody-MS2 viral capsid conjugates in breast cancer models. *Mol. Pharm.* **13**, 3764–3772 (2016).
5. Farkas, M. E., Aanei, I. L., Behrens, C. R., Tong, G. J., Murphy, S. T., Neil, J. P. O. & Francis, M. B. PET Imaging and Biodistribution of Chemically Modified Bacteriophage MS2. *Mol. Pharm.* **10**, 69–76 (2013).
6. Aanei, I. L., Huynh, T., Seo, Y. & Francis, M. B. Vascular Cell Adhesion Molecule-Targeted MS2 Viral Capsids for the Detection of Early-Stage Atherosclerotic Plaques. *Bioconjug. Chem.* **29**, 2526–2530 (2018).
7. Zhang, L., Sun, Y., Chang, L., Jia, T., Wang, G., Zhang, R., Zhang, K. & Li, J. A novel method to produce armored double-stranded DNA by encapsulation of MS2 viral capsids. Appl. *Microbiol. Biotechnol.* **99**, 7047–7057 (2015).
8. Stephanopoulos, N., Tong, G. J., Hsiao, S. C. & Francis, M. B. Dual-surface modified virus capsids for targeted delivery of photodynamic agents to cancer cells. *ACS Nano* **4**, 6014–6020 (2010).
9. Kovacs, E. W., Hooker, J. M., Romanini, D. W., Holder, P. G., Berry, K. E. & Francis, M. B. Dual-surface-modified bacteriophage MS2 as an ideal scaffold for a viral capsid-based drug delivery system. *Bioconjug. Chem.* **18**, 1140–1147 (2007).
10. Hooker, J. M., O'Neil, J. P., Romanini, D., Taylor, S. & Francis, M. B. Genome-free viral capsids as carriers for positron emission tomography radiolabels. *Mol. Imaging Biol*. **10**, 182–191 (2008).
11. Wu, W., Hsiao, S. C., Carrico, Z. M. & Francis, M. B. Genome-free viral capsids as multivalent carriers for taxol delivery. *Angew. Chemie - Int. Ed.* **48**, 9493–9497 (2009).
12. Tong, G. J., Hsiao, S. C., Carrico, Z. M. & Francis, M. B. Viral capsid DNA aptamer conjugates as multivalent cell-targeting vehicles. *J. Am. Chem. Soc.* **131**, 11174–11178 (2009).
13. Finbloom, J., Ozawa, T., Aanei, I., Francis, M., Bernard, J., Elledge, S., Nicolaides, T., Berger, M., Han, K. & Klass, S. Evaluation of Three Morphologically Distinct Virus-Like Particles as Nanocarriers for Convection-Enhanced Drug Delivery to Glioblastoma. *Nanomaterials* **8**, 1007 (2018).
14. Smith, B. R., Kempen, P., Bouley, D., Xu, A., Liu, Z., Melosh, N., Dai, H., Sinclair, R. & Gambhir, S. S. Shape matters: Intravital microscopy reveals surprising geometrical dependence for nanoparticles in tumor models of extravasation. *Nano Lett.* **12**, 3369–3377 (2012).
15. Shah, P. N., Lin, T. Y., Aanei, I. L., Klass, S. H., Smith, B. R. & Shaqfeh, E. S. G. Extravasation of Brownian Spheroidal Nanoparticles through Vascular Pores. *Biophys. J.* **115**, 1103–1115 (2018).
16. Guo, Y., Zhao, S., Qiu, H., Wang, T., Zhao, Y., Han, M., Dong, Z. & Wang, X. Shape of Nanoparticles as a Design Parameter to Improve Docetaxel Antitumor Efficacy. *Bioconjug. Chem.* **29**, 1302–1311 (2018).
17. Finbloom, J., Ozawa, T., Aanei, I., Francis, M., Bernard, J., Elledge, S., Nicolaides, T., Berger, M., Han, K. & Klass, S. Evaluation of Three Morphologically Distinct Virus-Like Particles as Nanocarriers for Convection-Enhanced Drug Delivery to Glioblastoma. *Nanomaterials* **8**, 1007 (2018).
18. Asensio, M. A., Morella, N. M., Jakobson, C. M., Hartman, E. C., Glasgow, J. E., Sankaran, B., Zwart, P. H. & Tullman-Ercek, D. A Selection for Assembly Reveals That a Single Amino Acid Mutant of

the Bacteriophage MS2 Coat Protein Forms a Smaller Virus-like Particle. *Nano Lett.* **16**, 5944–5950 (2016).

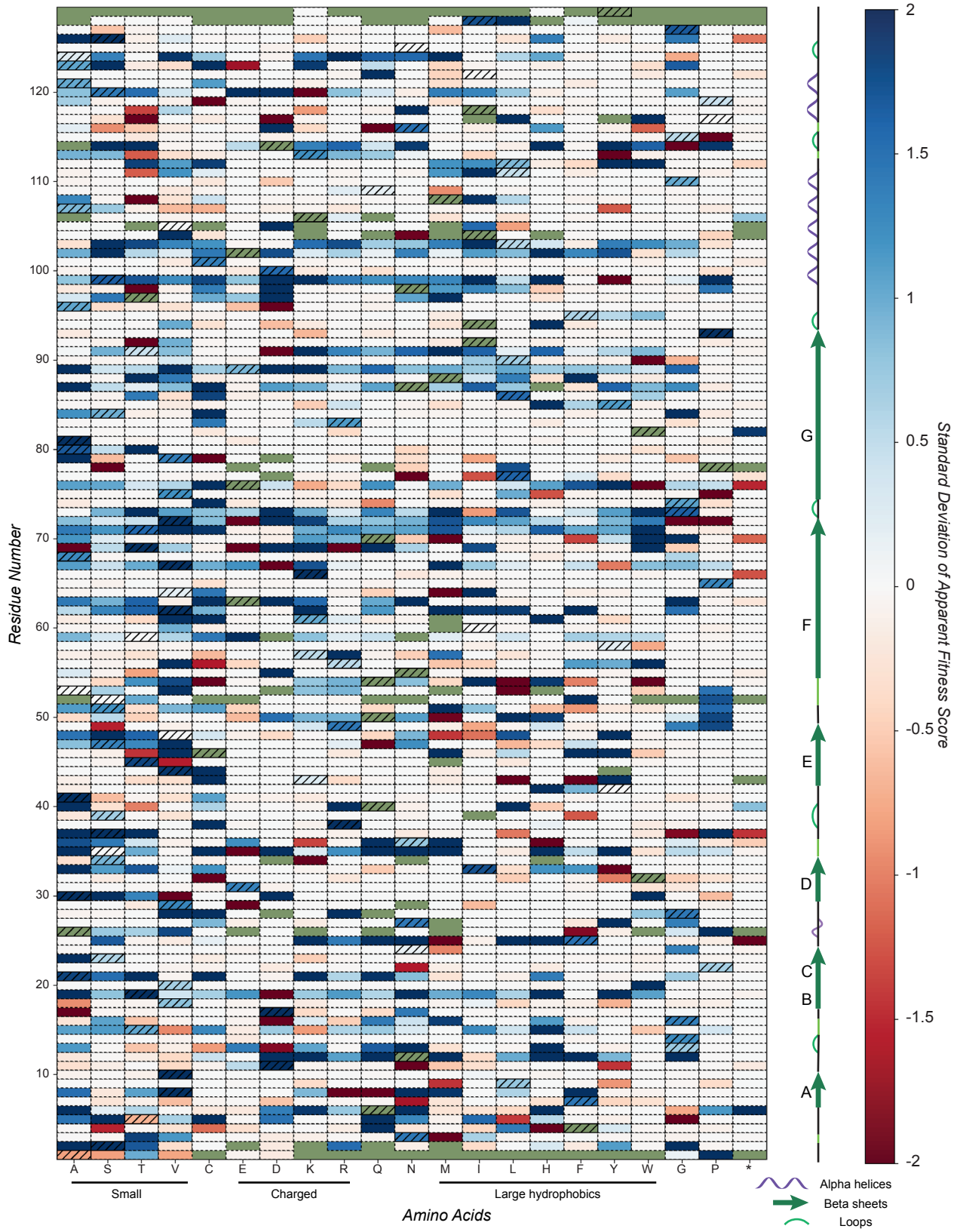
19. Mannige, R. V. & Brooks, C. L. Periodic table of virus capsids: Implications for natural selection and design. *PLoS One* **5**, 1–7 (2010).
20. Hartman, E. C., Jakobson, C. M., Favor, A. H., Lobba, M. J., Álvarez-Benedicto, E., Francis, M. B. & Tullman-Ercek, D. Quantitative characterization of all single amino acid variants of a viral capsid-based drug delivery vehicle. *Nat. Commun.* **9**, 1385 (2018).
21. Williamson, M. P. The structure and function of proline-rich regions in proteins. *Biochem*. *J*. **297**, 249–260 (1994).
22. Pang, A., Frank, S., Brown, I., Warren, M. J. & Pickersgill, R. W. Structural insights into higher order assembly and function of the bacterial micro compartment protein PduA. *J. Biol. Chem.* **289**, 22377–22384 (2014).
23. Glasgow, J. E., Capehart, S. L., Francis, M. B. & Tullman-Ercek, D. Osmolyte-mediated encapsulation of proteins inside MS2 viral capsids. *ACS Nano* **6**, 8658–8664 (2012).
24. Caliari, S. R. & Burdick, J. A. A practical guide to hydrogels for cell culture. *Nature Methods* **13**, 405–414 (2016).
25. Wang, W. & Malcolm, B. A. Two-Stage PCR Protocol Allowing Introduction of Multiple Mutations, Deletions and Insertions Using QuikChangeTM Site-Directed Mutagenesis. *Biotechniques* **26**, 680–682 (1999).
26. Larkin, M. A., Blackshields, G., Brown, N. P., Chenna, R., Mcgettigan, P. A., McWilliam, H., Valentin, F., Wallace, I. M., Wilm, A., Lopez, R., Thompson, J. D., Gibson, T. J. & Higgins, D. G. Clustal W and Clustal X version 2.0. *Bioinformatics* **23**, 2947–2948 (2007).
27. Hietpas, R. T., Jensen, J. D. & Bolon, D. N. A. Experimental illumination of a fitness landscape. *Proc. Natl. Acad. Sci. U. S. A.* **108**, 7896–7901 (2011).
28. Engler, C., Gruetzner, R., Kandzia, R. & Marillonnet, S. Golden gate shuffling: a one-pot DNA shuffling method based on type IIs restriction enzymes. *PLoS One* **4**, e5553 (2009).
29. Pinto, F., Thapper, A., Sontheim, W. & Lindblad, P. Analysis of current and alternative phenol based RNA extraction methodologies for cyanobacteria. *BMC Mol. Biol.* **10**, 79 (2009).
30. Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: A flexible trimmer for Illumina sequence data. *Bioinformatics* **30**, 2114–2120 (2014).
31. Magoč, T. & Salzberg, S. L. FLASH: Fast length adjustment of short reads to improve genome assemblies. *Bioinformatics* **27**, 2957–2963 (2011).
32. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754–1760 (2009).
33. Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G. & Durbin, R. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).

Appendix 1: Supplementary Figures

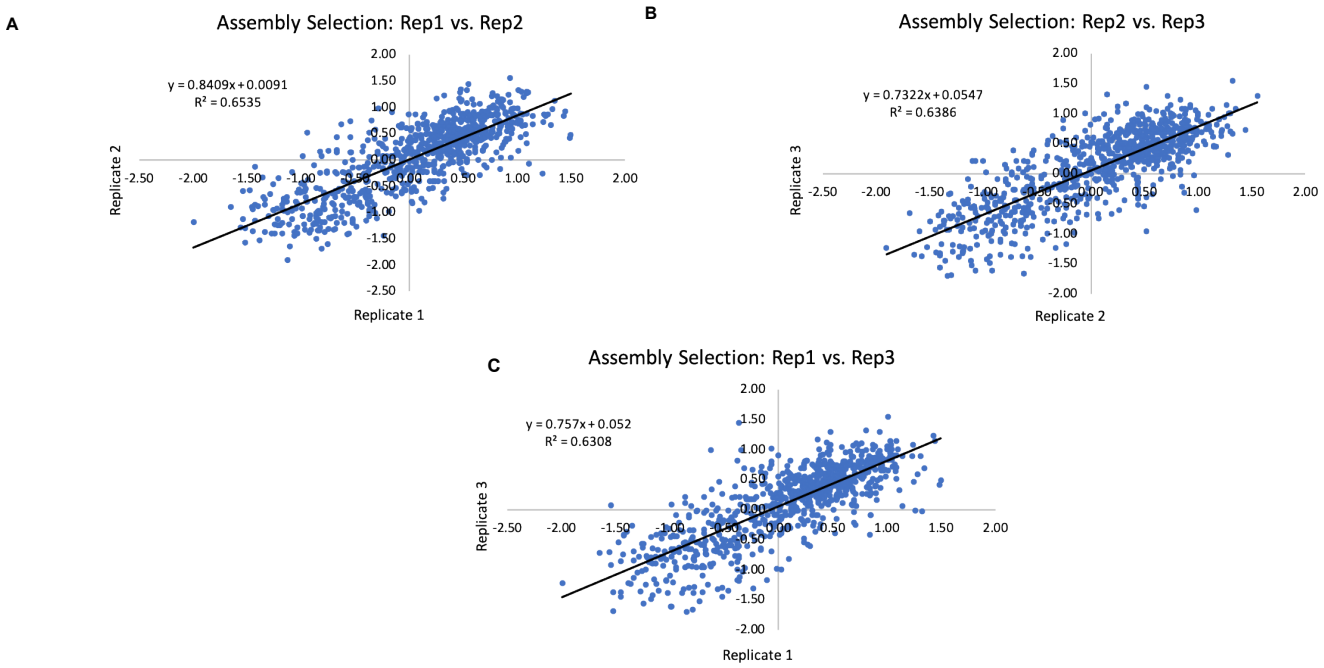
A1.1 Chapter 2 Supplementary Figures



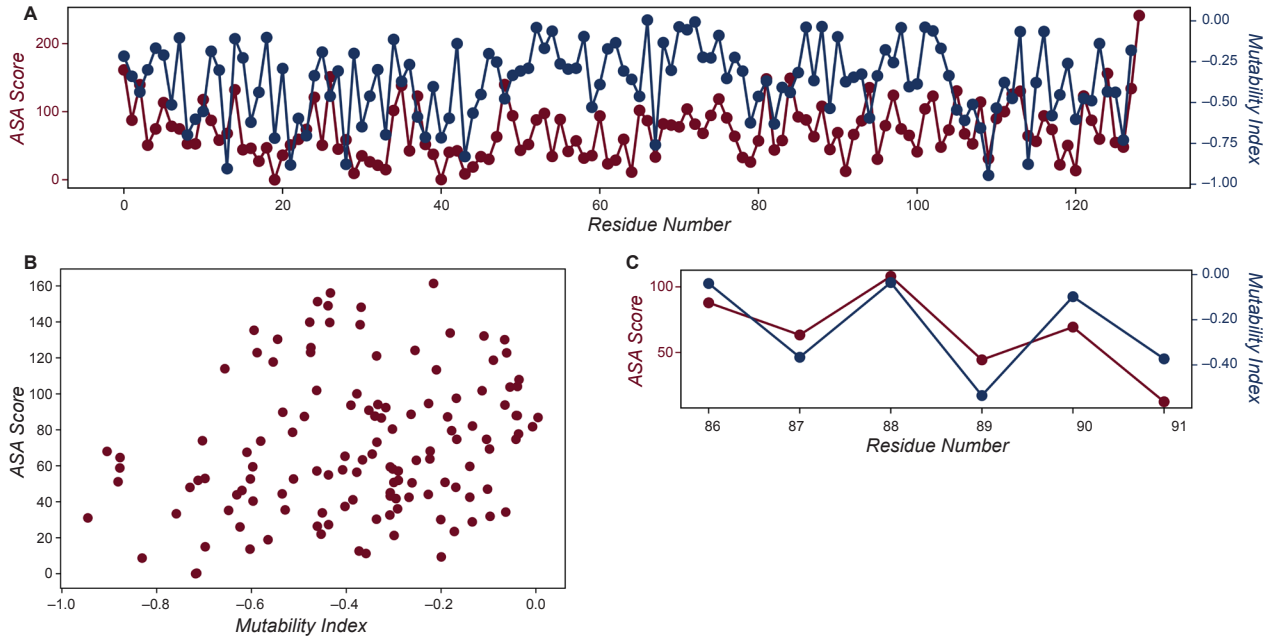
Supplementary Figure 2.1. CP[WT], CP[T19Stop], and CP[Non-assembling] VLPs were tested for RNA after the selection process using PCR. Lane 1: CP[T19Stop]; Lane 2: CP[Non-assembling]; Lane 3: CP[WT].



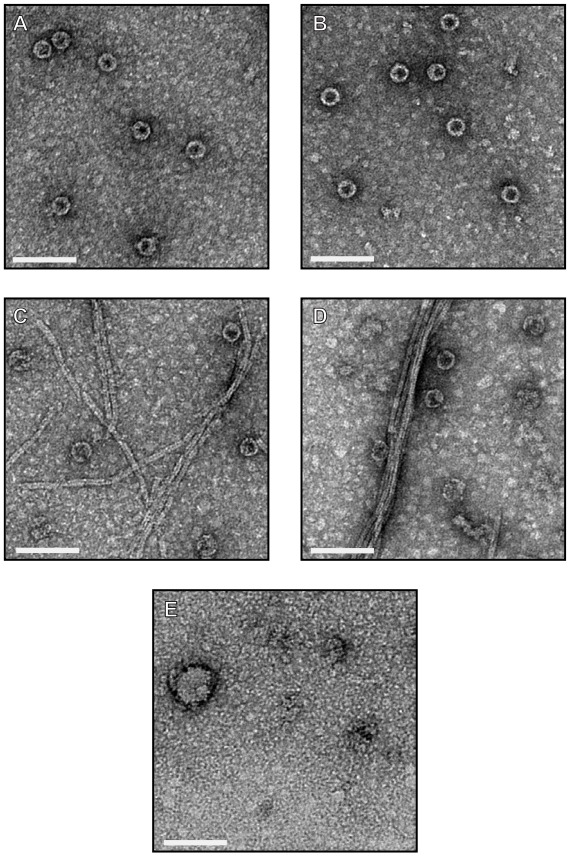
Supplementary Figure 2.2. Standard deviations of Apparent Fitness Scores (n=3). Logarithmic values are reported, where blue values are large standard deviations and red values are small standard deviations. Wild-type residues are indicated with hatches, and missing values are green.



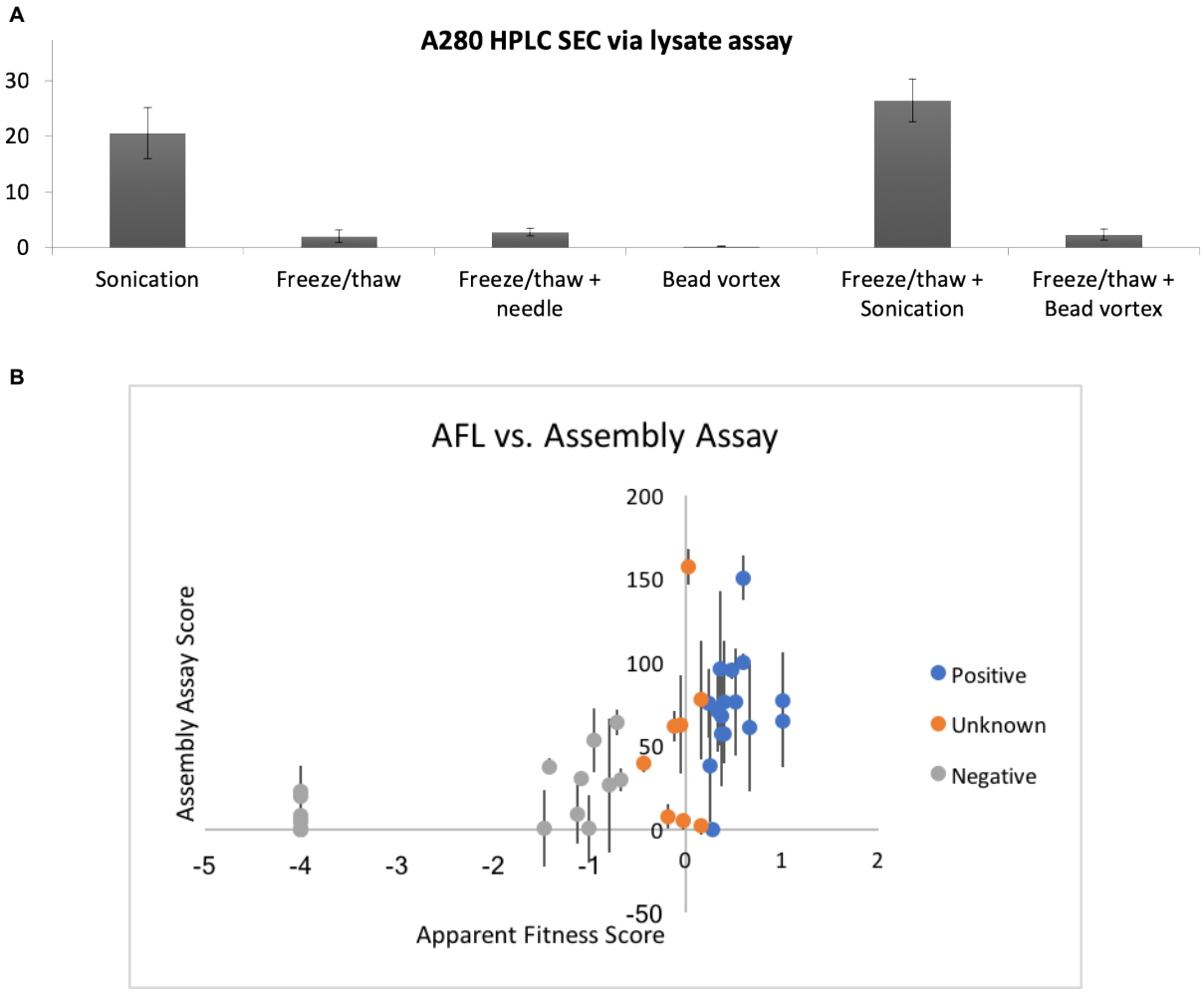
Supplementary Figure 2.3. Correlation analyses for all replicates. All scores with a –4 or nan value in any replicate are exluded. A.) replicate 1 vs. replicate 2 is shown, including the best fit equation and R^2 values. The same information is shown for (B) replicate 2 vs. replicate 3; and (C) replicate 1 vs. replicate 3.



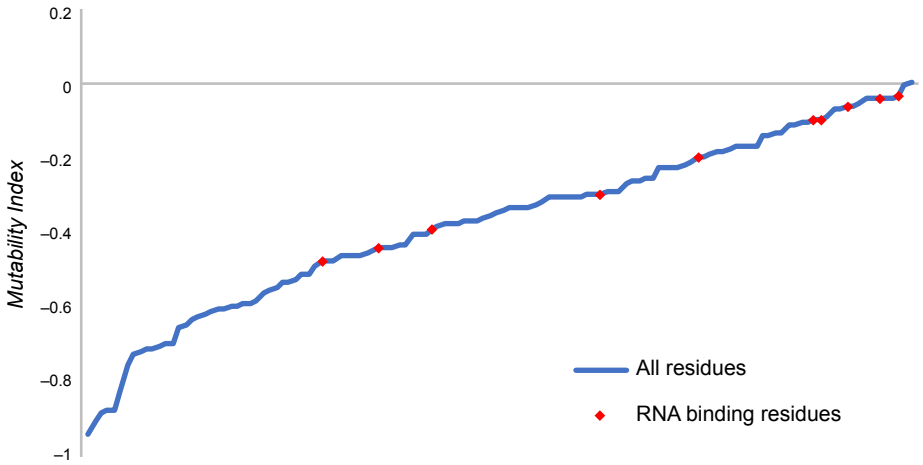
Supplementary Figure 2.4. Accessible surface area (ASA) compared to Mutability Index (MI). The ASA scores and MI values are plotted A) by residue number, and B) as a scatterplot. C) Beta sheet G ASA values are compared to the MI values for each residue.



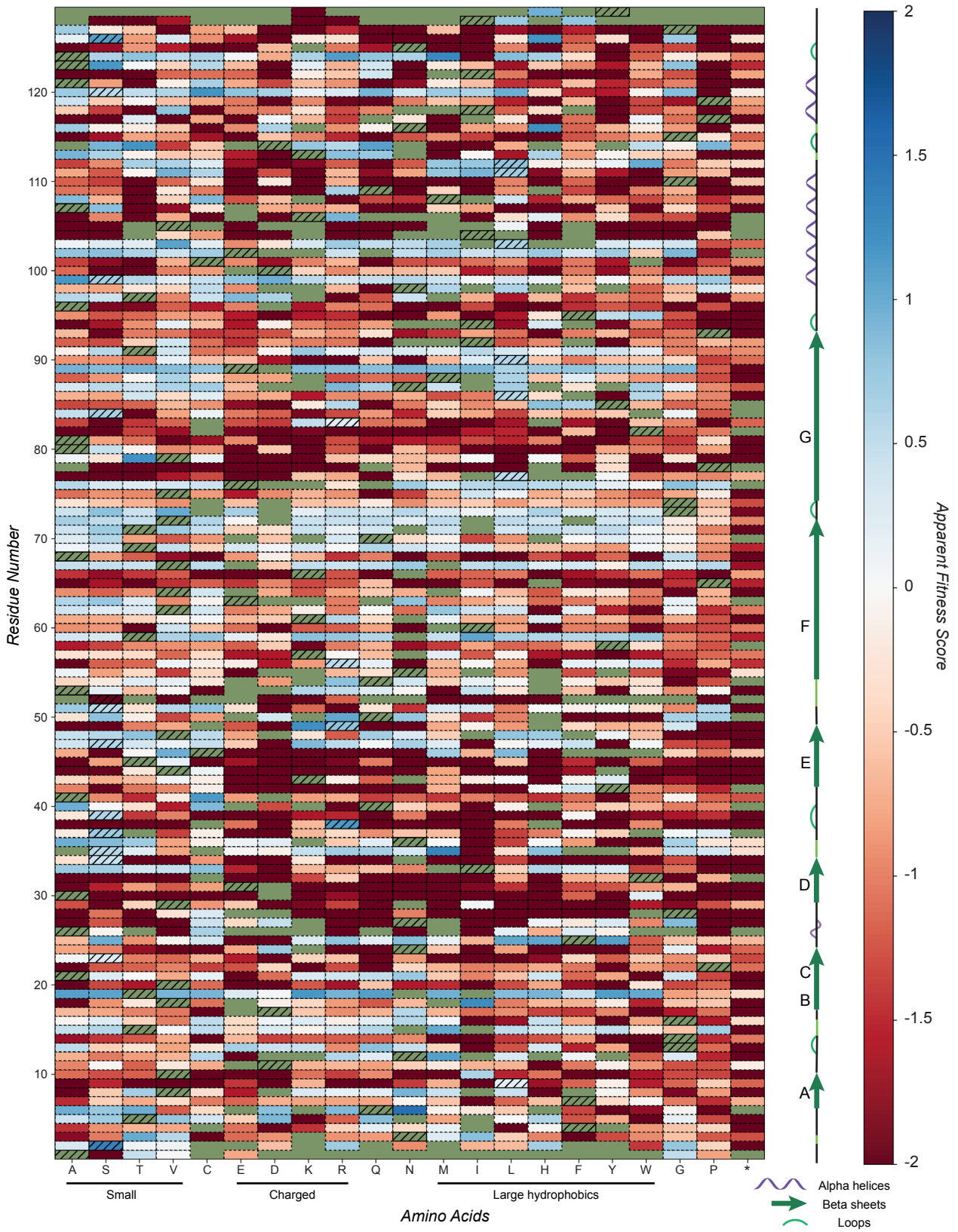
Supplementary Figure 2.5. TEM images of MS2 VLPs. A) CP[WT], B) CP[T71H], C) CP[T91C], D) CP[Q50C], and E) CP[F4V] are imaged using a $UO_2(Ac)_2$ negative stain. Scale bars indicate 100 nm.



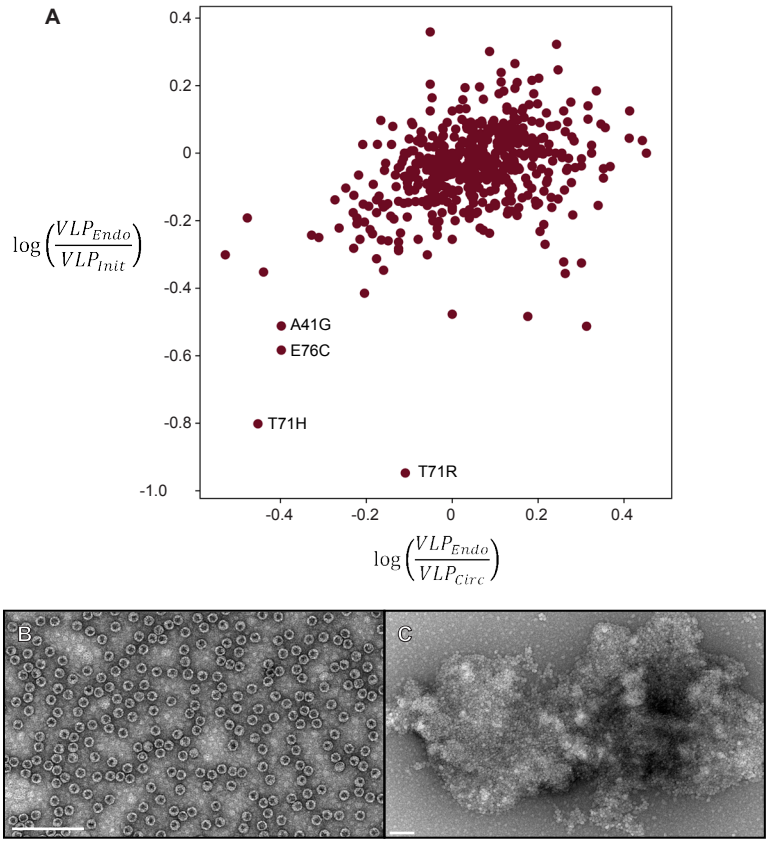
Supplementary Figure 2.6. Assembly screen to validate the Apparent Fitness Landscape. A.) Six different lysis methods are compared to quantify yield of VLP. B.) The optimized assembly assay is used to quantify all 20 variants at position 49 and 91.



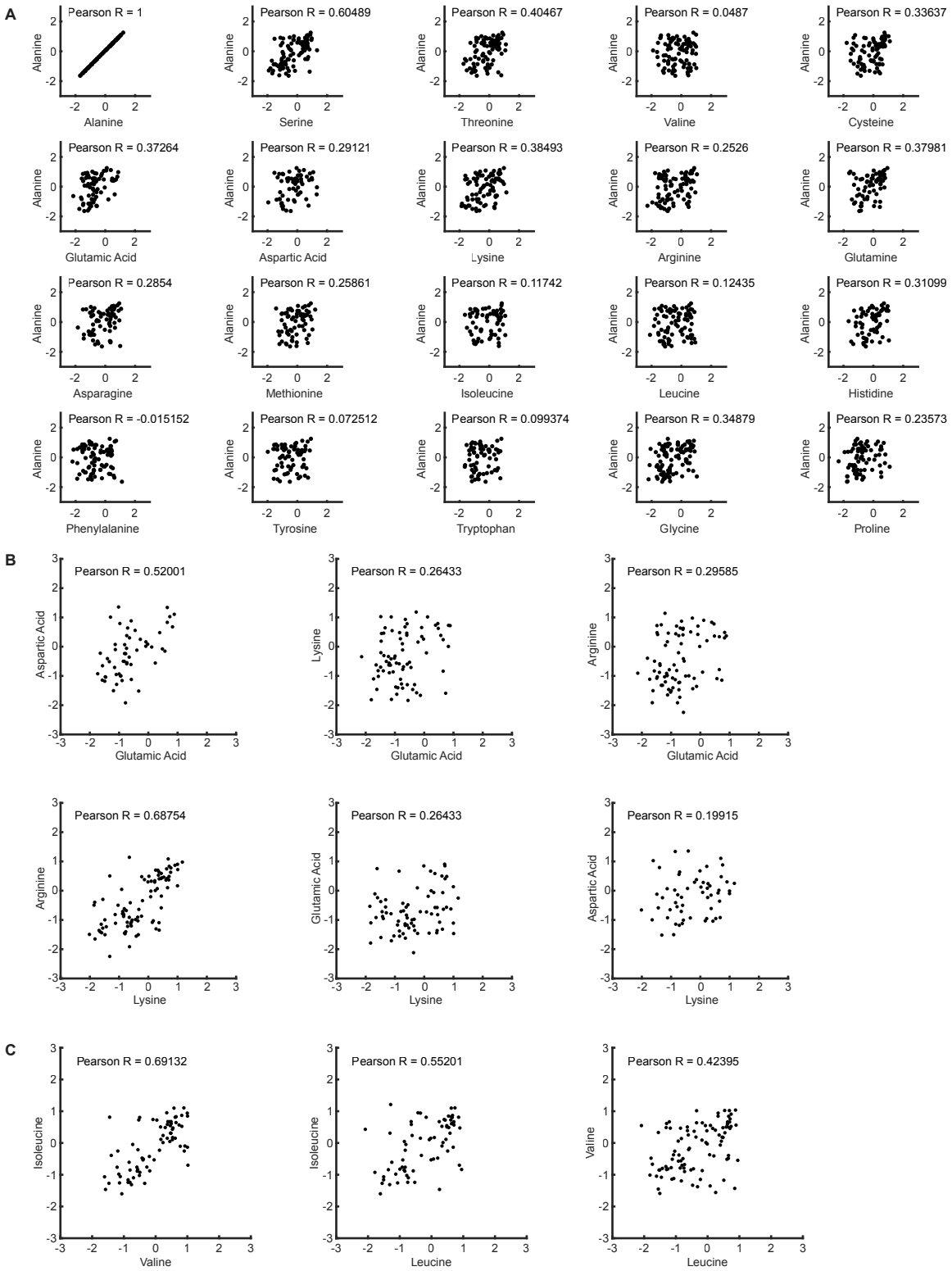
Supplementary Figure 2.7. Residues arranged and plotted by increasing Mutability Index. Mutability Index is the differential Shannon Entropy between the started and selected libraries and is used as a measure of permitted diversity. RNA binding residues are indicated in red.



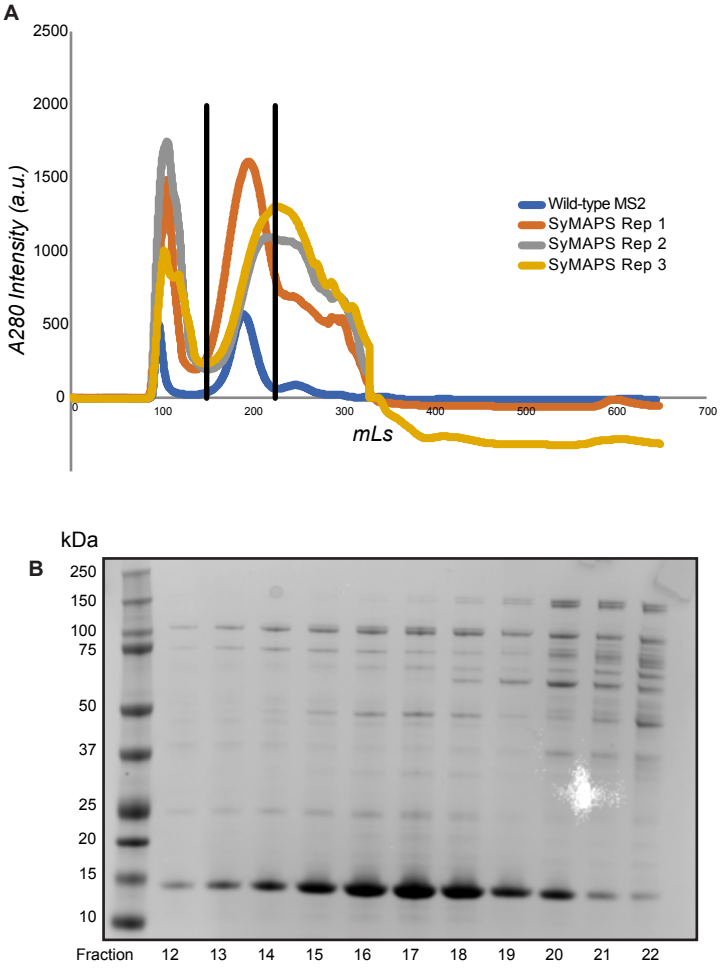
Supplementary Figure 2.8. Apparent Fitness Scores (AFS) where single base pair mutations are eliminated to evaluate the effects of sequencing read errors. Wild-type residues are indicated with hatches, and missing values are green. Dark red variants were sequenced before selection but absent following selection.



Supplementary Figure 2.9. Acid tolerance of MS2 CP variants. A) The population of VLPs remaining after 4 h of incubation at pH 5.0, 37 ºC (VLP_{Endo}) is compared to the population maintained at pH 7.3, 37 ºC (VLP_{Circ}) and the starting VLP library stored at pH 7.3, 4 ºC (VLP_{Init}). Variants of interest are CP[E76C] (–0.40, –0.58) and CP[T71H] (–0.45, –0.80). B) CP[WT] and C) CP[T71H/E76C] are imaged with TEM using a $UO_2(Ac)_2$ negative stain at pH 3.6. Scale bars indicate 200 nm.

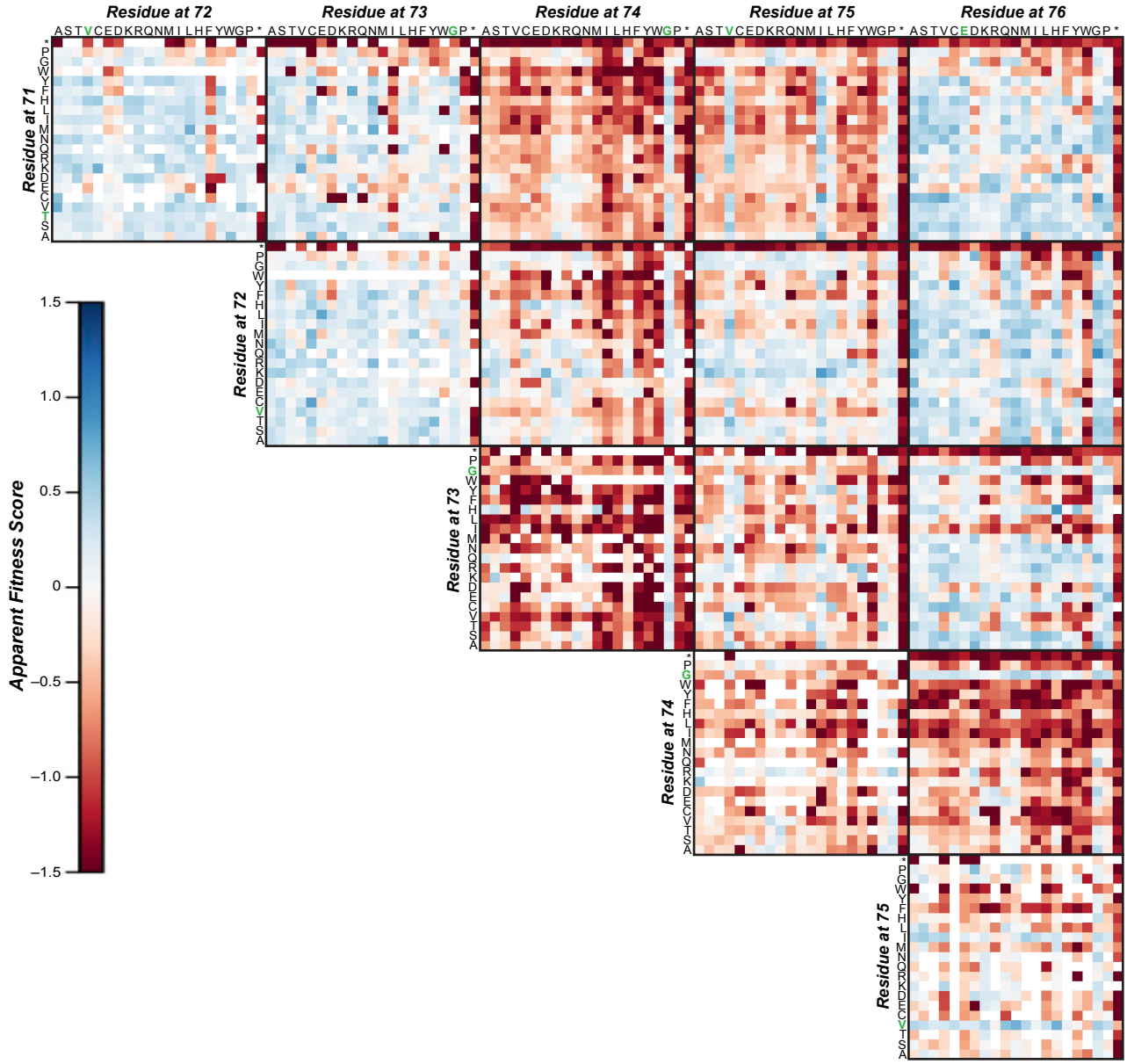


Supplementary Figure 2.10. AFS values correlations between substituted amino acids. AFS correlations between A) Alanine and all other amino acids, B) Charged amino acids, and C) Selected branched, hydrophobic amino acids. Residues where the AFS value of either amino acid was –4 (sequenced before selection but not after) were excluded from this analysis.

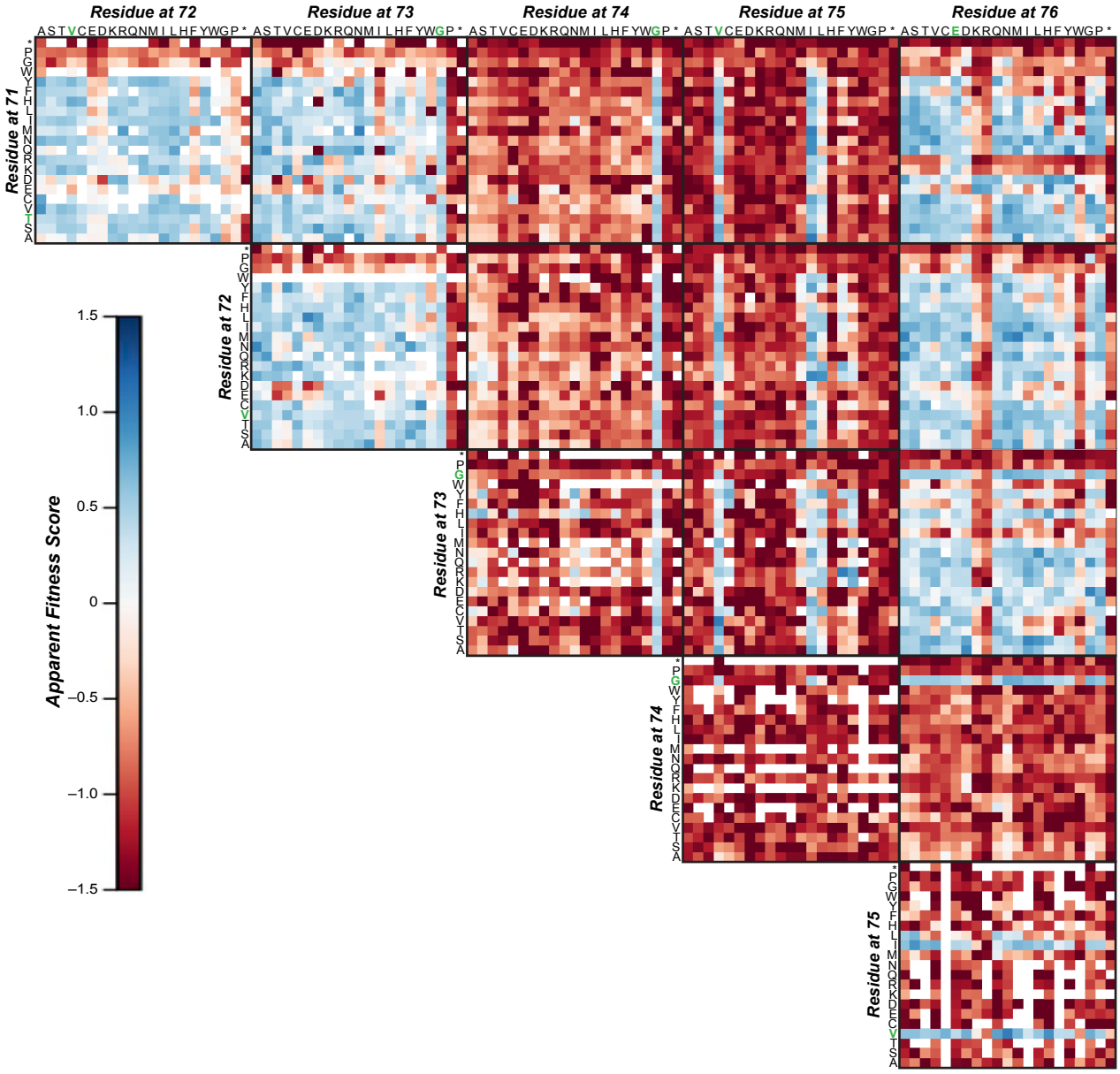


Supplementary Figure 2.11. SyMAPS selection using FPLC SEC. A) Traces of three SyMAPS replicates, where FPLC SEC is used to enrich for well-formed VLPs, in comparison to CP[WT]. Black lines indicate fractions that were collected for high-throughput sequencing analysis. B) SDS-PAGE gel of fractions 12 through 22 collected from CP[WT] size selection, showing enrichment for the MS2 CP band.

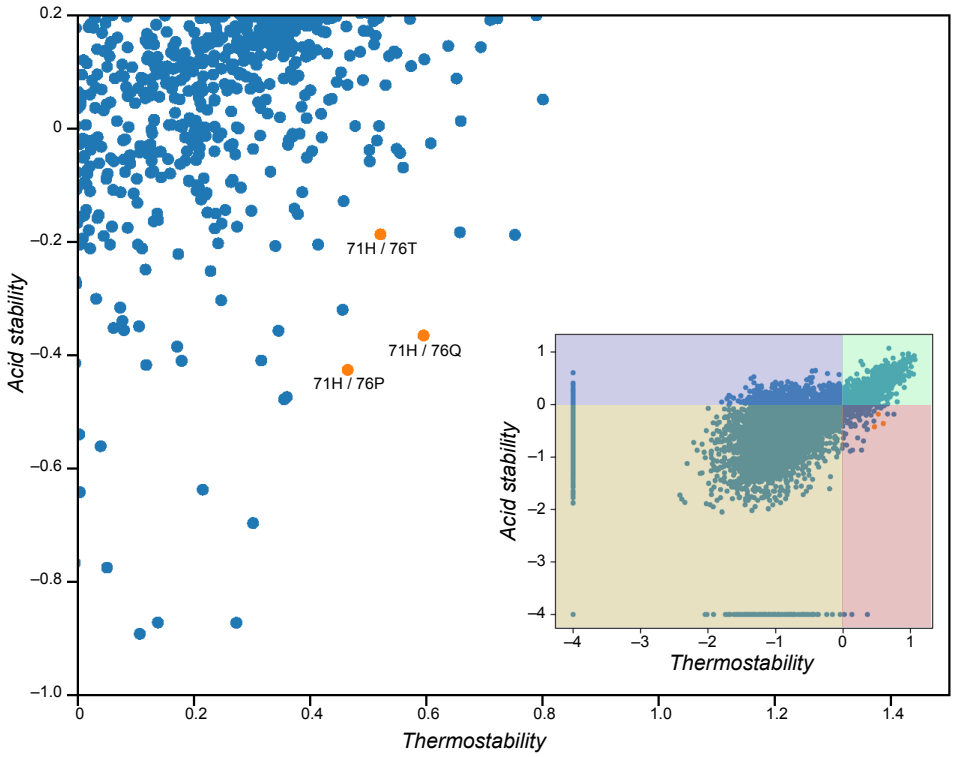
A1.2 Chapter 3 Supplementary Figures



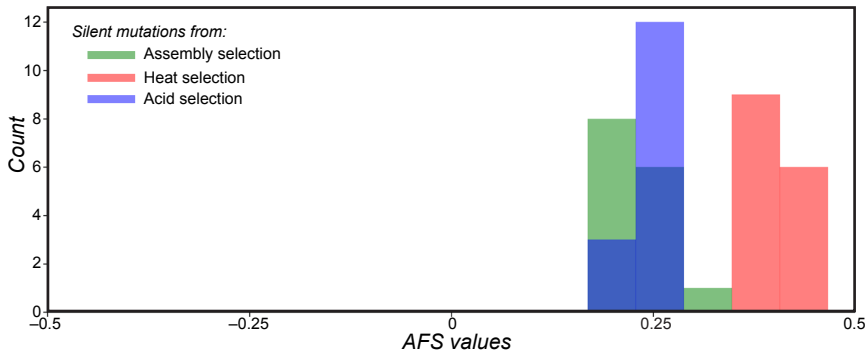
Supplementary Figure 3.1. Assembly-selected epistatic landscape of the MS2 CP FG loop. All possible two amino acid variants were characterized for assembly competency. Blue indicates variants that were enriched following the assembly selection, and red indicates variants that were less abundant following the selection. Dark red indicates variants that were sequenced in the plasmid library but absent in the VLP library.



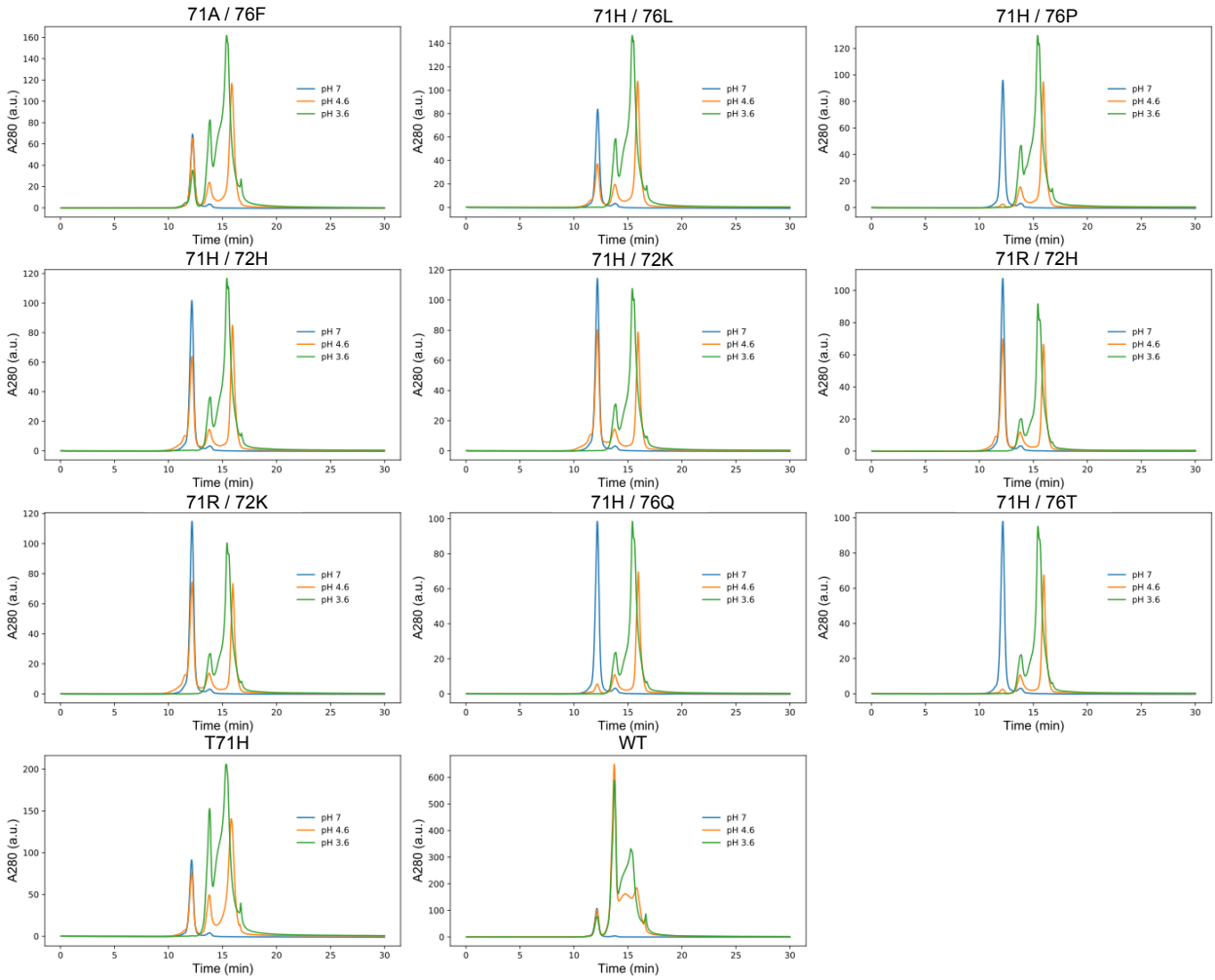
Supplementary Figure 2.2. Epistatic landscape of the FG loop following a heat challenge at 50 ºC for 10 min. Blue indicates variants that were enriched following the heat selection, and red indicates variants that were less abundant following the selection. Dark red indicates variants that were sequenced in the plasmid library but absent in the heat-challenged library. Wild-type amino acids are indicated in green for the one-letter codes.



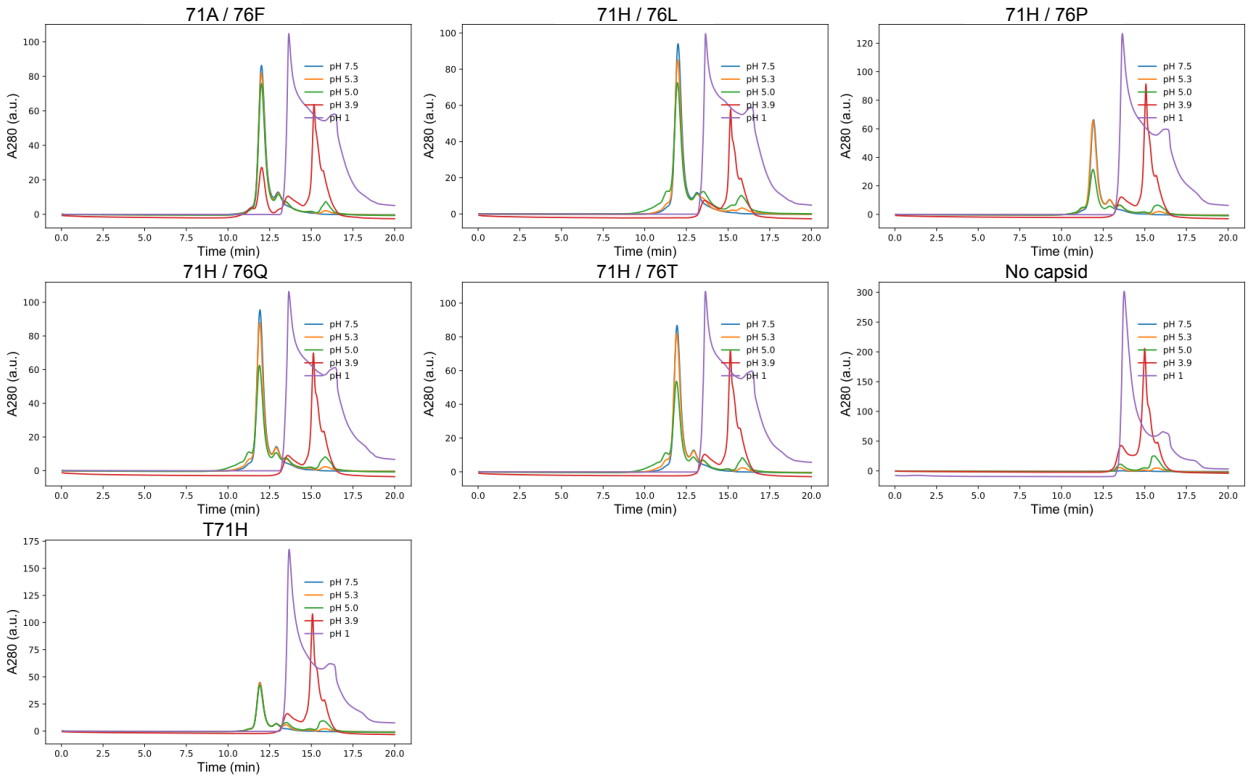
Supplementary Figure 3.3. Thermostable VLPs were compared to acid-stable VLPs to identify candidate variants with lowered acid stability and uncompromised thermostability. Variants of interest have high thermostability and low acid stability scores, which corresponds to the region highlighted in red. High acid stability and high thermostability (green), high acid stability and low thermostability (purple), and low acid stability and low thermostability (beige) are all less desirable variants. Several variants, indicated in orange, had strong positive scores in the heat-selected 2D-AFL with far reduced abundances following acidic pressure.



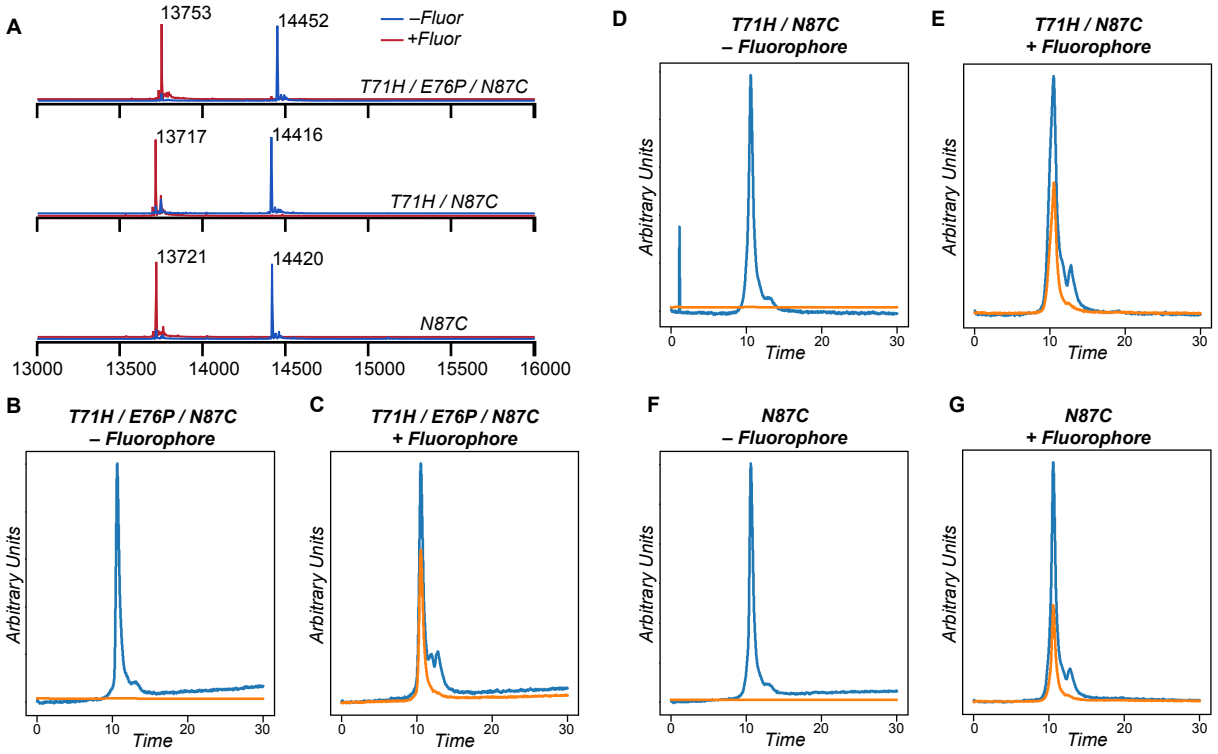
Supplementary Figure 3.4. Silent mutations from the assembly, heat, and acid selections. In each selection, all 15 instances of unmuated VLPs score higher than 0.2. In the heat selection, CP[WT] scores increase, likely because many mutants are not tolerant to the heat selection, resulting in increased relative percent abundances.



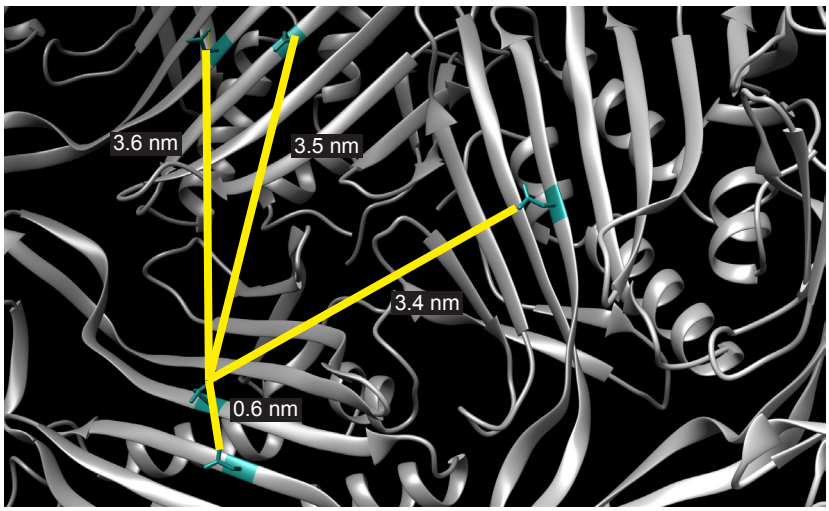
Supplementary Figure 3.5. Variants screened for sensitivity to pH 3.6 and 4.6. HPLC SEC was used to detect sensitivity to pH 4.6 and 3.6. Peak height at 11.2 min was used as a proxy for VLP formation following overnight incubation at the indicated pH.



Supplementary Figure 3.6. Variants screened for sensitivity to pH 5.3, 5.0, 3.9, and 1. HPLC SEC was used to evaluate sensitivity to a wider pH range. Peak height at the characteristic elution time for MS2 capsids, 11.2 min, was used to measure VLP formation after overnight incubation at the indicated pH.

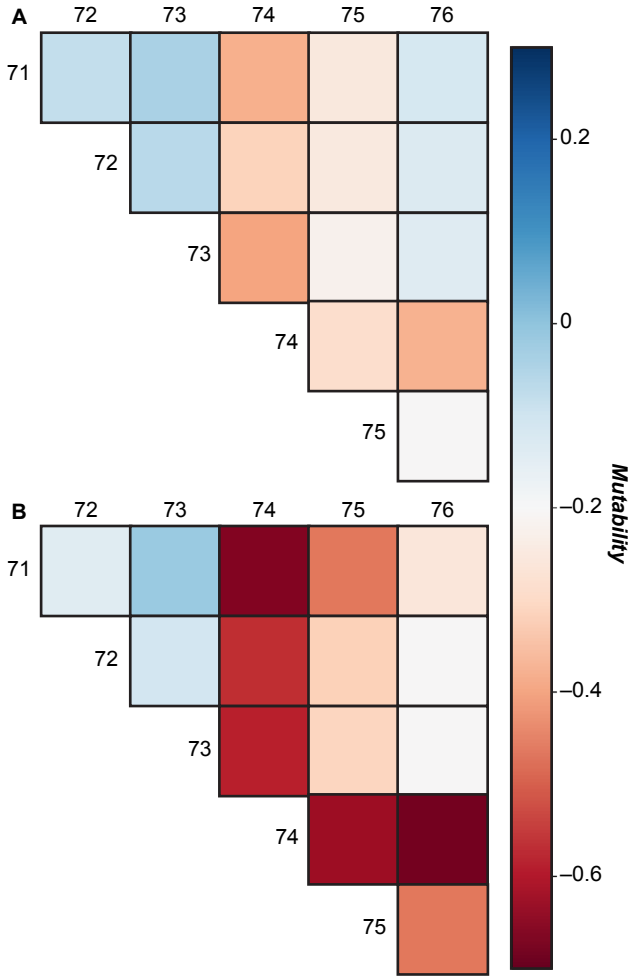


Supplementary Figure 3.7. Chemical conjugation of N87C. A) Mass spectrometry indicates near complete modification of the interior cysteine in all cases. HPLC SEC traces of CP[T71H/E76P/N87C] (B,C), CP[T71H/N87C] (D,E), and CP[N87C] (F,G) with and without AlexaFluor488 maleimide, respective. Orange indicates absorbance at 488 nm, while blue indicates absorbance at 280 nm.

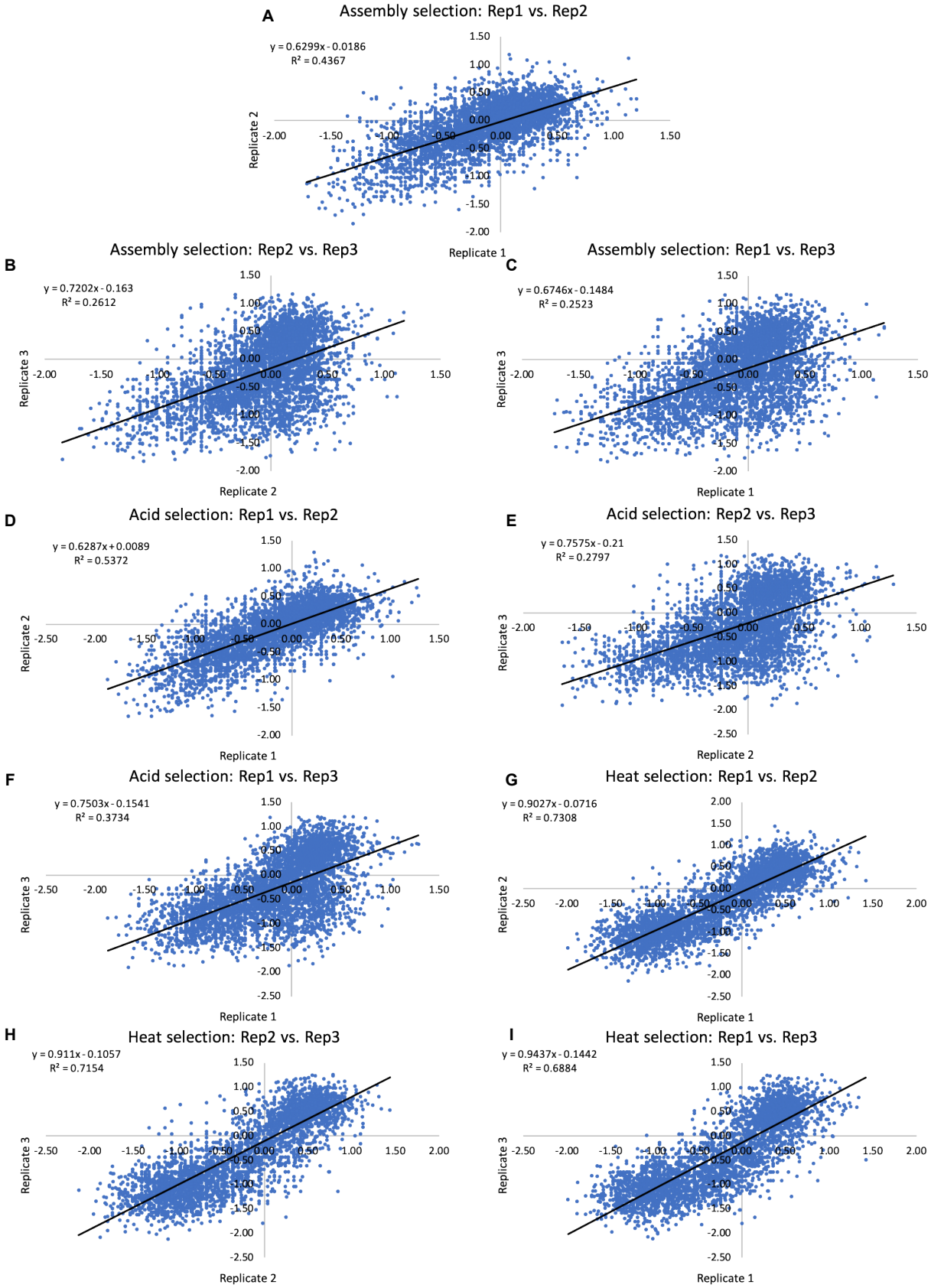


Supplementary Figure 3.8. Distance between cysteines in CP[N87C]. Five cysteines in the interior or the VLP are within the distance required for efficient Förster resonance energy transfer (FRET) quenching.

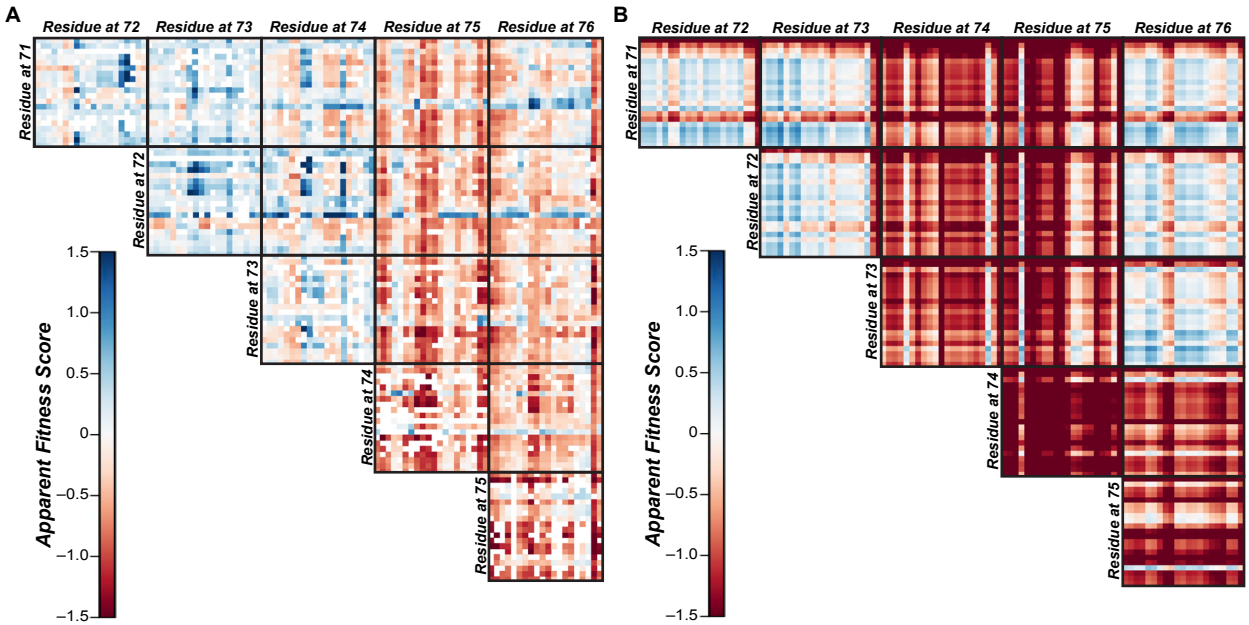
Supplementary Figure 3.9. Emission spectra of quencher and fluorophore labeled variants. Fluorophore (AlexaFluor647) and quencher (Tide 5WS) were combined at indicated concentrations during maleimide modification. Protein concentration was constant at 5 μM, excluding the no protein control. Fluorophore amounts vary in (A), (C), and (E) while quencher amounts vary in (B), (D), and (F).



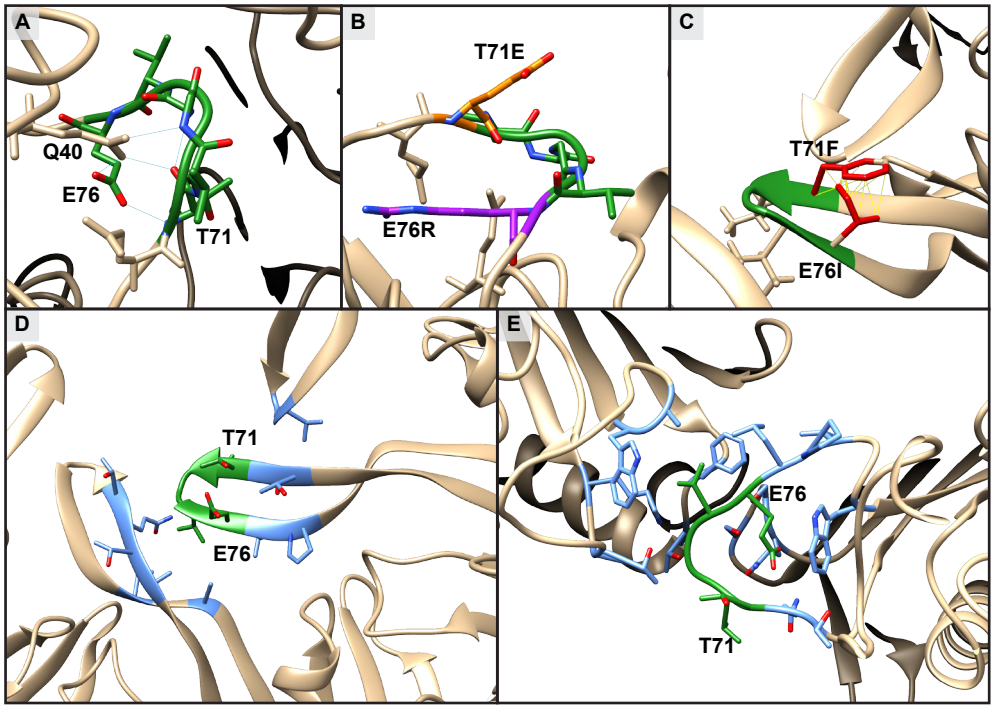
Supplementary Figure 3.10. Shannon Entropy is used to calculate the mutability of each pair of residues following A) assembly selection and B) heat selection.



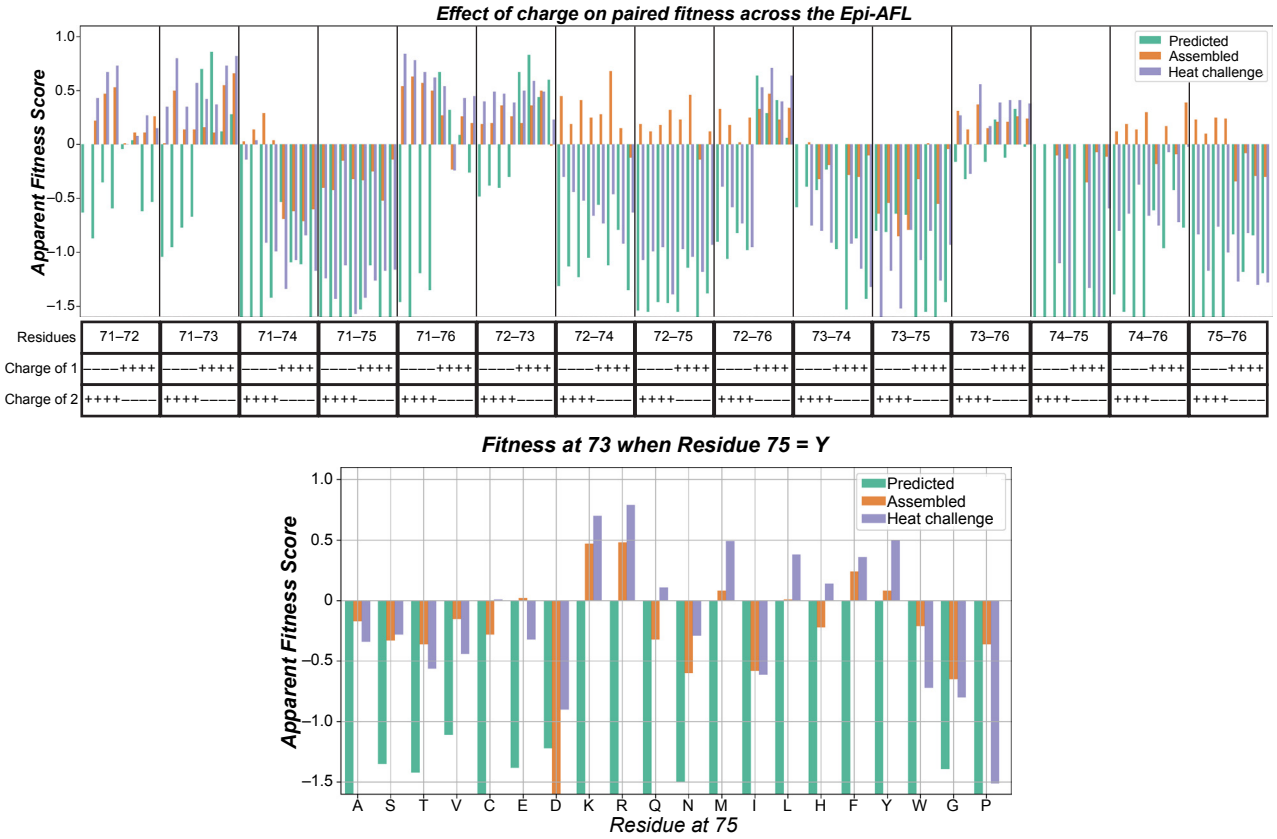
Supplementary Figure 3.11. Correlation analyses from the datasets in this study. Each dataset is compared to the other two biological replicates for the (A-C) assembly, (D-F) acid, and (G-I) heat selections. Best fit lines and R^2 values are indicated on each graph.



Supplementary Figure 3.12. Predicted 2D-AFLs generated from 1D-AFL data using A) a convolutional neural network and B) a simple additive method. The values indicated in B appear as the green bars in Figure 6 of the main

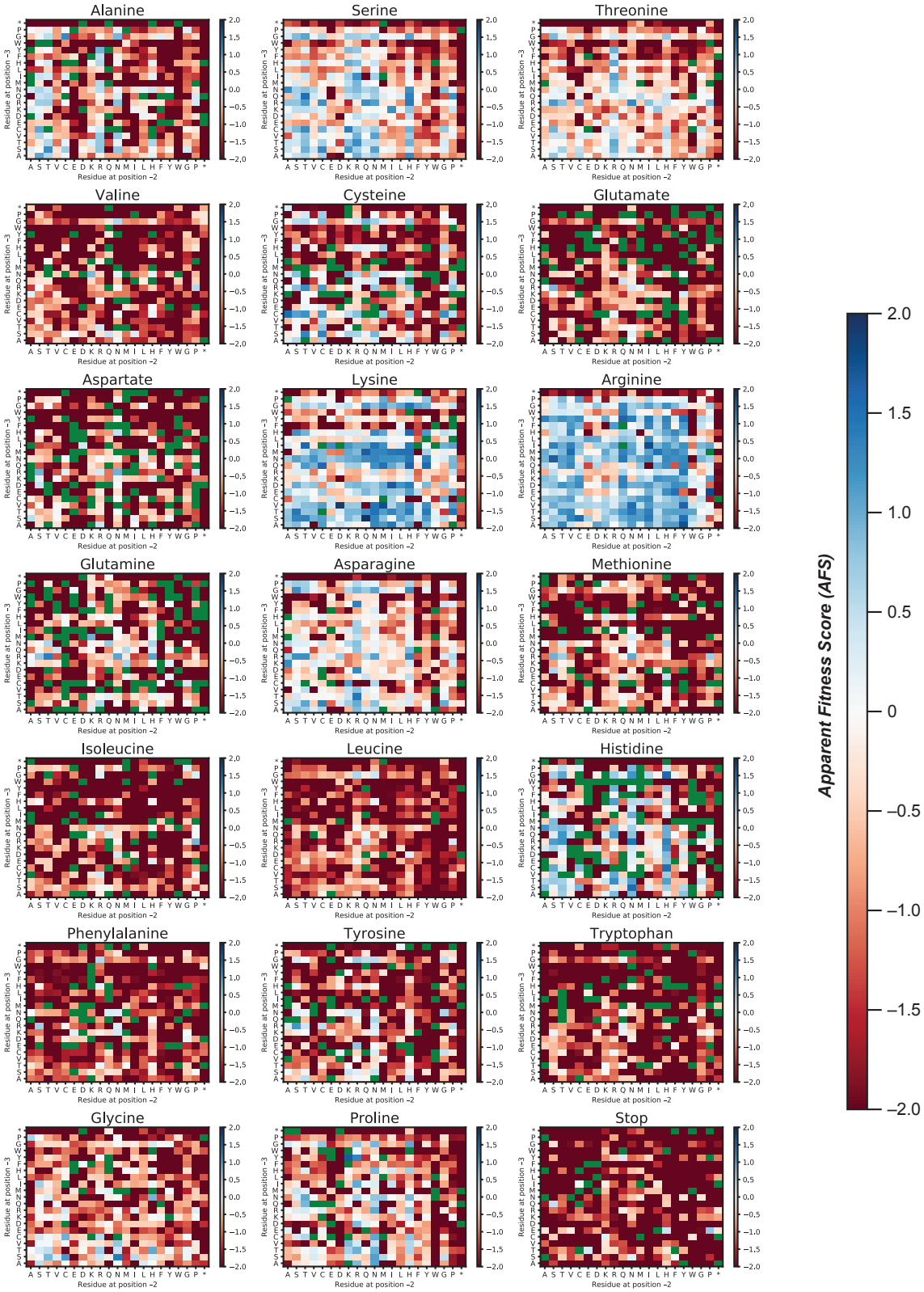


Supplementary Figure 3.13. Chimera analysis of mutations with a large effect on VLP formation. In each structure, T71 and E76 are indicated with labels, and the FG loop is shown in green. A) Hydrogen bond networks are shown in the B form of the FG loop. B) CP[T71E / E76R] is visualized at the B form. T71E is shown in orange, while E76R is shown in purple. C) CP[T71F / E76I] is shown at the C form, where clashes are indicated with yellow lines. D,E) The local environment of the FG loop is shown for (D) the A/C form and (E) B form. In each case, residues within 5 Å are indicated in light blue.

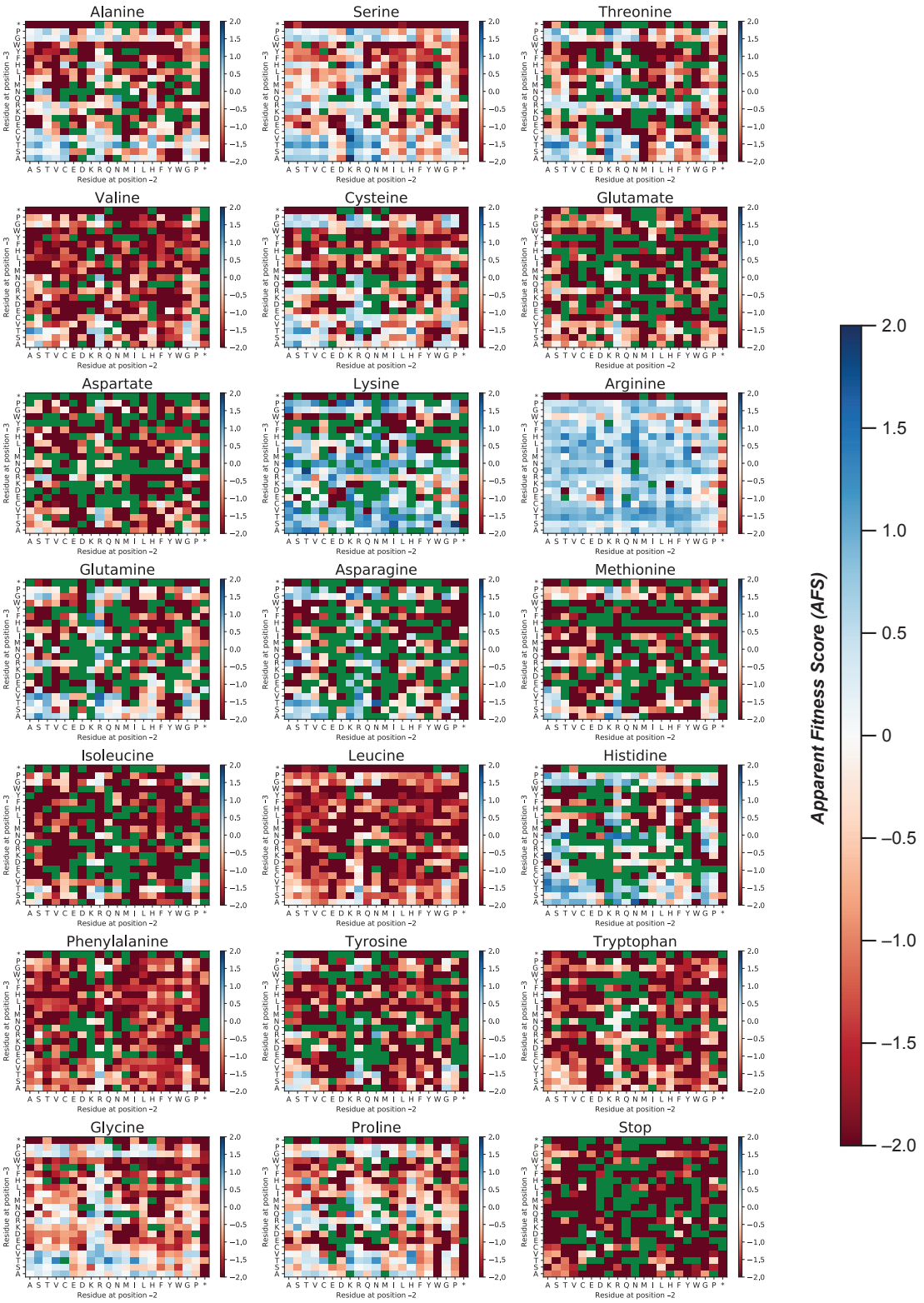


Supplementary Figure 3.14. Instances of positive and negative sign epistasis. A) The effect of two mutations to charged residues is compared across the FG loop. B) Combinations of mutations at position G73 and V75 are evaluated for epistasis. In these graphs, AFS values predicted from the 1D AFL data are shown in green. AFS values are also shown for the measured assembly-selected (orange) and heat-selected (purple) datasets. Differences between the predicted and measured scores indicate epistatic interactions.

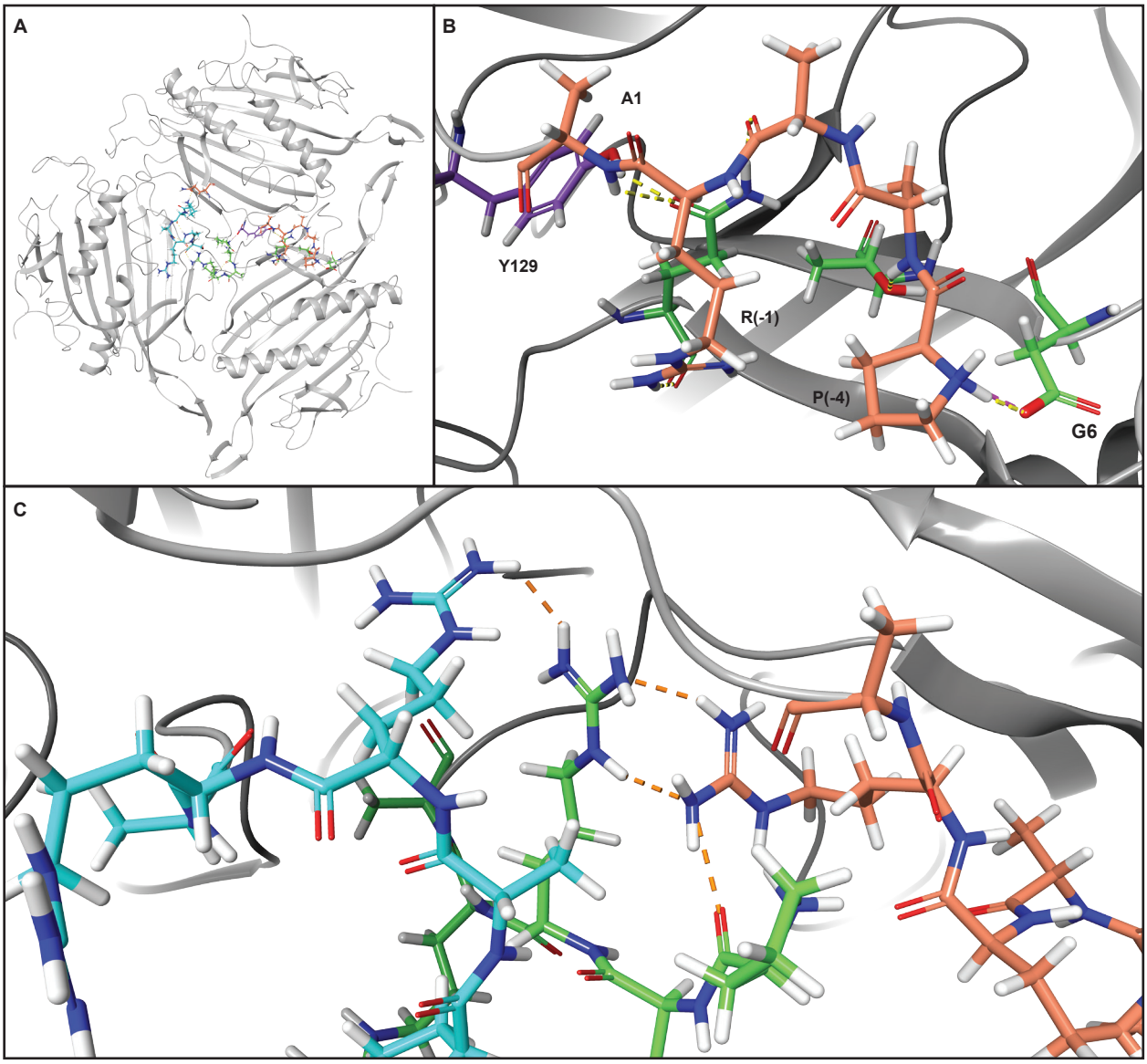
A1.3 Chapter 4 Supplementary Figures



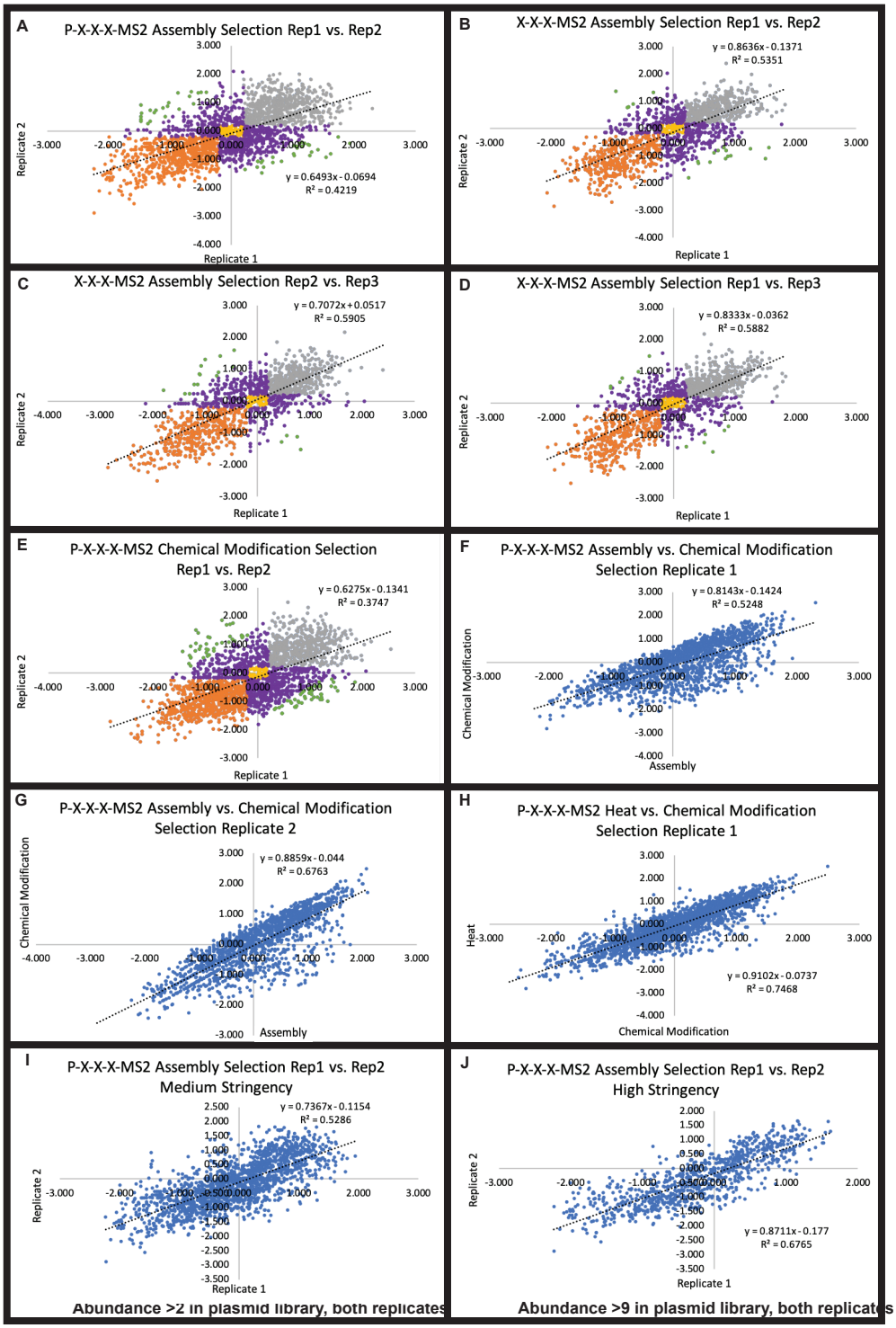
Supplementary Figure 4.1. Labeled full assembly-selected AFL of N-terminal extensions with the pattern P-X-X-X-MS2, as shown in **Figure 4.3**.



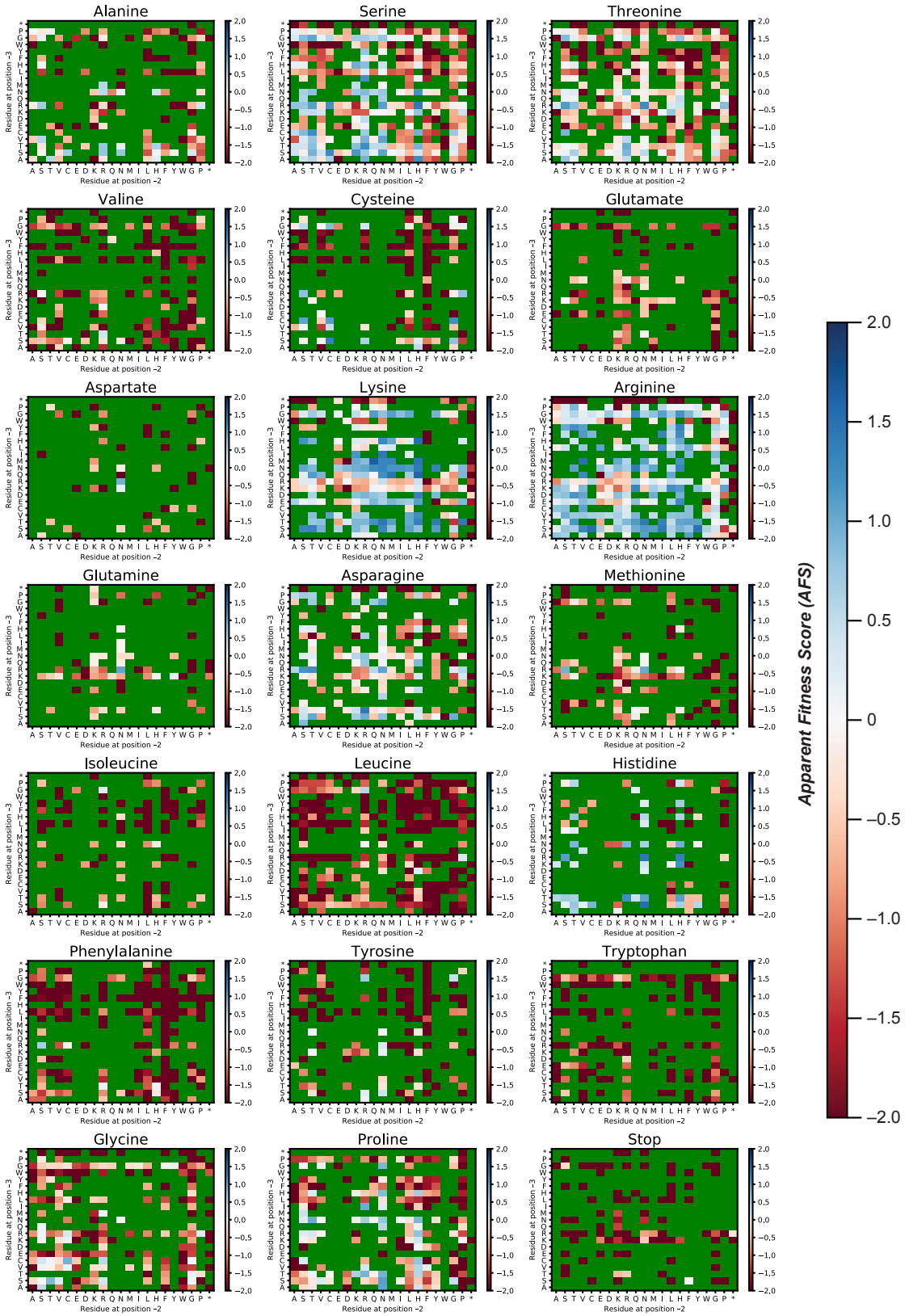
Supplementary Figure 4.2. Assembly-selected AFL of N-terminal extensions with the pattern X-X-X-MS2, without a proline at the –4 position. Blue indicates enriched amino acids, while red indicates combinations that are not enriched. Variants that were present in the plasmid library but absent in the VLP library are indicated in dark red while missing values are shown in green. Nonsense mutations are marked with an asterisk or "Stop".



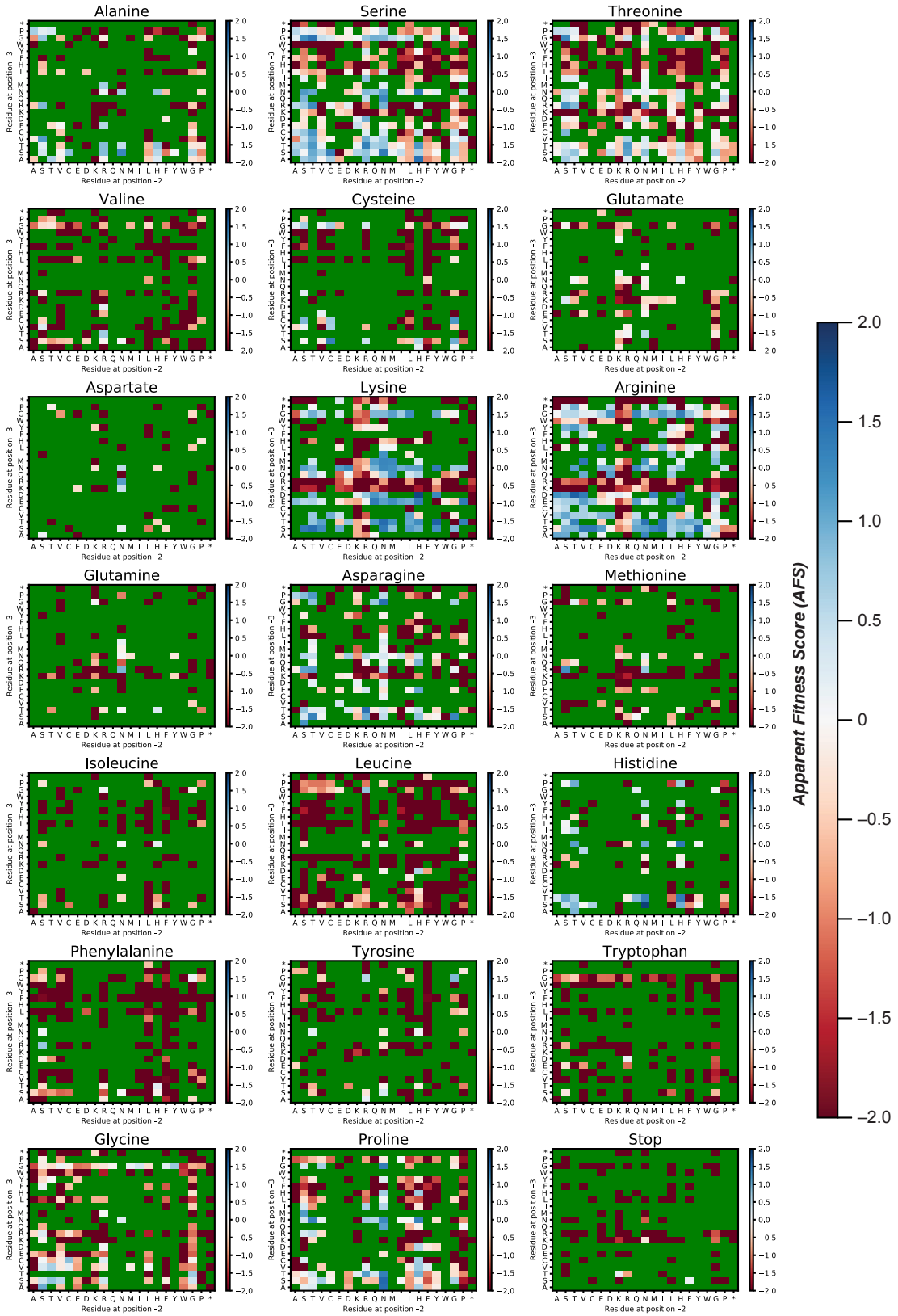
Supplementary Figure 4.3. Modeling of N-terminally extended MS2 coat protein variants. A) Modeling of the extended variants uses a hexameric subunit of the MS2 capsid. The A, B, and C forms of the coat protein monomer are labeled in blue, orange, and green respectively. B) A view of P-A-A-R MS2 is shown with hydrogen bonds in yellow and salt bridges in pink. C) A close up of P-A-R-R MS2 is presented with van der Waals clashing interactions



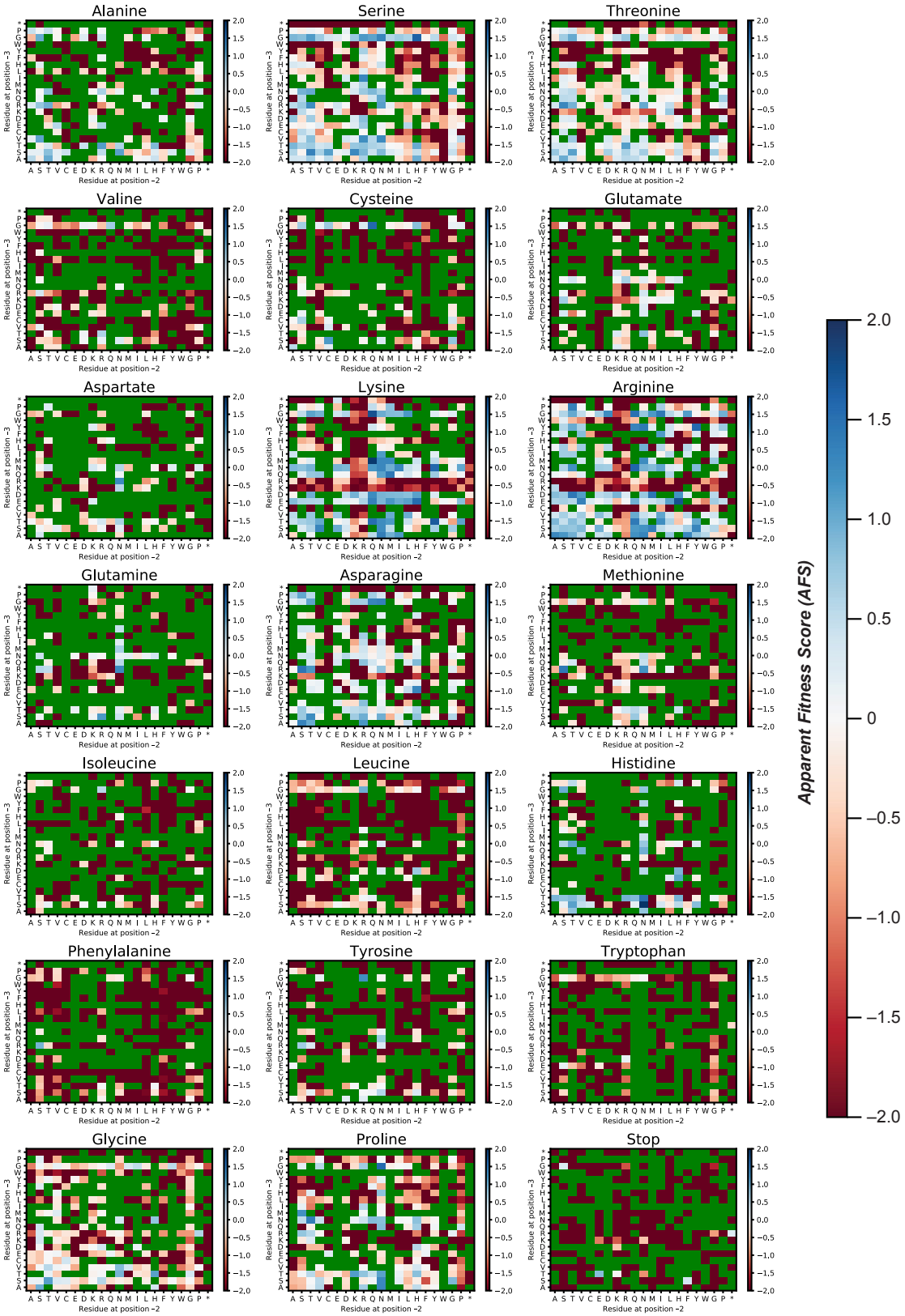
Supplementary Figure 4.4. Correlation analyses of libraries described in this work. In all cases, variants with –4 or nan values in either replicate were excluded. The correlations for A) P-X-X-X-MS2 Assembly Replicates 1-2 and B-D) X-X-X-MS2 Assembly replicates 1-3 are shown. E) Correlation of modification conditions for P-X-X-X-MS2 is shown. In each case, extensions with two positive scores (> 0.2) are shown in gray, and replicates with two negative scores (<–0.2) are shown in orange. Replicates with scores close from –0.2 to 0.2 in both replicates are shown in yellow, while green indicates extensions that are both inverted and greater than 1.5 apart. The remaining scores are in purple. F-H) correlation between various selections within a given replicate are shown. I) Medium and J) high stringency correlations for the P-X-X-X-MS2 library, determined by abundances (>2 and >9, respectively), are shown.



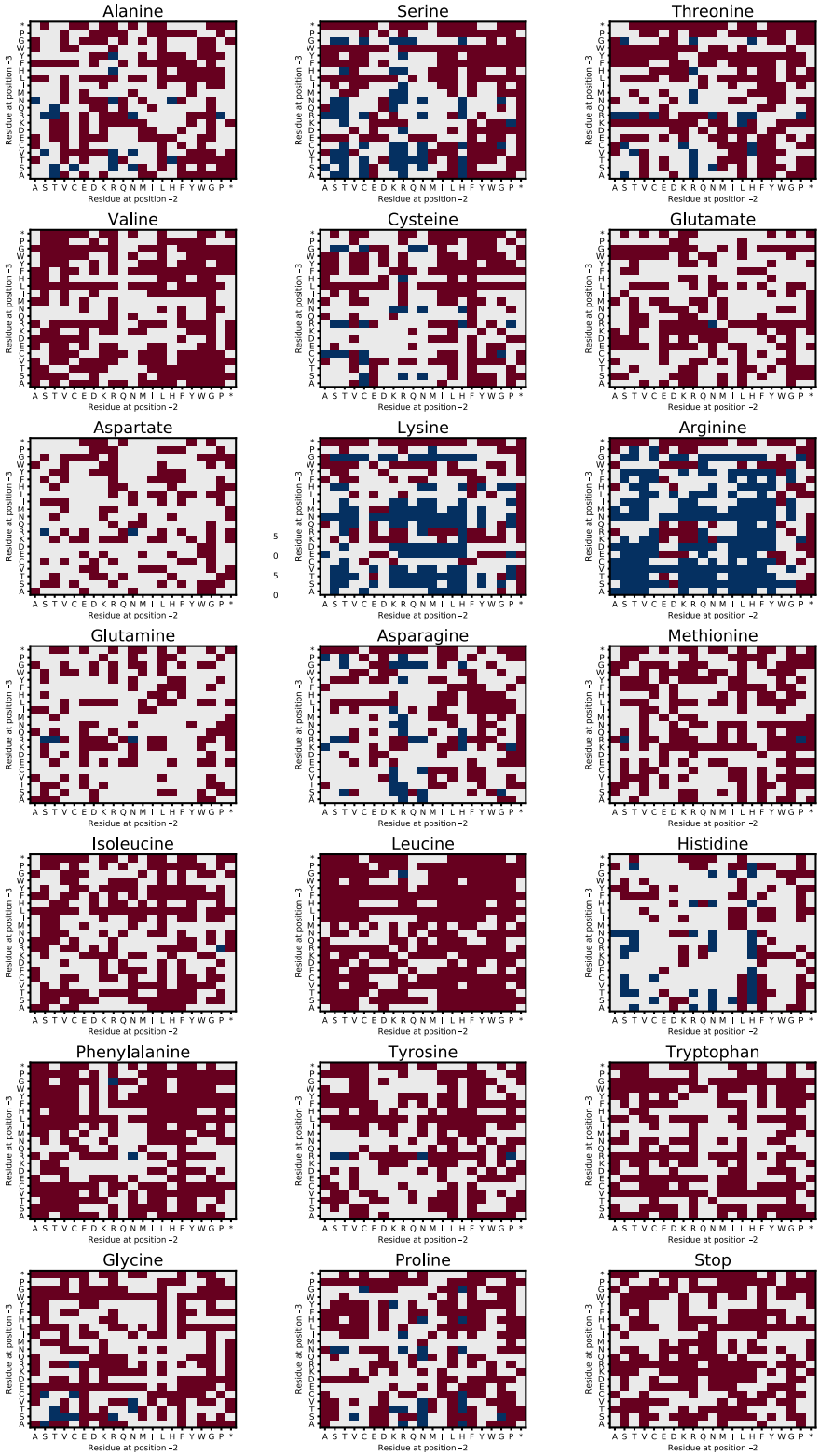
Supplementary Figure 4.5. Apparent Fitness Landscape of P-X-X-X-MS2 N-terminal extensions with low read filter. Variants with three or fewer reads in the plasmid library of either replicate were treated as a missing value and are shown in green.



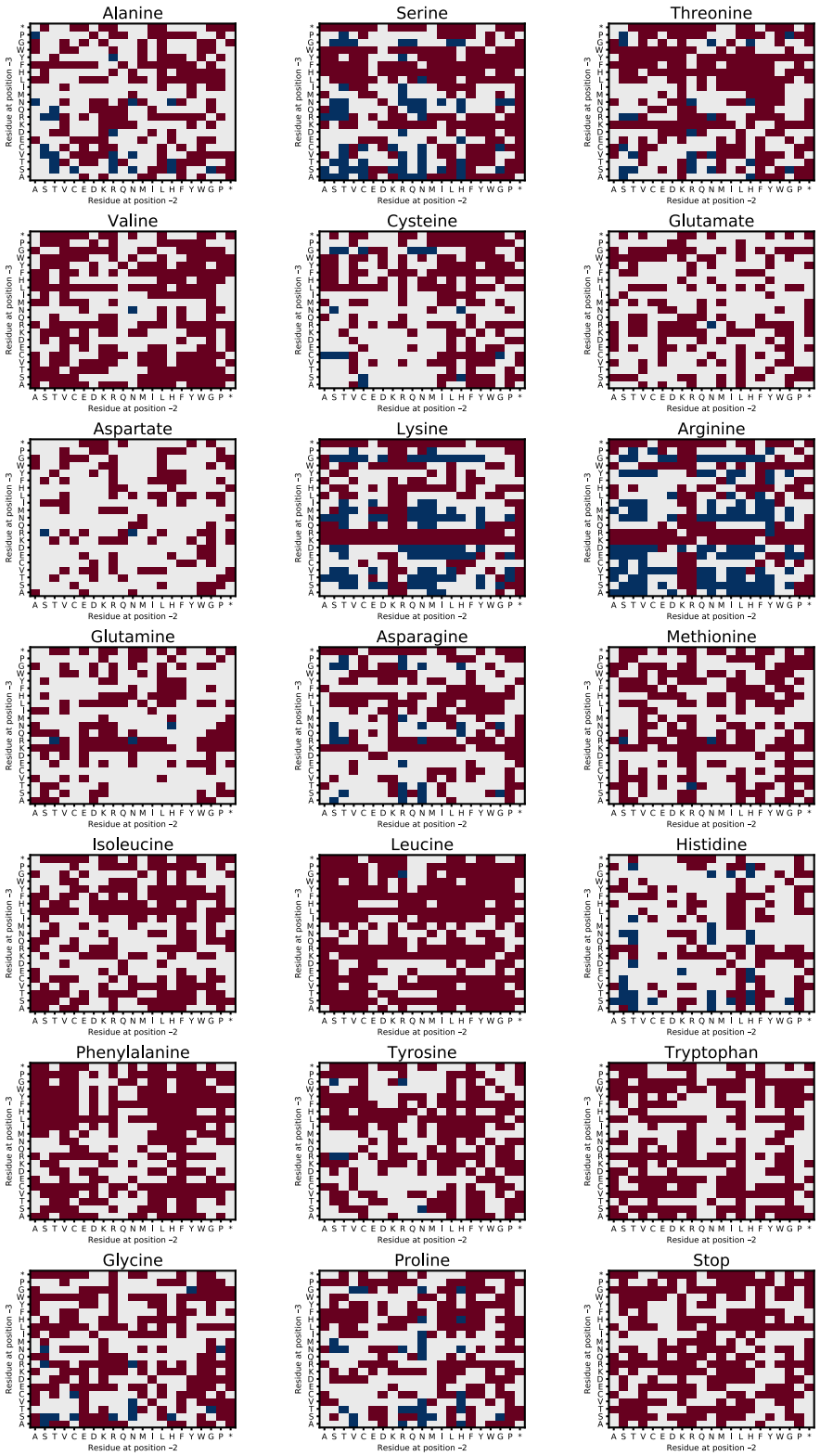
Supplementary Figure 4.6. Chemical challenge of P-X-X-X-MS2 N-terminal extensions with low read filter. Variants with three or fewer reads in the plasmid library of either replicate were treated as a missing value and are shown in green.



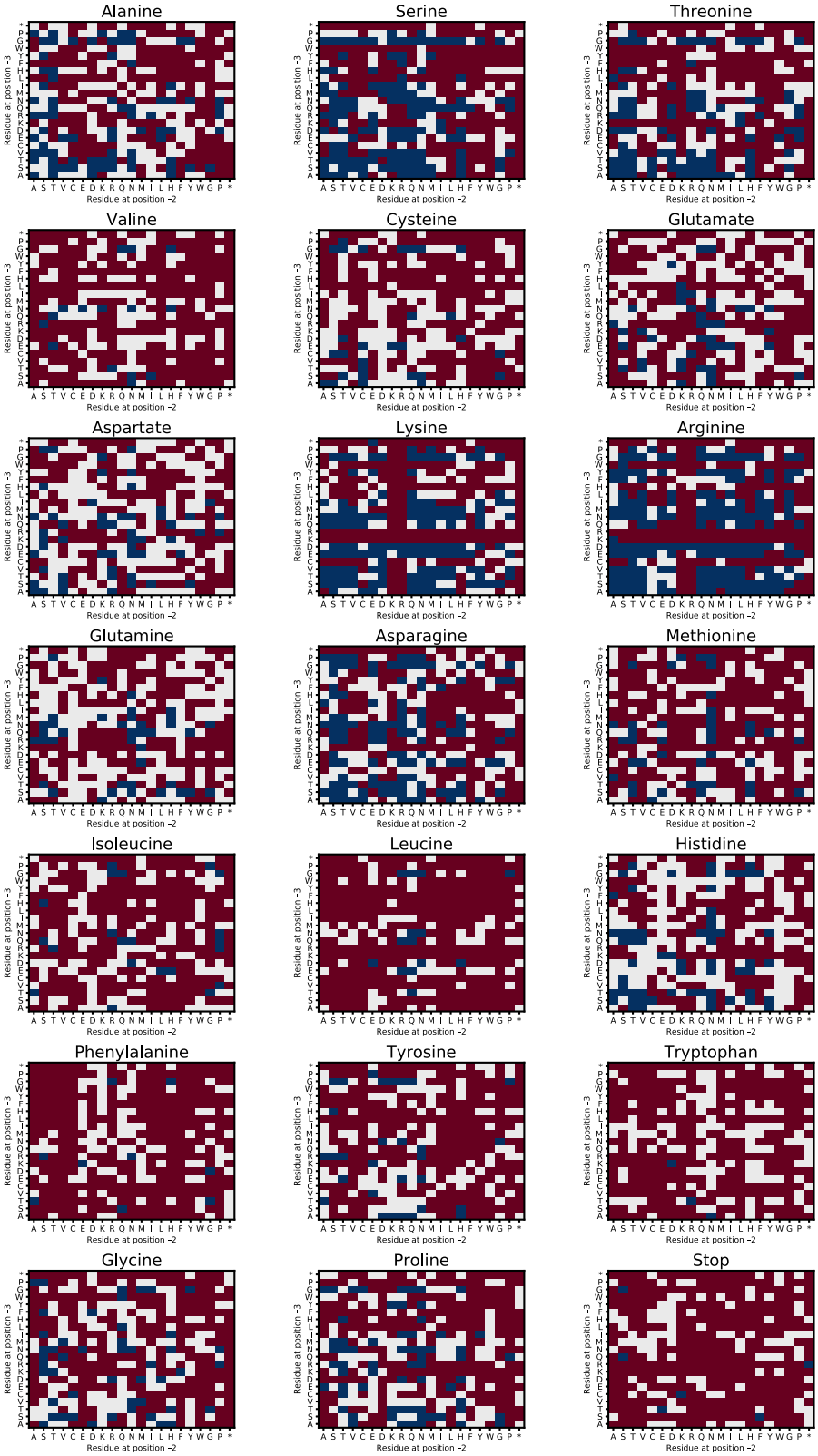
Supplementary Figure 4.7. Heat challenge of P-X-X-X-MS2 N-terminal extensions with low read filter. Variants with three or fewer reads in the plasmid library were treated as a missing value and are shown in green.



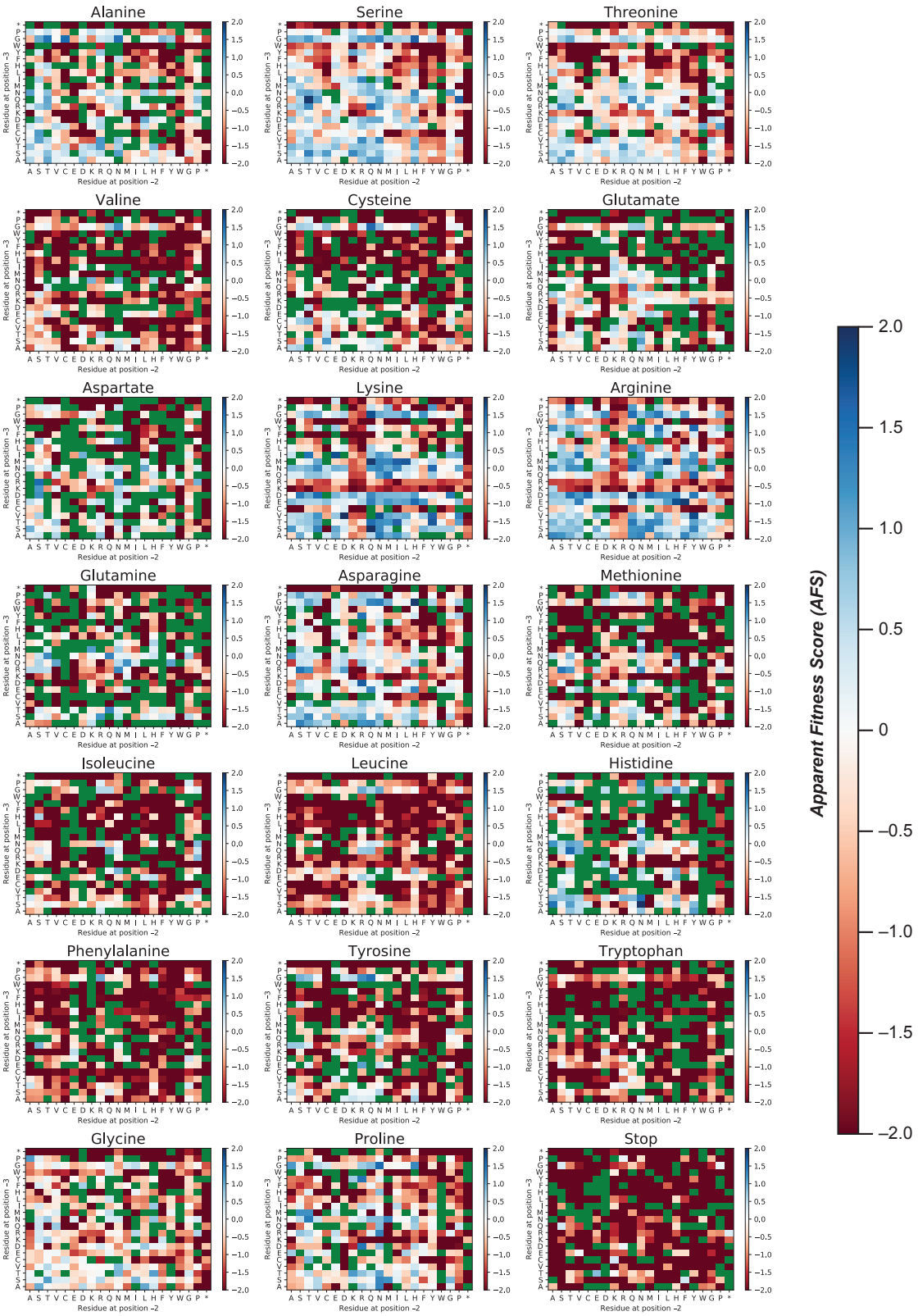
Supplementary Figure 4.8. Apparent Fitness Landscape of P-X-X-X-MS2 N-terminal extensions with alternative data processing. Variants with an AFL score between -0.2 and 0.2 in either replicate were removed and are shown in grey. Additionally, variants with AFL scores with opposite signs across replicates are shown in grey. Variants in which both replicate scores are >0.2 are shown in blue and variants with both replicate scores <-0.2 are shown in red.



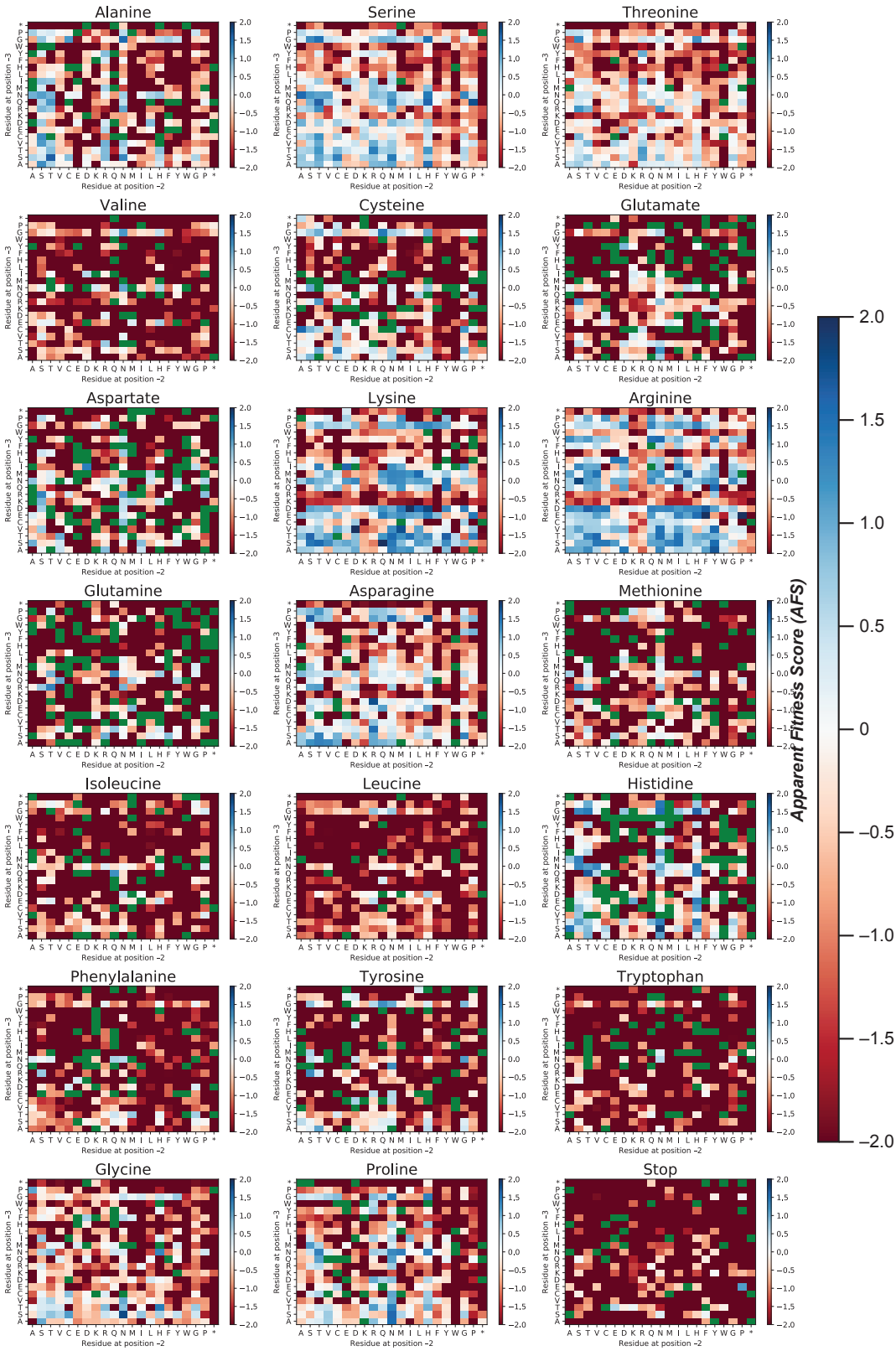
Supplementary Figure 4.9. Chemical challenge of P-X-X-X-MS2 N-terminal extensions with alternative data processing. Variants with chemical challenge score between -0.2 and 0.2 in either replicate were removed and are shown in grey. Additionally, variants with scores with opposite signs across replicates are shown in grey. Variants in which both replicate scores are >0.2 are shown in blue and variants with both replicate scores <-0.2 are shown in red.



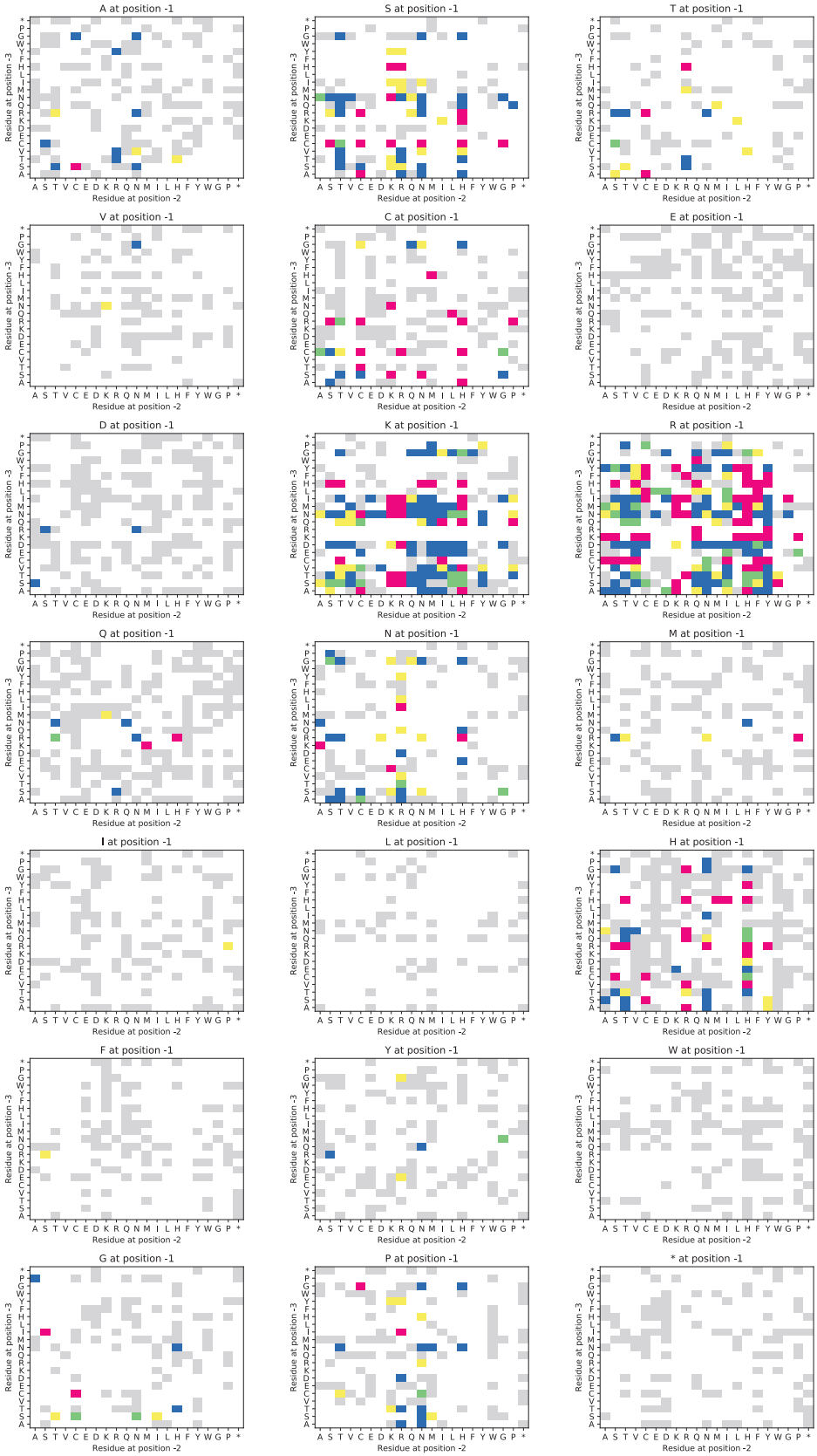
Supplementary Figure 4.10. Heat challenge of P-X-X-X-MS2 N-terminal extensions with alternative data processing. Variants with a score between -0.2 and 0.2 in were removed and are shown in grey. Variants with heat challenge scores >0.2 are shown in blue and variants with a score <-0.2 are shown in red.



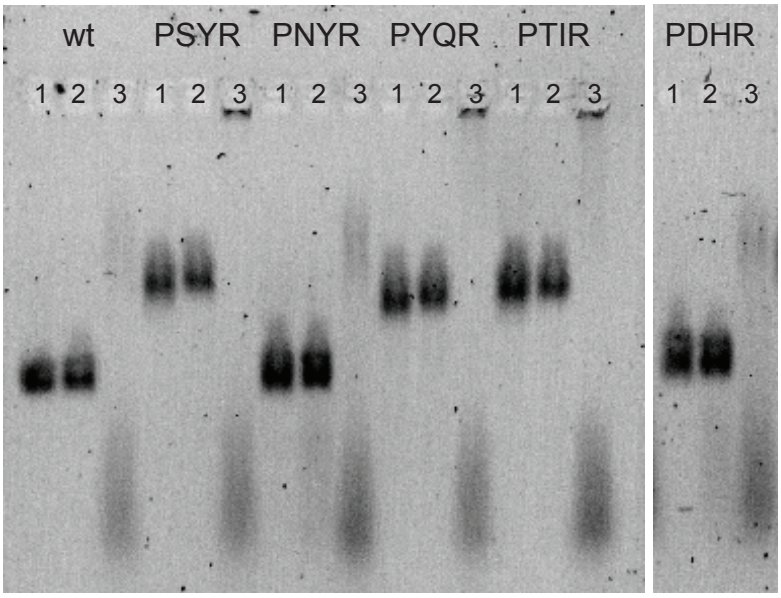
Supplementary Figure 4.11. Heat-selected AFL for the P-X-X-X-MS2 library. The library was subjected to 50 ºC for ten minutes, and assembled VLPs were enriched with semi-preparative HPLC size exclusion chromatography. Enriched amino acids are blue, variants that are not enriched are indicated in red, and missing values are shown in green.



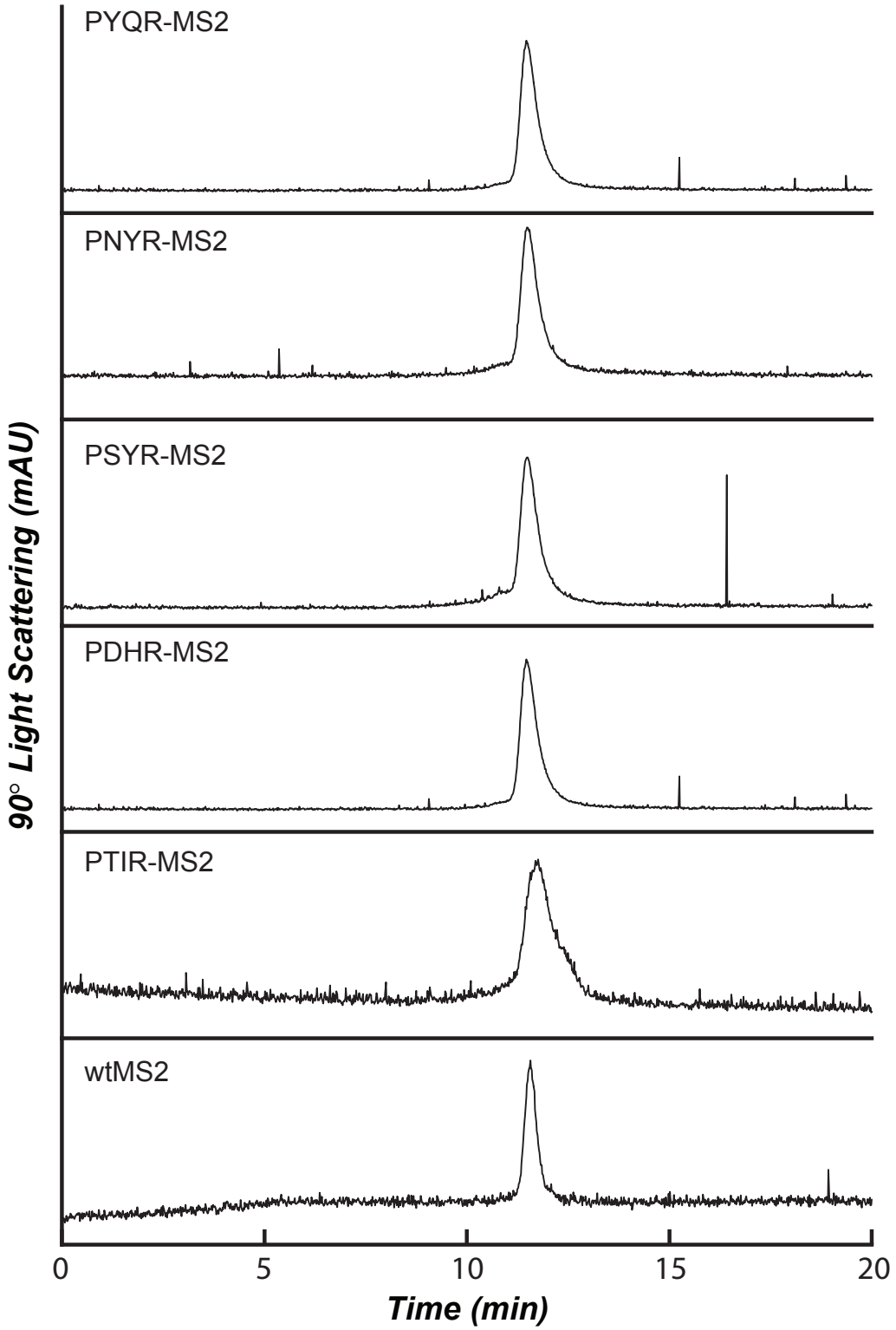
Supplementary Figure 4.12. Chemical modification AFL of the P-X-X-X-MS2 library. The library was subjected to chemical modification conditions, and assembled VLPs were enriched with semi-preparative HPLC size exclusion chromatography. Enriched combinations are blue, variants that are not enriched are shown as red, and missing values are green.



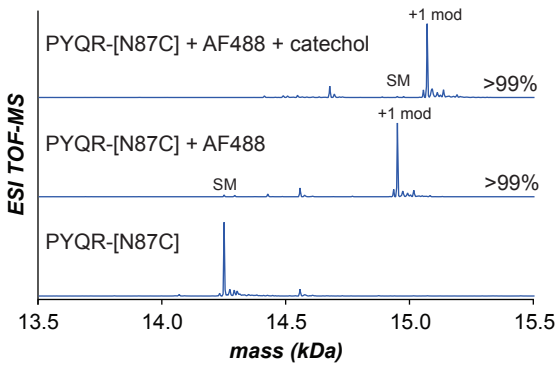
Supplementary Figure 4.13. An aggregated heatmap combining results from the assembly, thermal, and chemical modification selections. All combinations of P-X-X-X-MS2 are given a color based on the key shown in **Figure 5B**.



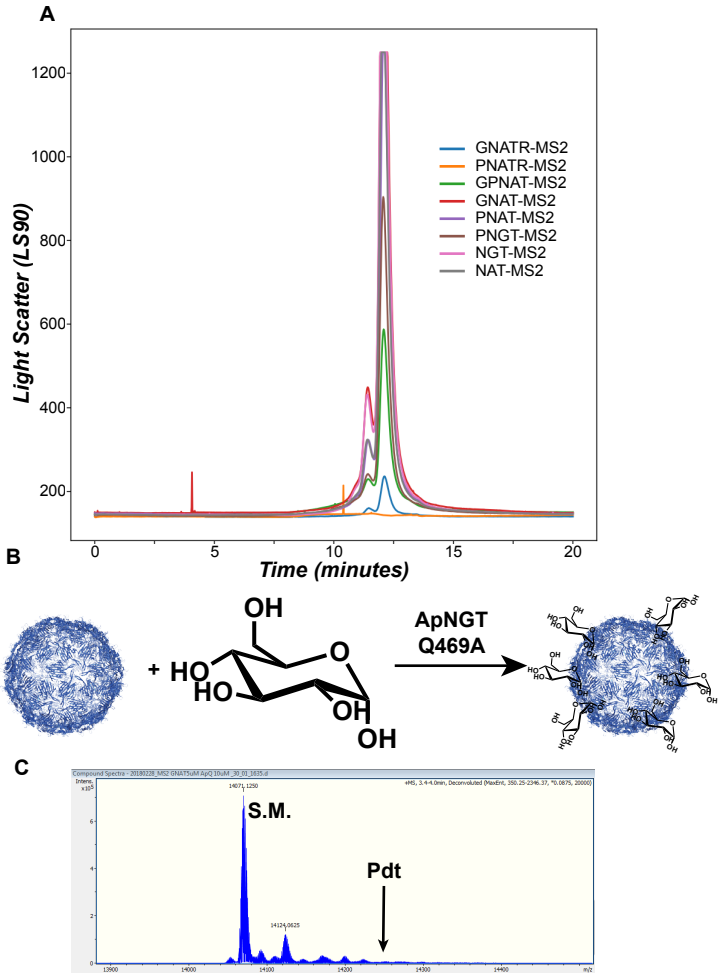
Supplementary Figure 4.14. Native agarose gel of HiPerX variants following a thermal challenge. Lanes 1-3 for each variant were incubated at room temperature, 50 °C, and 100 °C respectively.



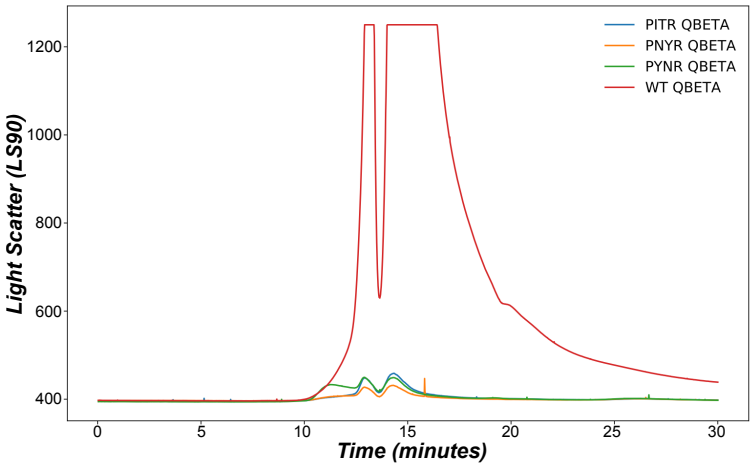
Supplementary Figure 4.15. HPLC SEC traces of CP[HiPerX] variants following $K_3Fe(CN)_6$-mediated oxidative coupling.



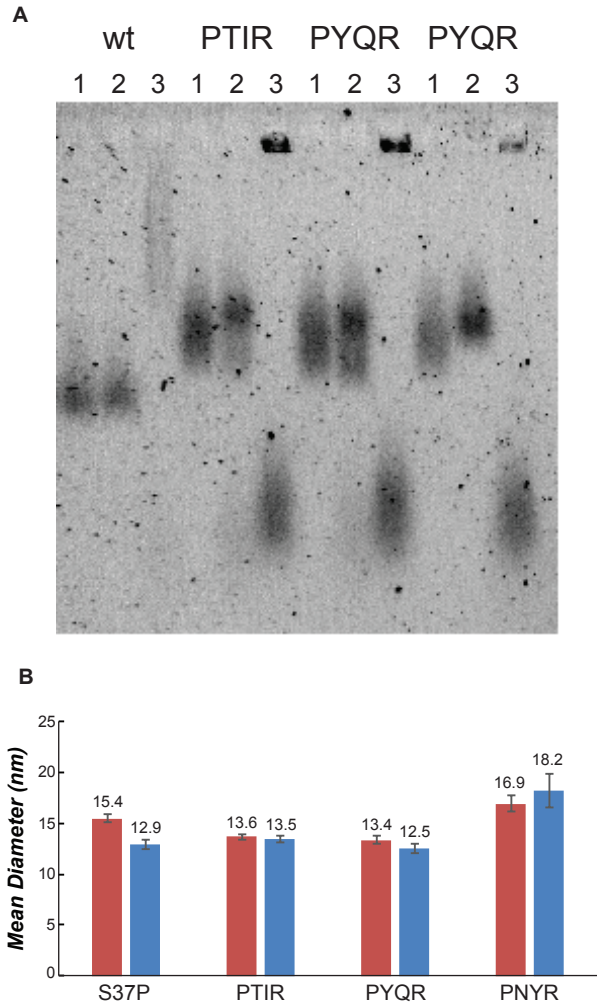
Supplementary Figure 4.16. Dual chemical modification of CP[PYQR-N87C] MS2.



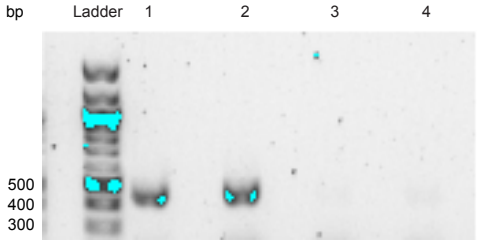
Supplementary Figure 4.17. MS2 compatibility with glycosylation sequences. A) Various glycosylation sequences are evaluated by HPLC SEC and permit assembly. B) Scheme of *in vitro* glycosylation experiment with ApNGT Q469A, a highly active enzyme that appends glucose at specific protein sequences. C) Representatitive deconvolution spectra showing no detectable modification. Starting materials (S.M.) and product (Pdt) are shown for GNAT-MS2.



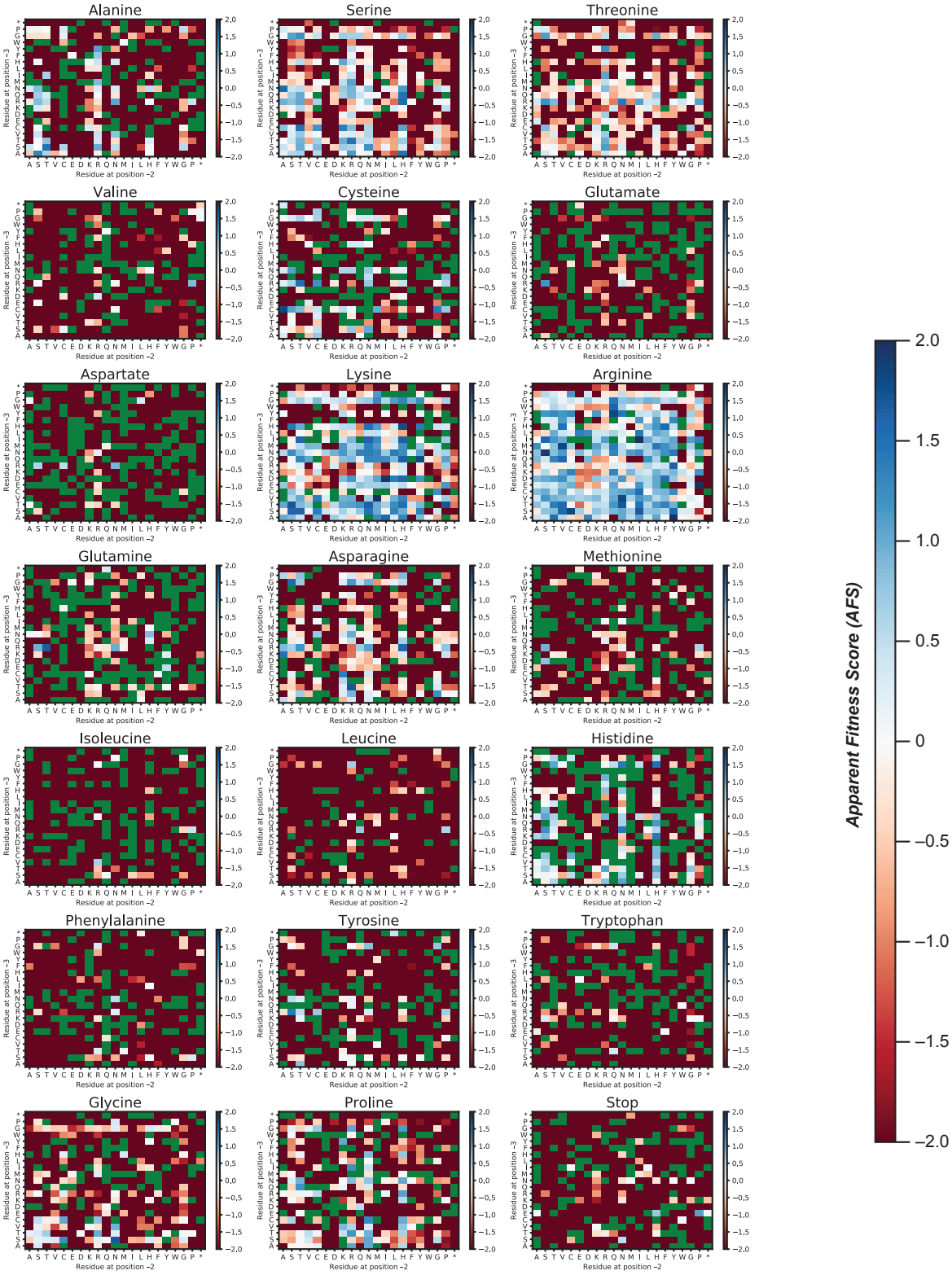
Supplementary Figure 4.18. Assembly of HiPerX variants on Qβ. HPLC SEC of all three HiPerX variants shows that the variants did not assemble into VLPs.



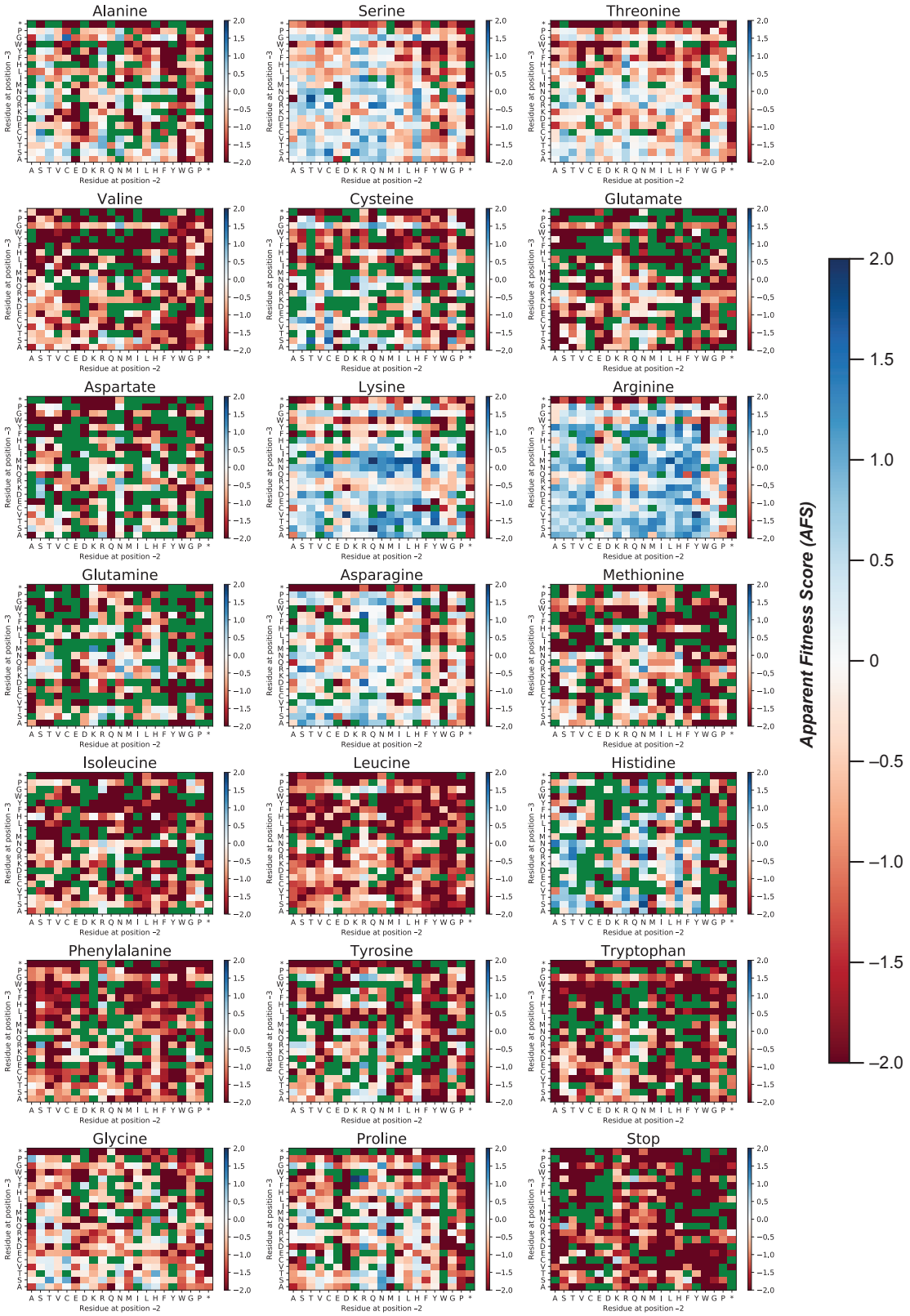
Supplementary Figure 4.19. CP[HiPerX–S37P] variants following thermal and modification challenges. A) Native agarose gel of CP[HiPerX-S37P] variants following a thermal challenge. Lanes 1-3 for each variant represent incubation at room temperature, 50° C, and 100° C respectively. B) DLS of CP[HiPerX-S37P] prior to small molecule oxidative coupling (red) and following modification (blue).



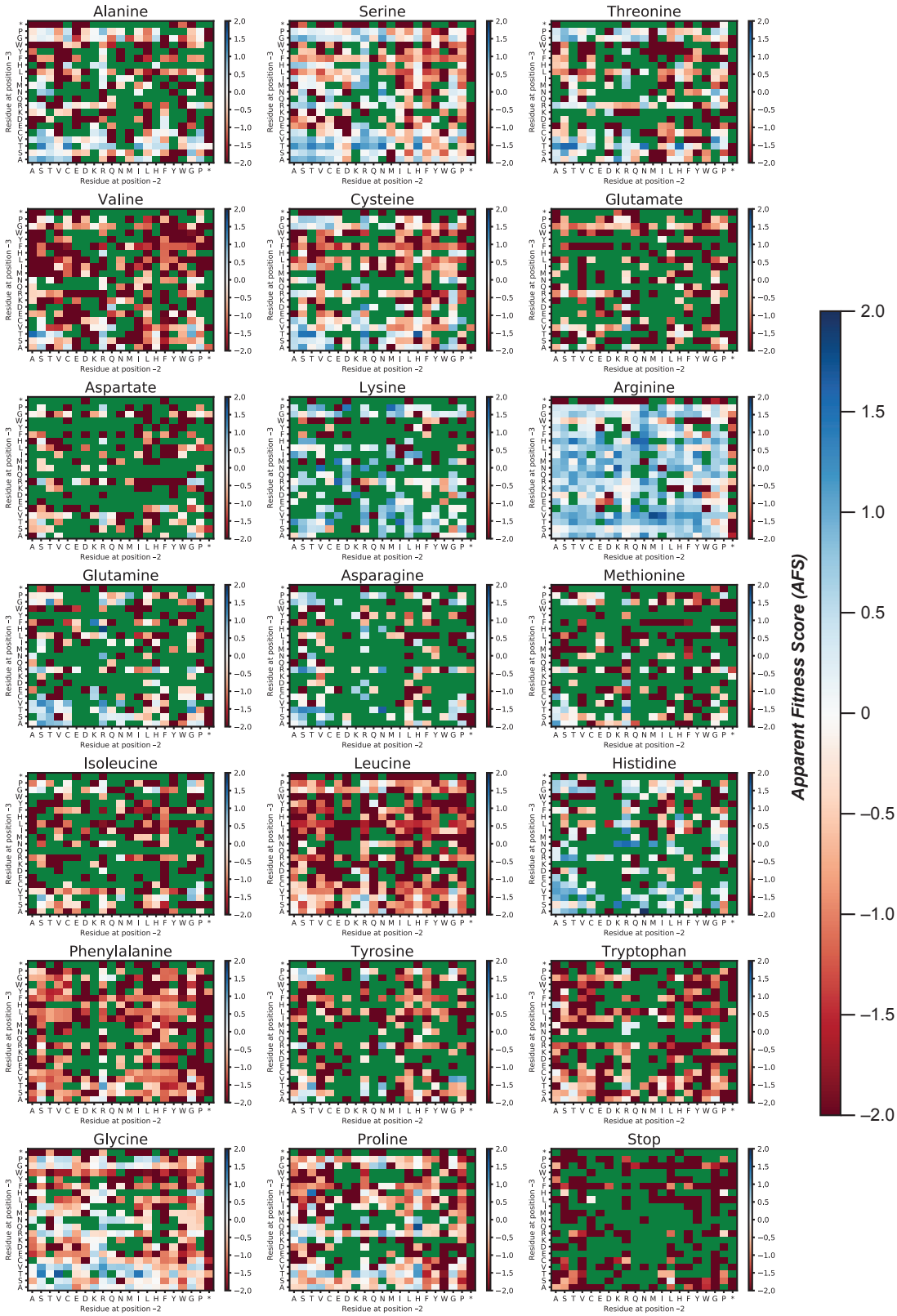
Supplementary Figure 4.20. Comparison of polyT and random hexamer primers for cDNA synthesis. The agarose gel shows DNA following cDNA synthesis and the first PCR amplification of barcoding. Lanes 1 and 2 contain DNA synthesized from wtMS2-derived RNA using polyT and random hexamer primers, respectively. Lanes 3 and 4 contain DNA synthesized from an assembly-incompetent MS2 variant using polyT and random hexamer primers, respectively. Image was cropped to remove primer dimer bands.



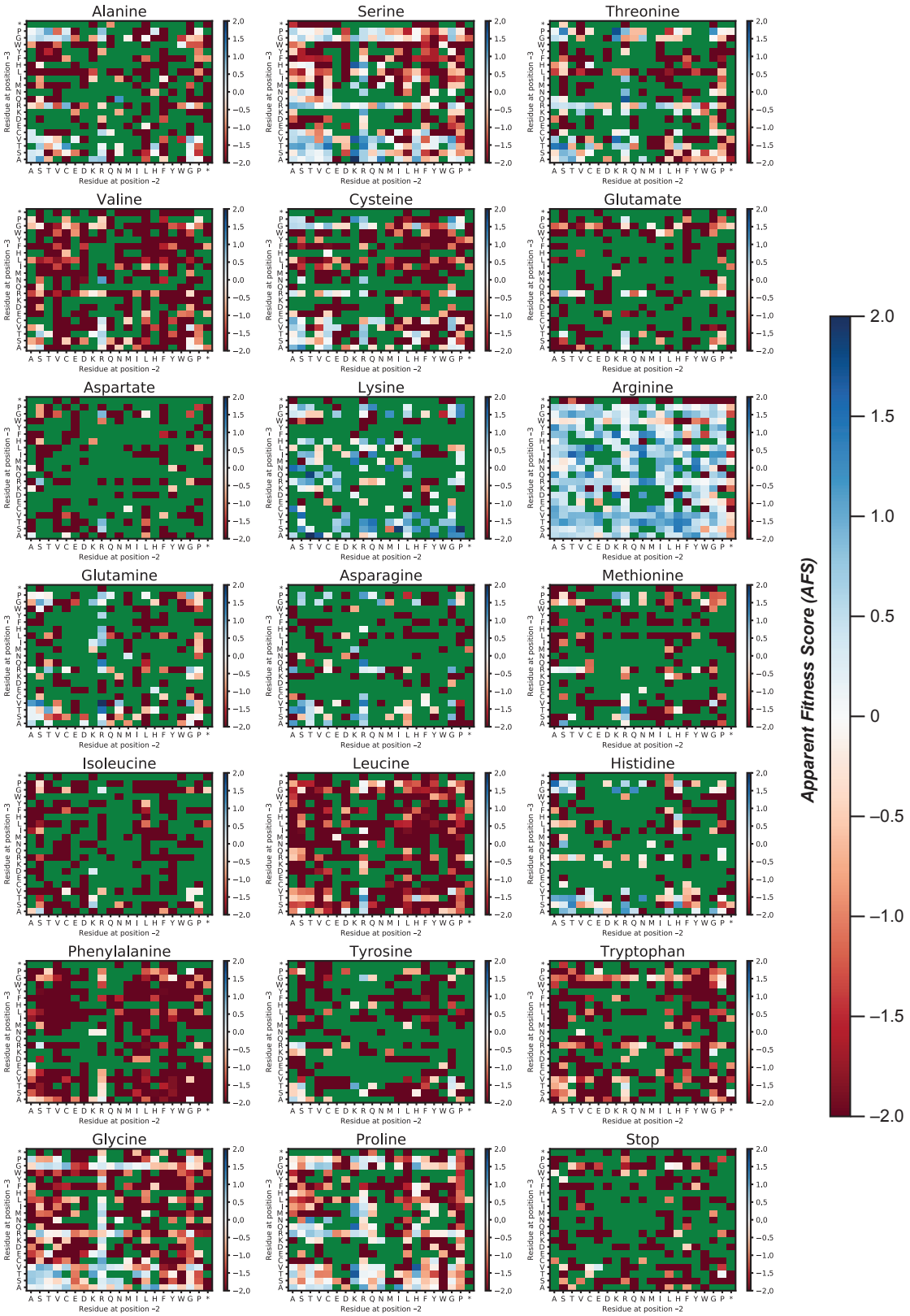
Supplementary Figure 4.21. Replicate one of the assembly-selected AFL of N-terminal extensions with the pattern P-X-X-X-MS2.



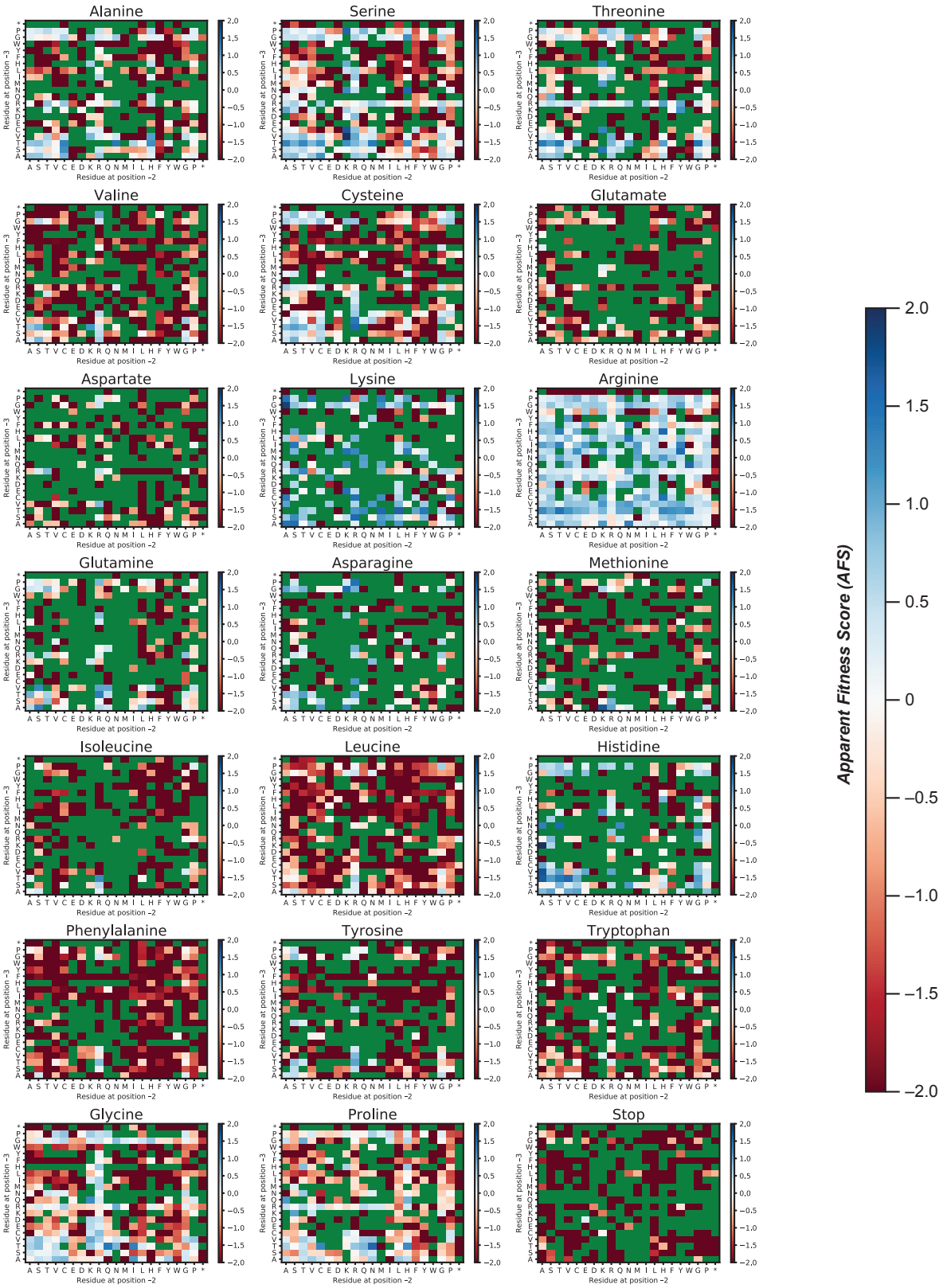
Supplementary Figure 4.22. Replicate two of the assembly-selected AFL of N-terminal extensions with the pattern P-X-X-X-MS2.



Supplementary Figure 4.23. Replicate one of the assembly-selected AFL of N-terminal extensions with the pattern X-X-X-MS2.



Supplementary Figure 4.24. Replicate two of the assembly-selected AFL of N-terminal extensions with the pattern X-X-X-MS2.

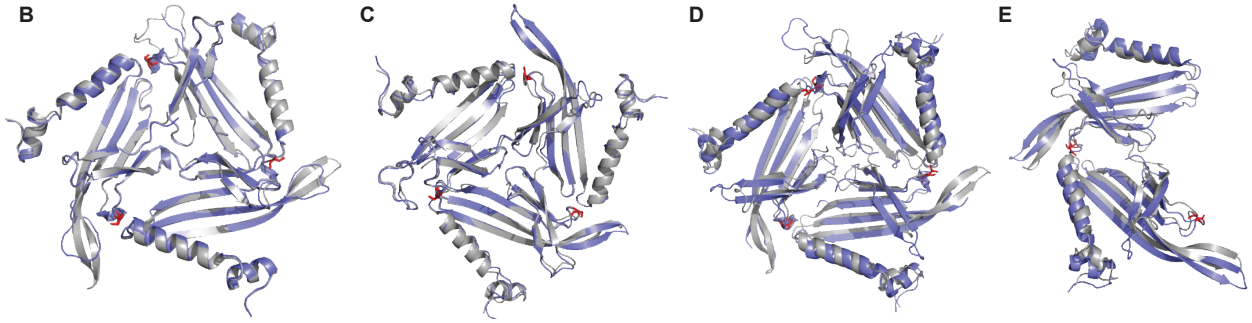


Supplementary Figure 4.25. Replicate three of the assembly-selected AFL of N-terminal extensions with the pattern X-X-X-MS2.

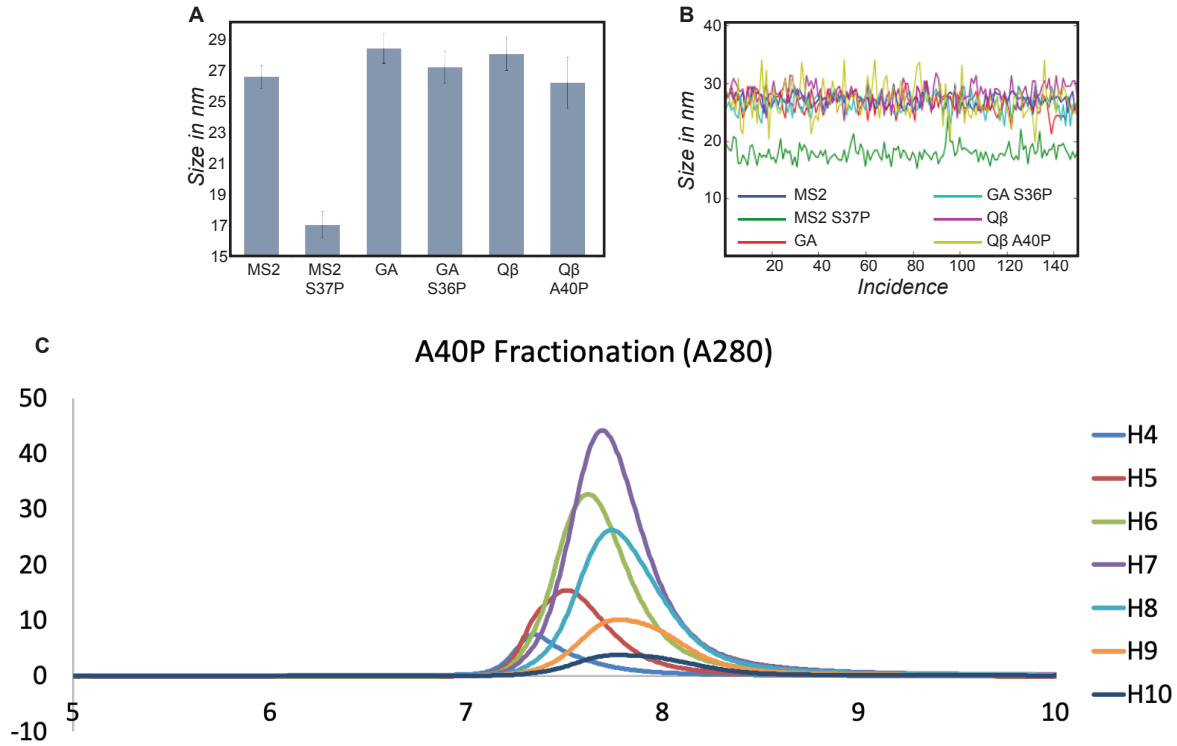
A1.4 Chapter 5 Supplementary Figures

A

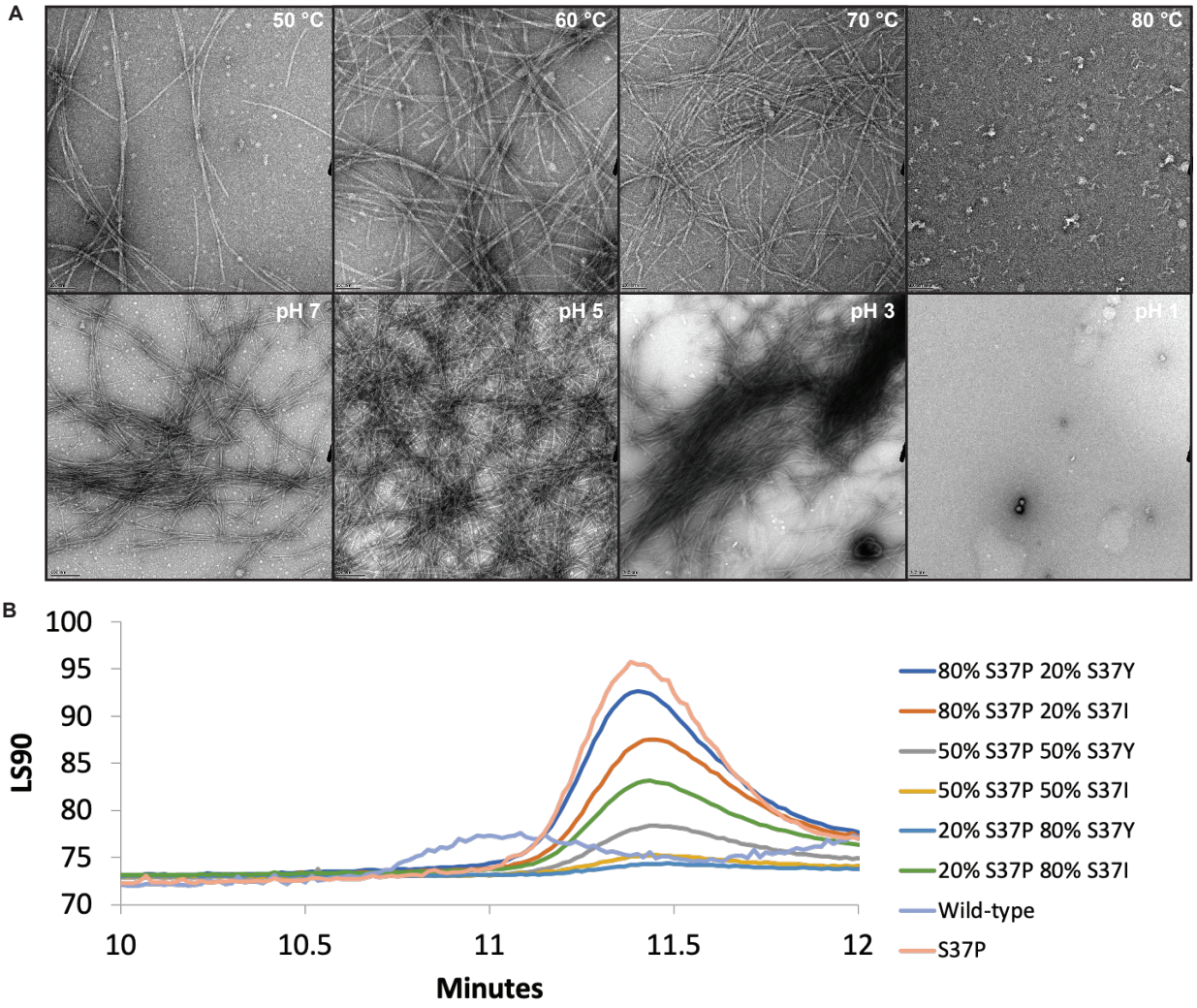
```
                % ID to MS2
MS2             100.00   MASNFTQFVLVDNGGTGD--VTVAPSN--FANGVAEWISSNS-RSQAYKVTCSVRQSSAQ 55
Fr              86.92    MASNFEEFVLVDNGGTGD--VKVAPSN--FANGVAEWISSNS-RSQAYKVTCSVRQSSAN 55
GA              59.69    -MATLRSFVLVDNGGTGN--VTVVPVS--NANGVAEWLSNNS-RSQAYRVTASYRASGAD 54
Qβ              20.31    -MAKLETVTLGNIGKDGKQTLVLNPRGVNPTNGVASLSQACAYPALEKRVTVSVSQPSRN 59
Prr1            26.98    -MAQLQNLVLKDREATP-NDHTFVPRDIR--DNVGEVVESTGYPIGESRFTISLRKTSNG 56
                           : :  ..* :           . * .    :.*..  .  .      :.* *    .

MS2             100.00   NRKYTIKVEVPKVATQTVGGVELPVAAWRSYLNMELTIPIFATNSDCELIVKAMQGLLKD 115
Fr              86.92    NRKYTVKVEVPKVATQVQGGVELPVAAWRSYMNMELTIPVFATNDDCALIVKALQGTFKT 115
GA              59.69    KRKYAIKLEVPKIVTQVVNGVELPGSAWKAYASIDLTIPIFAATDDVTVISKSLAGLFKV 114
Qβ              20.31    RKNYKVQVKIQNPTACTANGSCDPSVTRQAYADVTFSFTQYSTDEERAFVRTELAALLAS 119
Prr1            26.98    RYKSTLKLVVPVVQSQTVNGIVTPVVVRTSYVTVDFDYDARSTTKERNNFVGMIADALKA 116
                         . :  ::: :    : . .*   *  .  :*  : :     :: .:   .   :   :

MS2             100.00   GNP-IPSAIAANSGIY-    130
Fr              86.92    GNP-IATAIAANSGIY-    130
GA              59.69    GNP-IAEAISSQSGFYA    130
Qβ              20.31    PL--LIDAIDQLNPAY-    133
Prr1            26.98    DLMLVHDTIVNLQGVY-    132
                            :  :*   .  *
```



Supplementary Figure 5.1. Coat protein genes related to the MS2 bacteriohpage. A) Sequence alignment between MS2, Fr, GA, Qβ, and Prr1 bacteriophage. Sequence identity compared to the MS2 CP is shown, and positions analogous to 37 in MS2 are boxed in black. B-E) Structural alignment between MS2 and (B) Fr, (C) GA, (D) Qβ, and (E) Prr1 bacteriophage.



Supplementary Figure 5.2. Effect of A40P mutation in Qβ. A) Diameters and B) variance across 150 VLPs quantified from TEM images are shown. C) HPLC SEC of fractionated VLPs show that the size shift is retained. A low H indicates fractions that were collected earlier, and a high H indicates fractions that were collected later in an HPLC



Supplementary Figure 5.3. Fibril tolerance to thermal, acid, and assembly challenges. A) Fibrils were visible following 70 °C thermal challenge and pH 3, though morphological changes are visible. B) Disassembly–reassembly reactions with increasing amounts of fibril-forming mutant did not yield wild-type sized particles when combined with S37P.