Comprehensive Fitness Landscape of a Multi-Geometry Protein Capsid Informs Machine Learning Models of Assembly

Daniel D. Brauer,[†] Celine B. Santiago,[†] Zoe N. Merz,[†] Esther McCarthy,[†]

Danielle Tullman-Ercek,^{*,‡} and Matthew B. Francis^{*,¶,§}

[†]*Department of Chemistry, University of California Berkeley, Berkeley, CA*

[‡]*Department of Chemical and Biological Engineering, Northwestern University, Evanston,*

IL

[¶]*Department of Chemistry, University of California, Berkeley Berkeley, CA*

[§]*Materials Sciences Division, Lawrence Berkeley National Laboratories, Berkeley, CA*

E-mail: ercek@northwestern.edu; mbfrancis@berkeley.edu

Abstract

Virus-like particles (VLPs) are non-infectious viral-derived nanomaterials poised for biotechnological applications due to their well-defined, modular self-assembling architecture. Although progress has been made in understanding the complex effects that mutations may have on VLPs, nuanced understanding of the influence particle mutability has on quaternary structure has yet to be achieved. Here, we generate and compare the apparent fitness landscapes of two capsid geometries (T=3 and T=1 icosahedral) of the bacteriophage MS2 VLP. We find significant shifts in mutability at the symmetry interfaces of the T=1 capsid when compared to the wildtype T=3 assembly. Furthermore, we use the generated landscapes to benchmark the performance of *in silico* mutational scanning tools in capturing the effect of missense mutation on complex

particle assembly. Finding that predicted stability effects correlated relatively poorly with assembly phenotype, we used a combination of *de novo* features in tandem with *in silico* results to train machine learning algorithms for the classification of variant effects on assembly. Our findings not only reveal ways that assembly geometry affects the mutable landscape of a self-assembled particle, but also establish a template for the generation of predictive mutational models of self-assembled capsids using minimal empirical training data.

Introduction

Virus-like particles (VLPs) have emerged as promising scaffolds for many applications in biotechnology, including vaccine development,^{1–3} targeted drug delivery,^{4,5} and nanoreactor production.⁶ These well-defined, highly symmetrical closed-shell structures self-assemble from discrete protein building blocks that can be engineered to tune the physical and chemical features of the assembly.⁷ While great strides have been made in *de novo* design of self-assembling VLP mimetics,^{8–13} predicting the effects of mutations to the subunit building blocks of a nanocage remains challenging because even small changes can disrupt or drastically alter the complex network of interactions that drive assembly.^{14–16}

Systematic study of the effects of mutations on a given protein function has recently been made possible by the advent of deep-mutational scanning, a method that uses next-generation sequencing to assess $> 10^5$ protein variants in a single experiment.^{17,18} This technique has previously been applied to self-assembling viral structures such as AAV,^{19,20} influenza,²¹ and polio²² in order to generate fitness landscapes that describe the effects of mutations on viral infectivity. In order to probe the impact of mutations on assembly interactions more directly, we recently developed a strategy called SyMAPS (Systematic Mutagenesis and Assembled Particle Selection), which employs capsid self-assembly as a fitness readout for deep mutational scanning.²³ This method enables a quantitative, systematic understanding of how mutations to the subunits of a closed-shell particle affect the

final assembly state without relying on a selection that bundles many fitness criteria (e.g. infectivity requires successful viral attachment, replication, and assembly).²⁴ SyMAPS has been successfully utilized to quantify the effects of single mutants, epistatic interactions, loop insertions, and peptide extensions in bacteriophage MS2 VLPs, enabling tuned particle thermostability, acid lability, and chemical reactivity.^{25–27}

Here, we employ the SyMAPS approach to generate the apparent fitness landscape of a non-native assembly geometry of an icosahedral VLP. Prior work uncovered a point mutant (MS2[S37P]) in the coat protein of the well-studied MS2 bacteriophage VLP that results in a shift in quaternary assembly of the particle.¹⁶ This shift from T=3 to T=1 icosahedra dramatically reorganizes the capsid assembly while maintaining nearly identical primary, secondary, and tertiary structures of the constituent coat protein subunits (Figure 1).

Figure 1: Bacteriophage MS2 virus-like particle assemblies. (a) MS2[WT] T=3 icosahedral capsid is shown (PDB ID: 2MS2). C/C and A/B dimer conformers are shown in red and green/blue, respectively. (b) MS2[S37P] T=1 icosahedral capsid (PDB ID: 4ZOR). All dimers in this capsid geometry are equivalent. Assembly capsomers are outlined in black to show constituent symmetry interfaces of each particle geometry.

By generating a 1-dimensional AFL of MS2[S37P] we were able to uncover unprecedented insight into the underlying principles of assembled structure mutability — we find that core residues and residues at the dimer interface of both VLP geometries share similar mutabilities on average. On the other hand, residues mediating interface between dimer subunits or at multiple contact interfaces are significantly more mutable in the T=1 variant. Interestingly, this holds true for residues both at the quasi-6-fold symmetry interface, which only occurs in the wildtype VLP, and the 5-fold symmetry interface, which is present in both capsid phenotypes. We then use the comprehensive variant data to assess computational methods for determining changes to the folding free energy of a protein when point mutants are introduced, and find that correlation between the predicted folding $\Delta\Delta G$ and the experimental effect on assembly viability is relatively low. Lastly, we combine computational prediction results with experimental fitness landscape results to train machine learning models of assembly state classification (Figure 2). We find that minimal computational and experimental input may be used to generate well-performing classifiers for mutational effects on VLP assembly state.

Figure 2: Workflow for generation of capsid apparent fitness landscapes and subsequent assembly classifier generation.

Results

Deep Mutational Scan of a Non-Native Virus-like Particle Assembly

In order to explore the effect of quaternary structure on the mutable landscape of self-assembling VLPs, we generated a deep mutational scanning library of the MS2 coat protein with a fixed S37P backbone point mutation. This backbone, MS2[S37P], shifts the MS2 VLP assembly from a 27 nm, T=3 icosahedral geometry to a 17 nm, T=1 icosahedral assembly.¹⁶ The SyMAPS platform was used to construct triplicate libraries of the MS2[S37P], as well as the wildtype (MS2[WT]) backbone. Next-generation sequencing of the mutagenized plasmid libraries resulted in excellent coverage of the mutational space, with 94% and 96% of possible variants accounted for in the MS2[S37P] and MS2[WT] libraries respectively. In both cases the majority of variants missing from the library were at the first and last position of the coat protein, which is the result of the first 5 base pairs of the Illumina PE300 MiSeq reads falling below our quality score threshold. The mutagenized libraries were expressed and subjected to size-exclusion chromatography to distinguish between assembling T=1 capsids, assembling T=3 capsids, and non-assembling variants. Assembled capsids encapsidate their encoding mRNA, thus the genotype of each well-formed particle was retained for sequencing. A high-temperature challenge was also introduced to remove capsids with compromised thermostability. Following coat protein expression and capsid enrichment, 72% of MS2[S37P] variants and 84% of MS2[WT] variants were recovered via sequencing and assigned an Apparent Fitness Score (AFS) derived from the $\log_{10}$ change in DNA read abundance between the input plasmid library and assembled capsid library. Correlation of AFS between replicates of the MS2[WT] was relatively high ($r^2 = 0.70 - 0.83$), indicating strong reproducibility and minimal distortions from biological noise (Figure 3a). Interreplicate correlation in the MS2[S37P] libraries was somewhat lower ($r^2 = 0.32 - 0.56$), possibly due to a lower signal-to-noise ratio, as T=1 capsids encapsidate less genetic material

than the wildtype capsid (Figure 3b).²⁸ There was a low degree of correlation between the MS2[WT] and MS2[S37P] libraries, meaning that there are both shared mutability preferences and divergences in mutability between the two capsid geometries. A notable difference in AFS distribution of the libraries is derived from their tolerance to deleterious mutations. A large portion of poorly performing variants in the MS2[WT] library ($AFS = -2.0 - 0.0$) drop to the lowest possible fitness score in the MS2[S37P] library ($AFS = -4$) (Figure 3c). This suggests that the MS2[WT] capsid may retain some level of assembly competency in response to deleterious mutations, while the MS2[S37P] capsid assembly is completely ablated in such cases. This may however also result from lower mRNA packaging per T=1 capsid, thus dropping deleterious mutations below the detection limit.

Figure 3: Capsid apparent fitness landscape replicate correlation. Correlations between the three biological replicates of the (a) MS2[WT] and (b) MS2[S37P] libraries are shown. (c) Correlation between the mean apparent fitness scores of each variant in the MS2[S37P] and MS2[WT] libraries. Kernel density estimations of apparent fitness score distributions are shown outside of the axes.

To investigate how the mutational landscape of the T=1 assembly phenotype differs from the wildtype T=3 icosahedron further, Shannon Entropy calculations were performed for each position of the capsid monomer backbone.²⁹ This measure of diversity at a given residue has previously been used to generate a Mutability Index (MI) of each backbone position,²³ which may be used to determine engineering "hotspots" in the capsid coat protein

as well as conserved residues likely to mediate key interactions for successful protein folding and particle self-assembly. The difference in MI between the MS2[WT] and MS2[S37P] capsid monomers is shown in (Figure 4a). The majority of the coat protein backbone shows a relatively low difference in MI between the wildtype and S37P libraries, indicating that many of the essential folding and assembly interactions are preserved. Regions of defined secondary structure elements, such as the central portion of the alpha helix spanning residues 105-111, show low mutability in both the T=1 and T=3 capsids. Likewise, the well-studied FG loop region^{30,31} (residues 71-76) of the coat protein retains high mutability in both capsid assemblies. However, individual positions with a significant shift in MI (< -0.2 or > 0.2) span the length of the backbone. The difference in mean fitness at each backbone is shown in (Figure 4b-c). Regions of where the mean AFS of MS2[WT] is notably higher than MS2[S37P] are prevalent, while positions with a higher mean AFS in MS2[S37P] are relatively rare. Mapping of MI to the capsid structures reveals the variety of secondary structural motifs with shifts in mutability between the two capsid phenotypes — residues in beta sheets, alpha helices, turns, and disordered loops show unique preferences in the MS2[S37P] capsid (Figure 4d).

Surprisingly, it is much more common that positions become more permissive to mutations in the T=1 capsid than become more strongly conserved, which is reflected in a higher mean MI for the MS2[S37P] than MS2[WT] (-0.24 and -0.33, respectively). This may seem counterintuitive, given that the mean AFS of the MS2[S37P] library was lower than that of the MS2[WT] library (-1.14 and -0.92, respectively). However, the lower mean AFS of the MS2[S37P] library results from the abundance of mutants at the minimum -4.0 AFS.

Disaggregation of AFS distribution by structural context (core, interface, or surface) reveals that the capsid AFLs differ most at subunit interfaces (Figure 5a). The MS2[WT] capsid shows a hierarchy whereby core residues are most conserved, followed by subunit interface residues, with surface residues (at either the interior or exterior surface) being most permissive to mutation. These preferences are in line with previous trends determined

Figure 4: Global mutability trends in MS2 capsid assemblies. (a) Per-residue differences in the mutability index between MS2[WT] and MS2[S37P] capsid backbones. (b) Mean apparent fitness scores (AFS) of all mutations at a given backbone position for both mutagenic libraries. (c) Differences in mean AFS for both capsid libraries. (d) Differences in mutability index mapped to the MS2[S37P] capsid (PDB ID: 4ZOR) as well as its constituent dimer using ChimeraX.³² In all cases blue represents a higher value in MS2[S37P] and red represents a higher value in MS2[WT].

from multiple sequence alignment of T=3 VLPs.³³ MS2[S37P] capsids also possess a highly conserved core; however the mutability indices of the interface and surface residues are statistically similar. Closer examination of contact interface context reveals that interdimer contact residues have similar mutabilities in both MS2[WT] and MS2[S37P]. Meanwhile, residues along intradimer contacts or residues that participate in multiple interface contacts are more permissive to mutation in the MS2[S37P] capsid library (Figure 5b).

Figure 5: Capsid mutability trends based on structural context. (a) Boxplots of MS2[WT] and MS2[S37P] apparent fitness score (AFS) distributions by residue context. (b) Boxplot of AFS distributions by interface context, further separated in (c) by symmetry contact. MS2[WT] and MS2[S37P] distributions are shown in red and blue, respectively. *** indicates $p < 0.001$ and **** indicates $p < 0.0001$.

The symmetry elements of MS2[WT] and MS2[S37P] were next explored to understand better differences in the AFL of both capsid phenotypes. Casper and Klug's theory of quasi-

equivalence dictates that T=3 capsids will form from a mixture of pentameric and hexameric subunits at 5-fold and quasi-6-fold axes of symmetry, while T=1 icosahedral capsids form only from pentamer subunits at 5-fold symmetry axes.³⁴ Prior work has also isolated two key intermediates in the MS2[WT] assembly pathway: a 12mer subunit at the quasi-6-fold symmetry axis (6CP2) and a 20mer subunit at the 5-fold symmetry axis (10CP2).³⁵ While assembly intermediates of MS2[S37P] have not been directly observed, the 10CP2 subunit is a viable assembly intermediate, while the 6CP2 subunit presumably is not due to the geometry of the T=1 capsid. As such, we hypothesized that residues involved in the pentameric interfaces of the MS2 capsid would be similarly conserved in MS2[WT] and MS2[S37P], while the hexameric interfaces would be more flexible in MS2[S37P] as they do not mediate productive intermediates in the assembly pathway. Surprisingly, we find that both hexameric and pentameric interface residues are significantly more flexible in MS2[S37P]. This may suggest that the assembly pathway of the T=1 capsid may be more reliable than the T=3 capsid once initial folding and dimerization has occurred. There are indeed more potential misassembly pathways in T=3 icosahedra due to their larger number of unique interfaces,³³ although the relationship between the number of improper assembly pathways and the mutational flexibility of VLPs has not been previously established. These trends and observations demonstrate how comprehensive fitness landscapes may reveal mutability preferences that challenge expectations from sequence conservation in nature or known assembly pathways.

Benchmarking Computational Residue Scanning Methods for Variant Fitness Prediction

We used these datasets to benchmark computational tools that calculate predicted changes in folding free energy of protein variants. *In silico* deep mutational scans of the MS2[WT] and MS2[S37P] coat proteins were performed using BioLuminate,³⁶ the DynaMut2 web server³⁷ and the PoPMuSiC 2.1 web server.³⁸ As modeling the monomer subunit only accounts for one potential mode of assembly failure (protein misfolding), computational scans of the dimeric,

6CP2, and 10CP2 subunits of the assembled capsid were performed to achieve better coverage of the assembly deficits related to changes in intra-dimer and inter-dimer affinity. The 6CP2 capsomer was not modeled for MS2[S37P] as it is not likely to be formed in the assembly pathway. Correlation between computationally-predicted $\Delta\Delta G$ of coat protein variants and experimentally determined AFS was low across all tested methods and modeled subunits. This suggests that calculated fitness values of the capsid assemblies depends on a more complex set of factors than are captured by the endpoint Gibbs free energy of its constituent subunits.

Although the absolute AFS is not well correlated with computational $\Delta\Delta G$ scoring, we hypothesized that these methods may be successful in a binary classification of variant assembly competency. Previously-established AFS thresholds for assembly were used to label variant phenotypes (AFS > 0.2 = “assembling”, AFS < -0.2 = “non-assembling”). Predictions were classified as “assembling” if $\Delta\Delta G \leq 0$ (i.e. computation predicts the mutation is neutral or stabilizing) and “non-assembling” if $\Delta\Delta G > 0$ (i.e. computation predicts the mutant is destabilizing). Binary classification results are shown in Figure 6. The accuracy of the tested classification methods was modest, ranging from 0.62-0.73. There were no large differences in accuracy between the three computational scanning methods. Gratifyingly, models of assembly subunits improved classification accuracy over models of the coat protein monomer for both MS2[WT] and MS2[S37P] in all three computational methods. Accuracy can, however, provide an overly-optimistic measure of performance in cases where the tested dataset has a class imbalance (i.e. there are many more non-assembling mutations than assembling mutations).³⁹ Thus, the Matthews correlation coefficient (MCC) was reported for each classification case, as this metric is a more reliable measure for classification of imbalanced datasets.^{40,41} MCC was calculated with values derived from Table 1 using equation 1. MCC ranges from -1 to +1, where MCC = +1 represents a perfect classifier, MCC = -1 represents perfect misclassification, and MCC = 0 represents the expected value of a coin toss classification.

Table 1: Confusion matrix describing performance of computational scanning classification.

		Fitness Landscape	
		Assembling	Non-Assembling
Computational Scan	Assembling	TP	FP
	Non-Assembling	FN	TN

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP) \times (TP + TN) \times (TN + FP) \times (TN + FN)}} \quad (1)$$

Evaluation by MCC reveals that the physics-based prediction via BioLuminate scored much higher than the other two methods (Figure 6b). While the overall accuracy of each method was relatively similar, BioLuminate correctly predicted the minority class (i.e., “assembling” mutation) in more cases than DynaMut2 and PoPMuSiC 2.1. Subsequent tuning efforts focused on using the predicted Δ Stability from BioLuminate, since predicting assembly-competent mutations is more valuable for the production of capsids with engineered properties or new sequence motifs. In order to improve classifier performance, the $\Delta\Delta G$ threshold for designation of a mutation as “assembling” or “non-assembling” was scanned to optimize the MCC scores (Figure 6c). Threshold scanning revealed that a higher $\Delta\Delta G$ cutoff ($\Delta\Delta G = +5-8$ kcal/mol) resulted in a higher MCC score for capsid phenotype classification. This indicates that MS2 capsid assembly can readily occur in spite of moderate increases to the folding free energy of the capsid subunits. Though tuning the cutoff threshold improved model performance, we sought to improve predictive power further by employing machine learning to capitalize on the rich feature set derived from these computational methods.

Figure 6: Computational residue scanning classification performance. The (a) accuracies and (b) Matthews correlation coefficients of three computational residue scanning methods are shown for various subunit models of the MS2[WT] and MS2[S37P] capsid. Predicted $\Delta\Delta G$ values from BioLuminate's Δ Stability parameter were compared to experimental apparent fitness scores. (c) Matthews correlation coefficients are shown as a function of classification $\Delta\Delta G$ threshold.

Development of Supervised Learning Models for Capsid Assembly Classification

A total of 106 features for the MS2[WT] set and 81 features for the MS2[S37P] set were prepared for each amino acid variant represented in the experimental fitness landscapes of each capsid. The description of each feature is provided in the Supporting Information. Initial models were prepared via four supervised learning algorithms (K-Nearest Neighbor,⁴² logistic regression,⁴³ random forest,⁴⁴ and XGBoost⁴⁵) These models were trained with default hyperparameters using a stratified 75/25 train/test split of the landscape data. Models were evaluated by 10-fold cross validation, and binary classifier performance was measured using the MCC scores (Figure 7). The random forest and XGBoost models outperformed the K-nearest neighbor and logistic regression models in both training sets, as measured by MCC. Bayesian Optimization was subsequently implemented to tune the hyperparameters used for XGBoost or random forest model training.⁴⁶ The resultant tuned models showed strong predictive performance on the unseen test set of landscape data, with the MS2[WT] model slightly outperforming the MS2[S37P] model (accuracy = 0.86, MCC = 0.58 and accuracy = 0.82, MCC = 0.52, respectively).

Figure 7: Initial machine learning model classification performance. Four machine learning algorithms were benchmarked for predicting the assembly state of (a) MS2[WT] or (b) MS2[S37P] capsid variants. Boxplots of the resultant Matthews correlation coefficients from 10-fold cross validation of models are shown.

Ensemble tree-based algorithms, such as XGBoost and random forest, furthermore enable calculation of the importance of each feature with respect to prediction of the target variable, thus offering interpretability of the final MS2[WT] and MS2[S37P] models (Figure 8).^{47,48} Across all four models, the most important feature for prediction was the Δ Stability of the 10CP2 subunit calculated using Bioluminate. Many *de novo* features, such as residue solvent-accessible surface area, hydropathy, potential energy, and internal energy, contributed significantly to the final models. In the MS2[WT] models, Δ Stability of the 6CP2 subunit was

Figure 8: Final model performance and feature importance. Performance of XGBoost and random forest models of (a,b) MS2[WT] and (c,d) MS2[S37P] assembly classifications are shown. The top 10 most informative features for classification are shown. Model-derived and *de novo* features are shown in red and blue, respectively.

much less important than the 10CP2 subunit, which may indicate either that the endpoint stability of the 6CP2 complex does not capture inhibition of the subunit assembly pathway, or that this complex is more resilient to perturbation than the 10CP2 complex. Interestingly, no intersubunit affinity terms ranked highly in importance, while intradimer affinity ranked in the top ten of 2 out of 4 of the final models.

We next sought to establish the minimal input data set necessary for model training, as reduction of the training set size or initial computational modeling steps may enable accurate construction of assembled capsid fitness landscapes with reduced experimental and computational cost. Training features were reduced to include only *de novo* amino acid features as well as the most predictive feature, Δ Stability derived from the 10CP2 Bioluminate modeling. A random forest model was trained on randomly-generated subsets of the fitness landscape (1-2000 training samples) and evaluated by predicting the remaining landscape assembly states. Learning curves of both models displaying the gain in MCC and accuracy as a function of training samples are shown in Figure 9. Interestingly, the model achieves reasonable accuracy after very few samples, and only marginally improves as the sample numbers increase. Viewing MCC gain, however, displays a sharp improvement that plateaus after roughly 500 samples. This suggests that small sublibraries of self-assembled protein capsids may enable reasonably accurate predictions of capsid fitness landscapes, likely by learning that the majority of mutations inhibit capsid assembly. To obtain predictions that correctly classify assembly-competent mutants, a moderate sampling of the full mutational space of the VLP (about 20% of variants) will be needed.

Figure 9: Determination of minimized input model performance. (a) MCC gain plot and (b) accuracy gain plot of both capsid models as functions of training set size. Models were generated using only Δ Stability BL 10CP2 and *de novo* parameters for training. Random subsets of the fitness landscape data were used to simulate low-sample training information. MS2[WT] and MS2[S37P] learning curves are shown in red and blue, respectively.

Conclusion

Here, we have characterized the apparent fitness landscape of a non-native (T=1 icosahedral) VLP assembly geometry. Comparison to the apparent fitness landscape of the wild type (T=3 icosahedral) particle yielded insight into the influences quaternary structure has on the mutability of a self-assembling material. Residues at the core of the monomer subunit, as well as residues mediating the dimer interfaces of the capsid, showed similar mutability preferences in both capsid libraries. Surprisingly, residues along both the five-fold and quasi-six-fold symmetry interfaces of the particle were significantly more mutable in the T=1 capsid than the wildtype assembly, suggesting that the T=1 capsid may be a more versatile platform for engineering efforts. Further, the deep mutational scanning data produced were used to benchmark the performance of various computational residue scanning tools for estimating the effects of missense mutations on protein folding energetics. While correlations between calculated $\Delta\Delta G$ values and measured assembly competency were relatively low, modeling known assembly intermediates using BioLuminate's MM/GBSA method offered the best overall performance. Finally, machine learning models were trained using *de novo* structural features in tandem with computational modeling data to produce predictive assembly state classifiers for both capsid assemblies. The relationship between input training size and classifier performance demonstrates that relatively small input datasets can be used to produce predictive classifiers for self-assembly with strong performance. As such, this study may represent a generalizable strategy for semi-rational engineering and systematic mutation of proteinaceous self-assembled nanomaterials.

Experimental

Strains

All library experiments were conducted with MegaX DH10B *E. coli* electrocompetent cells (ThermoFisher Scientific, Cat# C640003). Overnight growths from a single colony were incubated for 18 h at 37 °C with 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 with shaking at 200 RPM.

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.

HPLC SEC

Variants or libraries were analyzed on an Agilent 1290 Infinity HPLC with an Agilent Bio SEC-5 column (5 μ m, 2000A, 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.

Library generation

Deep mutational scanning libraries of both capsids were performed as previously described via SyMAPS.²³ SyMAPS uses the EMPIRIC cloning developed in the Bolon lab.⁴⁹ 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. Five previously-generated entry vectors each designed to accept a 26 codon region of the MS2 coat protein as an insert were used to prepare both the MS2[WT] and MS2[S37P] libraries. Four of the entry vectors bearing the position S37 codon were mutated to S37P via Quikchange mutagenesis for the MS2[S37P] library. Single-stranded DNA primers spanning each 26-codon region with an NNK codon at each position were pooled and diluted to a final concentration of 50 ng/μL. Touchdown PCR was used to amplify the mutagenic inserts and the double-stranded DNA was purified by PCR clean-up (Promega, Cat# A9282). Cleaned dsDNA inserts were diluted to 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 electrocompetent DH10B E. coli cells (Invitrogen, Cat# 18290015). Following electroporation and recovery, cells were plated onto LB-A containing 32 μg/mL chloramphenicol and grown overnight at 37 °C. Dilutions (100-fold) of each replicate were plated and counted to ensure a minimum of 3x colony-forming units relative to the number of possible variants.

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 °C. 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 T=3 or T=1

VLPs. Collected fractions were then subjected to 50 °C thermostability challenges in a H₂O bath 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 6.5 min for MS2[WT] or 8.0 min for MS2[S37P]) were combined, subjected to RNA extraction, barcoded, and identified via high-throughput sequencing.

High-throughput sequencing sample preparation

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.⁵⁰ 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, random hexamers primer).

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. Libraries were quantified with Qubit and combined in equal molar ratio. Libraries were analyzed by PE300 MiSeq in collaboration with the UC Davis Sequencing Facilities.

High-throughput sequencing data processing and analysis

Data were trimmed and processed described previously,²³ with minor variation. Briefly, data were trimmed with Trimmomatic⁵¹ 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) with a maximum overlap of 300 base pairs. Reads were sorted and indexed with SAMtools⁵² and unmapped reads were filtered with the Picard function CleanSam. Reads were processed with in-house code to produce various AFLs.

AFL calculations

Cleaned and filtered high-throughput sequencing reads were analyzed using in-house Python code (available at <https://github.com/dan-brauer/MS2-Assembly-Classification>). Briefly, the start codon and stop codons on either end of the MS2 gene were parsed and the correct gene length was verified. For MS2[S37P], codon encoding for the S37P mutation was also verified. Codons were then translated into amino acids to generate counts for each tripeptide combination

Calculations were repeated for all replicate plasmid pools and assembled selections. Relative abundances were calculated similarly to the previously described protocol.²⁶ Briefly, the relative abundance of a selection was calculated by dividing the abundance generated from an assembled VLP by the plasmid abundance from its respective biological replicate.

The mean relative abundance across three biological replicates was calculated for the final fitness landscape of each library. 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. The log₁₀ relative abundance value for an individual variant in a particular challenge is referred to as its Apparent Fitness Score (AFS).

Computational fitness scanning

Assembly subunits (monomer, dimer, 6CP2, and 10CP2) were constructed in the Maestro molecular modeling environment based on the crystal structures of MS2[WT] (PDB:2MS2) and MS2[S37P] (PDB:4ZOR). Resultant structures were preprocessed and minimized using

the OPLS3e forcefield. Computational scanning for minimized subunit structures via DynaMut2 and PoPMuSiC 2.1 was performed using their respective web servers. BioLuminate calculations were performed symmetrically across chains using side-chain prediction with backbone minimization at a 0.00 Å cutoff. Classifier performance was evaluated using in-house Python code (available at <https://github.com/dan-brauer/MS2-Assembly-Classification>).

Machine learning model training

Machine learning models were trained and assessed using the Tidymodels⁵³ framework in R (available at <https://github.com/dan-brauer/MS2-Assembly-Classification>). Continuous training variables were normalized, nominal variables (e.g. amino acid identity) were encoded as binary dummy variables, and zero variance variables were filtered in data preprocessing. Algorithm comparisons (K-Nearest Neighbor, logistic regression, random forest, and XGBoost) were performed using default hyperparameters. Initial random forest and XGBoost model training used 1000 trees. Cross validation (10-fold) was performed on a 75/25 split of the fitness landscaping data. Random forest and XGBoost models were further optimized through 100 rounds of Bayesian hyperparameter tuning, with 10 initial parameter sets and an early stop after 20 consecutive iterations with no improvement. Optimized parameters were used to train final models which were evaluated on the 25% of data held out for final evaluation. Model feature importance was determined using the R ‘vip’ package.⁵⁴

Acknowledgement

This work was funded in part by the National Science Foundation (grants CBET-2043973 to DTE and CBET-2044011 to MBF). The authors thank Dr. Kathleen Durkin at the UC Berkeley Molecular Graphics and Computation Facility for assistance in running BioLuminate computational Scanning through Maestro. The sequencing was carried out at the DNA Technologies and Expression Analysis Cores at the UC Davis Genome Center, supported by

NIH Shared Instrumentation Grant 1S10OD010786-01.

References

- (1) Glasgow, J.; Tullman-Ercek, D. Production and applications of engineered viral capsids. *Applied Microbiology and Biotechnology* **2014**, *98*, 5847–5858.
- (2) Molino, N. M.; Wang, S.-W. Caged protein nanoparticles for drug delivery. *Current Opinion in Biotechnology* **2014**, *28*, 75–82.
- (3) Garg, H.; Mehmetoglu-Gurbuz, T.; Joshi, A. Virus Like Particles (VLP) as multivalent vaccine candidate against Chikungunya, Japanese Encephalitis, Yellow Fever and Zika Virus. *Scientific Reports* **2020**, *10*, 4017.
- (4) Nooraei, S.; Bahrulolum, H.; Hoseini, Z. S.; Katalani, C.; Hajizade, A.; Easton, A. J.; Ahmadian, G. Virus-like particles: preparation, immunogenicity and their roles as nanovaccines and drug nanocarriers. *Journal of Nanobiotechnology* **2021**, *19*, 59.
- (5) Rohovie, M. J.; Nagasawa, M.; Swartz, J. R. Virus-like particles: Next-generation nanoparticles for targeted therapeutic delivery. *Bioengineering & Translational Medicine* **2017**, *2*, 43–57.
- (6) Wilkerson, J. W.; Yang, S.-O.; Funk, P. J.; Stanley, S. K.; Bundy, B. C. Nanoreactors: Strategies to encapsulate enzyme biocatalysts in virus-like particles. *New Biotechnology* **2018**, *44*, 59–63.
- (7) Lázaro, E.; Arribas, M.; Cabanillas, L.; Román, I.; Acosta, E. Evolutionary adaptation of an RNA bacteriophage to the simultaneous increase in the within-host and extracellular temperatures. *Scientific Reports* **2018**, *8*, 8080.
- (8) Tetter, S.; Terasaka, N.; Steinauer, A.; Bingham, R. J.; Clark, S.; Scott, A. J. P.; Patel, N.; Leibundgut, M.; Wroblewski, E.; Ban, N.; Stockley, P. G.; Twarock, R.; Hil-

- vert, D. Evolution of a virus-like architecture and packaging mechanism in a repurposed bacterial protein. *Science* **2021**, *372*, 1220–1224.
- (9) Khmelinskaia, A.; Wargacki, A.; King, N. P. Structure-based design of novel polyhedral protein nanomaterials. *Current Opinion in Microbiology* **2021**, *61*, 51–57.
- (10) Wargacki, A. J.; Wörner, T. P.; van de Waterbeemd, M.; Ellis, D.; Heck, A. J. R.; King, N. P. Complete and cooperative in vitro assembly of computationally designed self-assembling protein nanomaterials. *Nature Communications* **2021**, *12*, 883.
- (11) Hsia, Y. et al. Design of multi-scale protein complexes by hierarchical building block fusion. *Nature Communications* **2021**, *12*, 2294.
- (12) Laniado, J.; Cannon, K. A.; Miller, J. E.; Sawaya, M. R.; McNamara, D. E.; Yeates, T. O. Geometric Lessons and Design Strategies for Nanoscale Protein Cages. *ACS Nano* **2021**, *15*, 4277–4286.
- (13) Laniado, J.; Meador, K.; Yeates, T. O. A fragment-based protein interface design algorithm for symmetric assemblies. *Protein engineering, design & selection: PEDS* **2021**, *34*, gzab008.
- (14) 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 (New York, N.Y.)* **1992**, *255*, 178–183.
- (15) Cordonnier, A.; Montagnier, L.; Emerman, M. Single amino-acid changes in HIV envelope affect viral tropism and receptor binding. *Nature* **1989**, *340*, 571–574.
- (16) 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 Letters* **2016**, *16*, 5944–5950.

- (17) Araya, C. L.; Fowler, D. M. Deep mutational scanning: assessing protein function on a massive scale. *Trends in biotechnology* **2011**, *29*, 435–442.
- (18) Starita, L. M.; Fields, S. Deep Mutational Scanning: A Highly Parallel Method to Measure the Effects of Mutation on Protein Function. *Cold Spring Harbor Protocols* **2015**, *2015*.
- (19) 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* **2014**, *11*, 124.
- (20) Ferguson, A. L.; Mann, J. K.; Omarjee, S.; Ndung'u, T.; Walker, B. D.; Chakraborty, A. K. Translating HIV sequences into quantitative fitness landscapes predicts viral vulnerabilities for rational immunogen design. *Immunity* **2013**, *38*, 606–617.
- (21) Visher, E.; Whitefield, S. E.; McCrone, J. T.; Fitzsimmons, W.; Llaure, A. S. The Mutational Robustness of Influenza A Virus. *PLOS Pathogens* **2016**, *12*.
- (22) Acevedo, A.; Brodsky, L.; Andino, R. Mutational and fitness landscapes of an RNA virus revealed through population sequencing. *Nature* **2014**, *505*, 686–690.
- (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. *Nature Communications* **2018**, *9*, 1385.
- (24) Qi, H.; Wu, N. C.; Du, Y.; Wu, T.-T.; Sun, R. High-resolution genetic profile of viral genomes: why it matters. *Current Opinion in Virology* **2015**, *14*, 62–70.
- (25) 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* **2019**, *58*, 1527–1538.

- (26) Brauer, D. D.; Hartman, E. C.; Bader, D. L. V.; Merz, Z. N.; Tullman-Ercek, D.; Francis, M. B. Systematic Engineering of a Protein Nanocage for High-Yield, Site-Specific Modification. *Journal of the American Chemical Society* **2019**, *141*, 3875–3884.
- (27) Robinson, S. A.; Hartman, E. C.; Ikwuagwu, B. C.; Francis, M. B.; Tullman-Ercek, D. Engineering a Virus-like Particle to Display Peptide Insertions Using an Apparent Fitness Landscape. *Biomacromolecules* **2020**, *21*, 4194–4204.
- (28) Kafader, J. O. et al. Multiplexed mass spectrometry of individual ions improves measurement of proteoforms and their complexes. *Nature Methods* **2020**, *17*, 391–394.
- (29) 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. *Molecular Immunology* **1997**, *34*, 1067–1082.
- (30) Stonehouse, N. J.; Vålegård, 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. *Journal of Molecular Biology* **1996**, *256*, 330–339.
- (31) 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. *Nature Communications* **2016**, *7*, 12524.
- (32) Pettersen, E. F.; Goddard, T. D.; Huang, C. C.; Meng, E. C.; Couch, G. S.; Croll, T. I.; Morris, J. H.; Ferrin, T. E. UCSF ChimeraX: Structure visualization for researchers, educators, and developers. *Protein Science: A Publication of the Protein Society* **2021**, *30*, 70–82.
- (33) Bahadur, R. P.; Janin, J. Residue conservation in viral capsid assembly. *Proteins: Structure, Function, and Bioinformatics* **2008**, *71*, 407–414.

- (34) Caspar, D. L. D.; Klug, A. Physical Principles in the Construction of Regular Viruses. *Cold Spring Harbor Symposia on Quantitative Biology* **1962**, *27*, 1–24.
- (35) Knapman, T. W.; Morton, V. L.; Stonehouse, N. J.; Stockley, P. G.; Ashcroft, A. E. Determining the topology of virus assembly intermediates using ion mobility spectrometry–mass spectrometry. *Rapid communications in mass spectrometry : RCM* **2010**, *24*, 3033–3042.
- (36) Beard, H.; Cholleti, A.; Pearlman, D.; Sherman, W.; Loving, K. A. Applying Physics-Based Scoring to Calculate Free Energies of Binding for Single Amino Acid Mutations in Protein-Protein Complexes. *PLOS ONE* **2013**, *8*.
- (37) Rodrigues, C. H.; Pires, D. E.; Ascher, D. B. DynaMut2: Assessing changes in stability and flexibility upon single and multiple point missense mutations. *Protein Science* **2021**, *30*, 60–69.
- (38) Dehouck, Y.; Kwasigroch, J. M.; Gilis, D.; Rooman, M. PoPMuSiC 2.1: a web server for the estimation of protein stability changes upon mutation and sequence optimality. *BMC Bioinformatics* **2011**, *12*, 151.
- (39) Galar, M.; Fernandez, A.; Barrenechea, E.; Bustince, H.; Herrera, F. A Review on Ensembles for the Class Imbalance Problem: Bagging-, Boosting-, and Hybrid-Based Approaches. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)* **2012**, *42*, 463–484.
- (40) Chicco, D.; Jurman, G. The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. *BMC Genomics* **2020**, *21*, 6.
- (41) Boughorbel, S.; Jarray, F.; El-Anbari, M. Optimal classifier for imbalanced data using Matthews Correlation Coefficient metric. *PLOS ONE* **2017**, *12*.

- (42) Altman, N. S. An Introduction to Kernel and Nearest-Neighbor Nonparametric Regression. *The American Statistician* **1992**, *46*, 175–185.
- (43) Walker, S., H.; Duncan, D. B. Estimation of the probability of an event as a function of several independent variables. *Biometrika* **1967**, *54*, 167–179.
- (44) Breiman, L. Random Forests. *Machine Learning* **2001**, *45*, 5–32.
- (45) Chen, T.; Guestrin, C. XGBoost: A Scalable Tree Boosting System. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York, NY, USA, 2016; pp 785–794.
- (46) Snoek, J.; Larochelle, H.; Adams, R. P. Practical Bayesian Optimization of Machine Learning Algorithms. *Advances in Neural Information Processing Systems*. 2012.
- (47) Strobl, C.; Boulesteix, A.-L.; Kneib, T.; Augustin, T.; Zeileis, A. Conditional variable importance for random forests. *BMC Bioinformatics* **2008**, *9*, 307.
- (48) Auret, L.; Aldrich, C. Empirical comparison of tree ensemble variable importance measures. *Chemometrics and Intelligent Laboratory Systems* **2011**, *105*, 157–170.
- (49) Hietpas, R. T.; Jensen, J. D.; Bolon, D. N. A. Experimental illumination of a fitness landscape. *Proceedings of the National Academy of Sciences* **2011**, *108*, 7896–7901.
- (50) Pinto, F. L.; Thapper, A.; Sontheim, W.; Lindblad, P. Analysis of current and alternative phenol based RNA extraction methodologies for cyanobacteria. *BMC Molecular Biology* **2009**, *10*, 79.
- (51) Bolger, A. M.; Lohse, M.; Usadel, B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **2014**, *30*, 2114–2120.
- (52) Li, H.; Handsaker, B.; Wysoker, A.; Fennell, T.; Ruan, J.; Homer, N.; Marth, G.; Abecasis, G.; Durbin, R.; 1000 Genome Project Data Processing Subgroup, The Sequence

Alignment/Map format and SAMtools. *Bioinformatics (Oxford, England)* **2009**, *25*, 2078–2079.

- (53) Kuhn, M.; Wickham, H. *Tidymodels: a collection of packages for modeling and machine learning using tidyverse principles.*; 2020.
- (54) Greenwell, B. M.; Boehmke, B. C. Variable Importance Plots—An Introduction to the vip Package. *The R Journal* **2020**, *12*, 343–366.