Role of Signal Transduction Domains in Histidine Kinase Evolution and Activity

by

Bruk Mensa

DISSERTATION

Submitted in partial satisfaction of the requirements for degree of

DOCTOR OF PHILOSOPHY

in

Chemistry and Chemical Biology

in the

GRADUATE DIVISION

of the

UNIVERSITY OF CALIFORNIA, SAN FRANCISCO

Approved:

DocuSigned by:

William Degrado

582727E949C7441...

William Degrado

Chair

DocuSigned by:

Jason E. Gestwicki

[illegible]4E5...

Jason E. Gestwicki

DocuSigned by:

Andrej Sali

8F7A6AB94F2C4F4...

Andrej Sali

Committee Members

Acknowledgements

I'd like to briefly thank the following people for their various contributions towards this thesis and my education at large. First and foremost, I'd like to thank my PI, Dr. William F DeGrado for his research ideas, mentorship, guidance, as well as consistent support, both professional and personal. Thank you for consistently elevating the quality of my work and ideas. I'd also like to acknowledge Dr. Gregory Caputo (Rowan university), my first lab mentor, for instilling confidence in my abilities and acumen early in my research career. I'm grateful to the members of my thesis committee, Drs. Jason Gestwicki and Andrej Sali, for incredibly useful discussions, insights, ideas and enthusiastic support of my various projects. I'd like to thank Drs. Kathleen Molnar, Nicholas Polizzi, Nathan Schmidt, Thomas Lemmin, Manasi Bhate and Bobo Dang, for their contribution to the genesis and refinement of various ideas presented in this thesis, as well as direct contributions to work produced during the course of my graduate career (Go team bacteria!). I'd also like to thank Drs. Iain C. Clark and Adam Abate for the opportunity to work with them on multiple projects, including one presented in chapter 3 of this thesis. I must recognize the people involved in my rotation projects (Drs. Charly Craik, Natalia Sevillano, Kathrin Zuberbühler, John Gross, Serena Sanulli) for their dutiful instruction and the useful skills that have contributed to my thesis. I also want to thank Drs. Mark Kelly, Yibing Wu and Manasi Bhate (NMR); Drs. James Holton, George Meigs (ALS 8.3.1 beamline) and Jessica Thomaston (Crystallography); Drs. Thomas Weiss and Tsutomu Matsui (SAXS, SSRL BL 4-2); and Dr. Gozde Ülas (CW EPR) for guiding me through various structural biology experiments. Julia Molla, Nicole Flowers, Miranda Paulos, Lili Li, Alma Agorilla (HR) and Kevin Long (ISSO) were instrumental in helping me navigate the various pitfalls of undergoing a PhD education as an international student. Finally, I'd like to

acknowledge members of the DeGrado lab, past and present as well as the students, faculty and staff of the UCSF CCB graduate program for all their support, ideas, discussions, contributions, encouragement and enthusiasm during this entire process. I look forward to working with you all in the future and attentively following your distinguished careers.

Contributions

Chapter 1 is a direct reproduction of a manuscript that has been recently submitted to Elife under the title, “An equilibrium allosteric coupling model for histidine kinase signaling”. The contributing authors to this work are Nicholas Polizzi, Kathleen M Molnar, Andrew Natale and Thomas Lemmin. I was responsible for the refinement of the central allosteric coupling model presented in the chapter, along with exploring several alternative models and hypotheses. I also doubled the available experimental dataset for model-fitting, supplementing the set with hand-picked mutations and combinations of mutations that would allow the fit to converge (successfully). I extended the findings of our model to other E. coli HKs and was responsible for all data-analysis and representation. I credit Drs. Kathleen Molnar and Manasi Bhate for the genesis of the allosteric coupling model that was explored, refined and experimentally fit in this work, which was first presented in Bhate MP, Molnar KS, Goulian M, DeGrado WF. Signal transduction in histidine kinases: insights from new structures. *Structure.* 2015 Jun 2;23(6):981-94. Kathleen Molnar and Andrew Natale also contributed experimental data towards the fitting of said allosteric model in Chapter 1. Dr. Nicholas Polizzi wrote the script utilized for data-fitting of our allosteric model and assessment of confidence intervals, and conducted iterative fitting as I refined the model, datasets and parameters. Dr. Thomas Lemmin generated the MD model of PhoQ in explicit membrane that was used for visualizations as well as structure-guided hypothesis generation. Both provided short methods for their respective contributions. Along with ideas, input and guidance throughout the process from my PI, Dr. William F DeGrado, I am otherwise responsible for the writing of Chapter 1.

Chapter 2 is a collaborative effort with Drs. Iain C. Clark (Abate lab, UCSF) and Yibing Wu. I was responsible for experimental and library design, library cloning and transformation. Dr. Clark conducted the FACS library sorting experiment. I also conducted the subsequent sequencing library building and data processing/analysis of the library, as well as writing, interpretation and data representation in Chapter 2. Dr. Wu is conducting Deep Learning analysis of the resulting data and contributed Figure 41, and details for Methods section. As always, my PI Dr. William F DeGrado provided ideas, input and guidance for the work presented.

Chapter 3 is a direct reproduction of a recently published manuscript, Clark IC, Mensa B, Ochs CJ, Schmidt NW, Mravic M, Quintana FJ, DeGrado WF, Abate AR. Protein design-scapes generated by microfluidic DNA assembly elucidate domain coupling in the bacterial histidine kinase CpxA. Proc Natl Acad Sci U S A. 2021 Mar 23;118(12):e2017719118. An introductory abstract and "contributions" section highlighting relevance to the topic of this thesis and my personal contribution respectively has been appended to the beginning of chapter 3.

Role of Signal Transduction Domains in Histidine Kinase Evolution and Activity

Bruk Mensa

Abstract

The process by which various upstream sensor and signal-transduction domains of bacterial histidine kinases (HKs) modulate the activity of the conserved autokinase domain remains poorly understood. Specifically, why do most HKs contain modularly inserted signal transduction domains? How do HKs robustly evolve and finetune the coupling between stimulus sensor domains and the conserved autokinase domain, which are often separated by 10s of nanometers? What is the role of these intervening domains in fine-tuning signaling parameters such as the minimum/maximum responsiveness, mid-point, and steepness of signal transition of an HK? In this work, we examine signal transduction through model *E. coli* HKs, PhoQ and CpxA, which contain one of the most abundant signal transduction domains in HKs, the HAMP domain. We first generate a large set of single-point mutants of PhoQ, and simultaneously measure the signaling state of the ligand-binding sensor and the kinase activity of the autokinase *in vitro,* at several inducing ligand concentrations to assess the coupling between these two domains. We demonstrate that point mutants in the HAMP signal transduction domain significantly modulate the coupled behavior of the sensor and autokinase, producing markedly varied ligand-dependent responses. We further use the insertion of poly-glycine motifs (Gly_7) to decouple domains from one another and qualitatively show that, intrinsically, the sensor domain has a drastically poor ligand-dependent state transition propensity, and similarly, the autokinase domain has a drastically high basal kinase activity. The HAMP domain strongly couples to both domains and is sufficient to adjust these propensities to what is observed in the full length PhoQ. We suggest

that signal transduction in PhoQ occurs by an allosteric coupling mechanism, in which the HAMP domain strongly couples to and acts in opposition the underlying signaling state equilibria of PhoQ such that it is maximally responsive to physiologically relevant ranges of stimuli. We demonstrate the same phenomenon in two other E. coli HAMP containing HKs, CpxA and BaeS, and suggest this may be a common theme in the evolution of signal transduction domains in HKs. In order to quantitatively examine the feasibility of modulating various ligand-dependent properties that inform HK function through evolution, we next establish and experimentally fit a three-domain, two-state equilibrium allosteric signaling model. We demonstrate that small changes to the HAMP domain sequence allow for robust modulation of the signaling ensemble and provide quantitative measures for the strong modulation of both sensor and autokinase domains by the HAMP, as well as the effects of point-mutations and Gly_7 insertions.

We more fully examine the ability of the HAMP to couple strongly and influence the sensor and autokinase domains of PhoQ by introducing a large library of variants in the HAMP four-helix bundle hydrophobic core, as well as the junction between the HAMP and autokinase domains (the S-Helix) and selecting for variants with high PhoQ activity. We find that destabilizing the HAMP four-helix bundle hydrophobic core does indeed lead to higher kinase activity. Furthermore, we find that the wild-type S-Helix sequence is enriched in the high-activity population, along with sequences with comparable polarity or poor helical propensity. Taken together, these observations lend credence to the hypothesis that the thermodynamically preferred signaling state of the HAMP behaves as a negative allosteric regulator of the autokinase, and that this regulation is alleviated by destabilizing the core helical bundle structure as well as the alpha-helical motif that connects it to the autokinase. We investigate this

relationship further using a deep learning method to establish sequence-activity predictive relationships and extract structural features that are essential for this behavior. Finally, we examine the question of whether the HAMP domain exists in two distinct structural states, or rather conformational ensembles that can be classified into one of two functional states. We examine signaling through the HAMP domain of an E. coli histidine kinase, CpxA, by constructing a small library of structurally diverse inputs into the HAMP domain and evaluate the resulting autokinase activity as a function of several S-helix point mutations. This analysis allows us to discern the relationship between different signal inputs into the HAMP domain as the linkage to the output domain (autokinase) is varied. We find that the HAMP seems to have a multiconformational landscape that is not explained by 2 unique structural conformations.

In this thesis, we show that the insertion of signal transduction domains in HKs can significantly alter both the intrinsic behaviors of sensor and autokinase domains, as well as the coupling between them. These properties can be well-described through a coupled two-state allosteric mechanism, and easily finetuned through simple mutations to the signal transduction domain and its linkage to adjacent domains to achieve the desired physiologically relevant activity profile.

Table of contents

List of figures

List of tables

Abbreviations

ADP – Adenosine diphosphate

AMP – Ampicillin

ATP – Adenosine triphosphate

bMe – *b*- Mercaptoethanol, reducing agent

BSA – Bovine Serum Albumin

DNA- Deoxyribonucleic acid

DTT – Dithiothreitol, reducing agent

ECL substrate –

EDTA – Ethylenediamine tetraacetic acid, divalent ion chelator

FACS – Fluorescent Assisted Cell Sorting

GAF – globular domain named after classes of proteins it was first identified in: cGMP-specific phosphodiesterases, Adenylyl cyclases and FhlA

GFP – green fluorescent protein

Gly$_7$ – An insertion of 7 Glycines in a row.

HAMP – Histidine kinases, Adenylate cyclases, Methyl transferases, Phosphatases, parallel four-helix bundle signal transduction element

HAMP Gly$_7$ – An insertion of 7 glycines between the HAMP and autokinase domains of a histidine kinase

HK – Histidine Kinase

HRP – Horse Raddish Peroxidase

ILVF – Isoleucine, Leucine, Valine, Phenylalanine; name of library in which these amino acids were introduced into variable residues

ILVFSPTA - Isoleucine, Leucine, Valine, Phenylalanine, Serine, Proline, Threonine, Alanine; name of library in which these amino acids were introduced into variable residues

IPTG – Isopropyl β-D-1-thiogalactopyranoside, lac promoter inducer

KAN – Kanamycin

KO – knock out

LacZ – *b*-Galactosidase gene

LB – Luria Bertani media (lysogeny broth)

LDS – Lithium dodecyl sulfate, detergent

mCherry – cherry red fluorescent protein

MCS – Multiple Cloning Site

MOPS - 3-(N-morpholino)propanesulfonic acid, buffer

NEM – N-Ethylmaleimide, cystine alkylating agent

OD – Optical Density

ONPG – *o*-Nitrophenyl-β-galactoside, colorimetric beta-galactosidase substrate

PAS – globular domain named after the first proteins it was identified in: period clock protein (PER), vertebrate aryl hydrocarbon receptor nuclear translocator (ARNT), and Drosophila single-minded protein (SIM)

pcpxP – promoter of the E. coli cpxP gene

PCR – Polymerase Chain Reaction

pmgrB – promoter of the E. coli mgrB gene pmgtA – promoter of the E. coli mgtA gene

pspy – promoter of the E. coli spy gene

RR – Response Regulator

S-Helix – Signaling Helix, a conserved polar coiled-coil motif with deviations from idealized coil-coil geometry, found in various classes of proteins including histidine kinases

SD – Standard deviation

SDS – Sodium dodecyl sulfate, detergent

SOC – Super Optimal broth with Catabolite repression

STE – Standard error

TBS-t – TRIS Buffered Saline with Tween-20

TM - Transmembrane

TM Gly$_7$ – An insertion of 7 Glycines between the transmembrane and HAMP domains of a histidine kinase

TRIS – Tris(Hydroxymethyl)aminomethane, buffer

YFP – yellow fluorescent protein

Z-buffer – beta-galactosidase buffer

Chapter 1: Examining the role of signal transduction domains in Histidine Kinase function and evolution

Introduction

Sensor Histidine Kinases (HKs) are a conserved signaling module in bacteria responsible for sensing a myriad of environmental stimuli and orchestrating transcriptional responses along with their cognate transcription factors (Response Regulator, RR) (1, 2). These sensors are implicated generally in environment sensing, and have been implicated in multi-drug resistance (3–5) and as master regulators of virulence programing in pathogenic bacteria (6, 7). While a lack of a full-length HK structure has hampered our understanding of the mechanism of signal transduction in these proteins, cytoplasmic domain structures have shed light particularly on the enzymatic core of this class of kinases. Several recent crystallographic snapshots of the autokinase domains of multiple HKs in various conformations (8–14), particularly CpxA and DesK, have shown distinct conformations involved in autophosphorylation, phosphotransfer and dephosphorylation that may be conserved across this family. While these works offer a conserved view of the catalytic cycle of the autokinase domain (15), the question of how these proteins couple a sensory event often several nanometers away to the modulation of this underlying activity remains unanswered. This question is especially perplexing in light of the various modular architectures of HKs, involving the insertion of one or more signal transduction domain between sensors and the conserved autokinase domain. It is abundantly clear that the same conserved autokinase domain that defines this protein class can be regulated by a myriad of structural inputs, ranging from short alpha-helical dimeric coiled coils, to well-folded tertiary folds, such as HAMP, PAS and GAF domains (16, 17) (**Figure 1.1**). Moreover, it is clear from the representation of these folds in diverse protein classes that these domains evolved independently of HKs and were co-opted pervasively into functioning HK architectures. Therefore, they are likely to serve a generalizable function that is robust to evolutionary selection, and the construction of physiologically relevant sensors optimally positioned to respond to environmental changes.

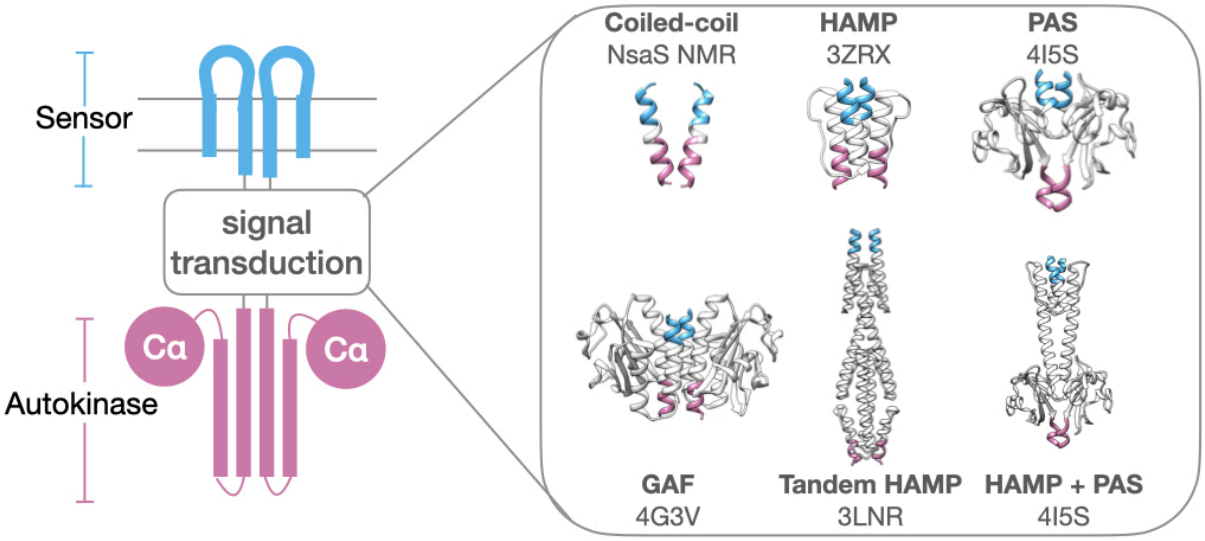


Figure 1.1 - Modular architecture of histidine kinases. Various protein folds and numbers of signal transduction domains are found inserted between sensor (blue) and autokinase (purple), including simple coiled-coils (NsaS), HAMP (AF1503), PAS (VicK), GAF (Nlh2), Tandem HAMP (Aer2) and HAMP/PAS domain (VicK). Pdb codes are provided in figure, except for NsaS (NMR structure).

In this work, we create and evaluate the activity of a large set of single-point mutants and poly-glycine insertions in a model Gram-negative HK, PhoQ(18), whose sensor and autokinase domains are separated by an intervening HAMP signal transduction domain. The PhoPQ two-component system is composed of a canonical transmembrane sensor HK that senses the presence of divalent cations (19, 20), and polycationic species such as antimicrobial peptides (21, 22), and a cognate response regulator, PhoP (18, 23), which controls regulons pertinent to cation transport and outer-membrane remodeling (24–33). The electrostatic repulsion between an acidic patch in the sensor domain and the negatively charged bacterial inner membrane in the absence of cations, and the bridging of these electrostatic interactions in their presence, is hypothesized to be the mechanism by which the sensor functions (34, 35). The kinase activity of PhoQ is repressed by divalent cation binding, whereas it is enhanced by the presence of antimicrobial peptides. PhoQ is additionally implicated in low pH sensing (36) via an interaction with the membrane protein UgtL (37), and has recently been suggested to respond to changes in osmolarity (38), although the mechanism is unclear. In previous work, our group generated distance restraints from disulfide scanning of the alpha-helical dimeric interface of the sensor and transmembrane domains of PhoQ. Using Bayesian inference, a two-state sensor conformational equilibrium was determined as the most parsimonious description of the population ensemble (39). Thus, PhoQ serves as an ideal system for investigating how the changes in the two-state conformational equilibrium of the sensor are propagated to modulate the kinase-phosphatase output of the autokinase domain.

We create a set of 35 different PhoQ constructs, containing either single point-mutations or poly-glycine insertions away from the ligand-binding site of the sensor and the autokinase domain, and simultaneously evaluate the sensor signaling-state and autokinase activity at 5 different concentrations of Mg^{2+}. This allows us to thoroughly interrogate signal coupling between the sensor and autokinase, and how it may be modulated by the intervening HAMP domain. We show that small changes to the HAMP domain, such as point mutations, can produce large changes in the coupling between sensor and autokinase. Furthermore, we use poly-glycine insertions between domains to uncouple the sensor and autokinase domains from the HAMP and show that the intrinsic behavior of these domains is considerably different from that in the fully coupled HK. Specifically, the sensor domain of PhoQ has a drastically reduced ability to transition to a Mg^{2+} bound signaling state when uncoupled from the HAMP. Similarly, the autokinase domain has high basal kinase activity when uncoupled from the HAMP. Coupling to the HAMP domain by itself is sufficient for restoring the behavior of the sensor and autokinase to that observed in the fully coupled HK. These experimental observations suggest that the HAMP domain significantly alters the underlying equilibria of the sensor and autokinase domains such that the fully coupled HK is maximally responsive to a physiologically relevant range of stimuli. The modular insertion of the HAMP allows for a robust modulation of PhoQ activity without changing the ligand-binding or enzymatic functions of the HK, providing a possible explanation for the ubiquity of similarly inserted signal-transduction domains in HK evolution.

Finally, we develop an equilibrium model to demonstrate how the HAMP domain robustly modulates PhoQ signaling and allows for the facile evolution of different ligand-dependent behaviors. This model describes PhoQ as an allosterically connected three-domain protein, in which each domain samples a two-state equilibrium. We use our single-point and poly-glycine insertion experimental data to mathematically fit this allosteric coupling model and demonstrate how the thermodynamically favored state of the HAMP couples strongly to both sensor and autokinase domains in the phosphatase state, producing an overall bistable sensor with a greater affinity for its ligand and lowered kinase activity. Changes in the intrinsic equilibrium of the HAMP, and its coupling strength adjacent domains dynamically adjusts the basal and maximal responsiveness of HK.

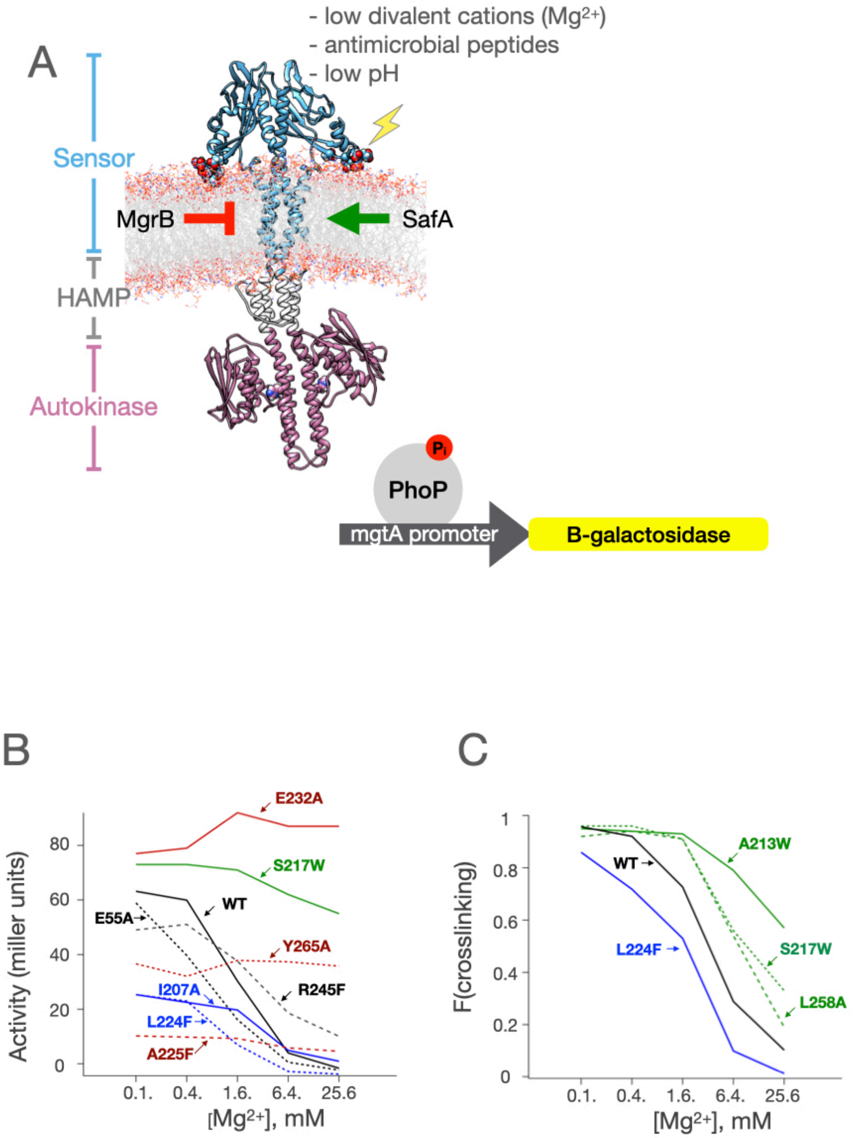


Figure 1.2 - PhoQ structure and function. (A) Molecular dynamics model of PhoQ highlighting the sensor (blue), signal transduction HAMP (white) and autokinase (purple) domains. Acidic patch residues involved in stimulus sensing in the sensor domain are shown in spheres. The catalytic residue (His$_{277}$) is shown in spheres in the autokinase domain. Upon phosphorylation by PhoQ, the response regulator, PhoP (grey) can modulate the activity of its cognate promoters, which can be monitored using a beta-galactosidase reporter fusion. **(B)** PhoQ autokinase has high reporter activity at low divalent ion concentrations, which is turned off by high concentrations of Mg^{2+}. Different point mutations of PhoQ produce various Mg^{2+} dependent response, including 'locked' phenotypes (red), enhanced kinase (green), enhanced phosphatase (blue) and wild type-like (black). **(C)** PhoQ sensor domain has Y60C high cross-linking efficiency at low [Mg^{2+}], which is reduced by Mg^{2+} binding. Different point mutants can shift Mg^{2+} dependent behavior to more (blue) or less (green) sensitive to ligand-based state transition.

Results

Concerted vs. Allosteric signal transduction mechanism

Binding of Mg^{2+} to the sensor domain of PhoQ results in the turning off of kinase activity in the autokinase in a ligand concentration dependent manner. In one limiting regime, signal transduction in PhoQ can be described in terms of a concerted structural rearrangement of the entire dimer that results from structural changes in the sensor domain due to ligand binding. In this regime, any structural changes in the sensor are *necessarily* propagated to the remainder of the HK, obligatorily tying the response behavior of the two domains. On the other end of this spectrum, signal transduction can occur not as concerted changes over the whole HK, but rather the mutual "coupling" of underlying two- or multi-state equilibria in adjacent, distinct domains. In this scenario, ligand binding would alter the conformational ensemble of the sensor, which in turn couples to the energetics of the conformational ensemble of the adjacent domain, the HAMP, with varying efficiencies. This in turn couples the intrinsic equilibrium of the autokinase, again with varying efficiencies. The modulation of the strength of these couplings would dictate the relative transition in activity of the autokinase in response to the ligand-driven transition of the sensor. Indeed, a concerted signaling mechanism is an extreme manifestation of an "allosteric coupling" model, in which a high degree of coupling results in, for all intents and purposes, a single conformational transition throughout the protein. If we are able to simultaneously observe the signaling states of the sensor and autokinase domains at various concentrations of ligand, we can assess where PhoQ lies in this spectrum of allosteric coupling.

Evaluating the activity of PhoQ sensor and autokinase domains

In order to effectively study the coupling in structural transitions between the sensor and catalytic domains of PhoQ, we needed to establish robust experimental methods to evaluate these 2 domains. The activity of the autokinase of PhoQ goes from a high kinase activity state which phosphorylates the cognate response regulator PhoP at low Mg^{2+} ('K' state) to a low kinase/high phosphatase activity state which dephosphorylates PhoP at high Mg^{2+} ('P' state). The phosphorylation of PhoP was measured using an established reporter gene assay, in which a beta-galactosidase gene was appended immediately downstream of a PhoP regulated promoter (p*mgtA*, a magnesium transporter). Phosphorylated PhoP drives higher transcription of the reporter gene from the mgtA promoter, and serves as a proxy for the overall kinase activity of the PhoQ autokinase.

The binding of ligand to the periplasmic domain is difficult to evaluate experimentally due to the complex binding mode of divalent cations in between the negatively charged surface of the bacterial inner membrane and the acidic patch residues of PhoQ. Fortunately, in the course of previous work (39), we identified multiple residues in the periplasmic domain of PhoQ that when mutated to cysteine showed Mg^{2+} dependent crosslinking efficiency. In particular, the mutant Y60C showed near-100% crosslinking efficiency at low Mg^{2+} concentrations and no detectible crosslinking at high Mg^{2+} concentrations. Therefore, Cys-60 crosslinking was used as a means of quantifying the two-state population partition of the PhoQ sensor domain. The ligand free 'K' state of the sensor has high Cys-60 crosslinking efficiency (and corresponds to high kinase activity), and the ligand-bound 'P' state of the sensor has low Cys60 crosslinking efficiency (and corresponds to low kinase/high phosphatase activity), and these two states can easily be quantified by western blot.

Equipped with a means of evaluating the functional states of both the sensor and autokinase domains of PhoQ, we are able to address the fidelity with which changes in the sensor-state are propagated to the autokinase. Initial comparison of the Mg^{2+} dependent activity of PhoQ suggest a strong coupling between the sensor- and autokinase- functional states, as seen in the similar Mg^{2+}-dependent transitions between states for both domains in 'wild type' (Y60C) PhoQ (**Figure 1.2B-C**). However, we can modulate the basal and maximal activity of phoQ at high and low Mg^{2+} as well as the midpoint of transition between the K and P states of the autokinase through the introduction of various single point-mutations within the HAMP domain or at the sensor-HAMP interface (**Figure 1.2B**). Since these mutations are not affecting the ligand binding surface of the PhoQ sensor or the catalytic domains of the autokinase, they must be modulating the extent to which the sensor is coupled to autokinase response. Similarly, we see that mutations in the HAMP can alter the favorability of the Mg^{2+} binding dependent transition of the sensor from the K to the P state without altering the ligand binding surface itself (**Figure 1.2C**). The range of possible behaviors strongly suggests that signal propagation occurs by an allosteric coupling mechanism in which the HAMP can buffer or potentiate the ligand binding-dependent K to P transition of PhoQ in a tunable manner.

Disrupting coupling between domains in PhoQ

Having established a seemingly strongly allosterically coupled wild-type HK, we sought to examine the effect of purposefully uncoupling the sensor from the autokinase at various domain-domain junctions. Since the sensor and HAMP domains are connected by a dimeric alpha helical coiled-coil between transmembrane helix 2 of the sensor and helix 1 of the HAMP, we inserted a 7-Glycine sequence which has very poor helicity at this junction (between PhoQ res. 219 and 220, "Sensor-HAMP Gly_7"). This allows us to observe the intrinsic ligand-dependent transition of the sensor domain when structurally decoupled from the HAMP and autokinase of PhoQ. We find that the Gly_7 insertion strongly destabilizes the 'P' state of the sensor, as indicated by its high cys-60 crosslinking even at high [Mg^{2+}] (**Figure 1.3A**). We then migrate the Gly_7 motif to the HAMP-autokinase junction (between PhoQ res. 260 and 261, "HAMP-autokinase Gly_7"), thereby restoring the alpha-helical coupling between the sensor and the HAMP. This is sufficient to restore the Mg^{2+} binding dependent transition of the sensor to the 'P' state, as indicated by the low Cys-60 crosslinking efficiency at high [Mg^{2+}] (**Figure 1.3A**). This suggests that the sensor domain is intrinsically very stable in the 'K' state, to the point that it becomes Mg^{2+} insensitive. The HAMP structurally couples to act in opposition to this preferred sensor 'K' state sufficiently strongly such that it can now become sensitive to Mg^{2+} again.

We next evaluated the effect of these two Gly_7 insertions on the activity of the autokinase domain. The HAMP-autokinase Gly_7 insertion disrupts the helical Signaling Helix ('S-Helix') motif that couples the HAMP to the autokinase, again due to poor helicity. When the autokinase is uncoupled in this manner, it has a high, ligand independent kinase activity, as indicated by high reporter gene transcription at both low and high [Mg^{2+}] (**Figure 1.3B**). This indicates that the autokinase prefers to be in the 'K' state when uncoupled from the sensor + HAMP. When the Gly_7 insertion is migrated to the sensor-HAMP junction, thereby restoring the alpha-helical coupling between the HAMP and the autokinase, the autokinase now has low kinase activity, as indicated by low reporter activity at both low and high [Mg^{2+}] (**Figure 1.3B**). It appears the HAMP domain again couples to act in opposition to the intrinsic activity of the adjacent autokinase domain.

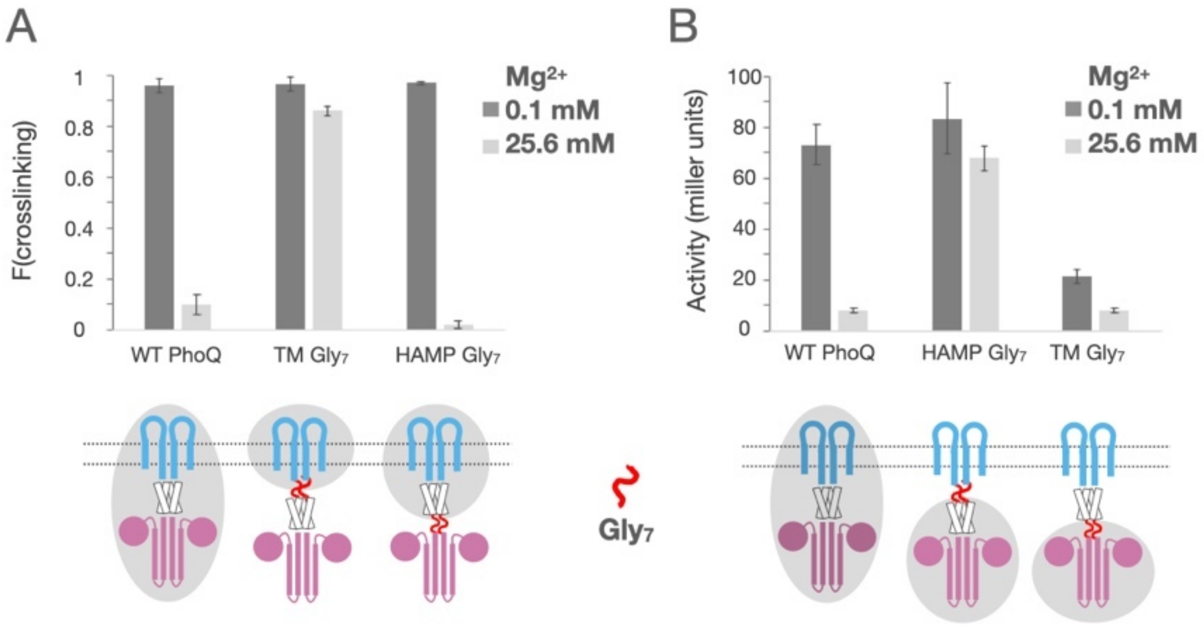


Figure 1.3 - Effect of disrupting allosteric coupling between domains. (A) Measured crosslinking of sensor at low and high Mg^{2+}. Sensor is Mg responsive in wild type PhoQ. The sensor in isolation remains at high crosslinking efficiency even at high [Mg^{2+}], but the addition of the HAMP domain back to the sensor is sufficient to restore wild type Mg^{2+} responsiveness. **(B)** Measured kinase activity at low and high [Mg^{2+}]. Kinase is responsive to Mg^{2+} in wild type PhoQ. The Kinase domain in isolation has high kinase activity, which is repressed by the addition of the HAMP domain. Error bars are ±standard deviation for indicated biological replicates.

Structural perspective of HAMP coupling to sensor and autokinase

The HAMP domain is a well-folded globular domain that has been co-opted in the many classes of proteins, for which it is named (40), rather than a structural motif that has repeatedly convergently evolved in various histidine kinases. In the few multidomain structures and high-resolution models of histidine kinases containing a HAMP, the sensor-HAMP junction features disruptions to the coiled-coil geometries of the region (39, 41–43). Previous work from our lab has shown that the crosslinking efficiency of cysteine mutants in the TM2 helices of PhoQ show a canonical coiled coil heptad repeat pattern of crosslinking that is disrupted at the TM helix2-HAMP helix1 junction, presumably due to local disorder or splaying apart of the coiled coil (39). Earlier work has also suggested the presence of a polar hydrated hemi-channel in the c-terminal end of the PhoQ TM bundle, coordinated by a buried polar residue in the TM bundle (Asn_{202}) (44). Furthermore, structures of sensory rhodopsin II transducing complex (HtrII) also show potential hydration of the c-terminal end of the TM bundle (45). Lastly, the strong sequence conservation of a proline residue in the transmembrane-HAMP junction (46, 47), as well as poly-glycine motifs similarly situated in chemotaxis proteins (43, 48, 49) suggest a local disruption of the helical junction. Indeed, the recently determined multi-domain structure of NarQ in apo and holo states shows a rigid body hinge-like motion about a conserved proline between the sensor-TM and HAMP domains (50). Similarly, the junction between the HAMP and autokinase domains is a conserved motif called the Signaling Helix (S-Helix), which is characterized by an apparent insertion of 1, 3 or 4 residues into the otherwise canonical coiled coil dimeric interface (51). This motif is found to be often rearranged structurally to allow for the symmetric-asymmetric transitions observed in several auto-kinase structures.

Allosteric coupling mechanism for signal transduction in PhoQ

Our data indicate that the sensor and autokinase domains of PhoQ have intrinsic equilibria that strongly favor the ‘K’ state but are destabilized by unfavorable coupling to the HAMP domain. This presumably results in a strained full-length protein that can sufficiently transition to the ‘P’ state through ligand binding. In order to further evaluate the effect of the HAMP domain in the bistability of PhoQ as a whole, we established a simple allosteric model for signal transduction, in which PhoQ is represented as 3 domains: the “sensor” comprising of the periplasmic and transmembrane domains, the “HAMP” signal transduction domain, and the catalytic “autokinase” domain, with each domain allowed to sample a two-state equilibrium **(Figure 1.4)**. Each domain equilibrium is then energetically coupled to the adjacent domain with varying efficiencies. The choice of a three-domain, two-state mechanism is informed both by experimental evidence for two-state equilibria in many HK domain structures, as well as the granularity of the question posed in this work. By definition, some change must occur in each domain for signal propagation, thus requiring a minimum of two signaling states in each domain, while additional states can be allowed. From a functional standpoint, a two-state definition offers intuitive classifications of what could be multiple structural states in a signaling ensemble; a sensor domain can be described as an apo ‘K’ state or holo ‘P’ state as discussed above, and similarly, a kinase domain can be described as high kinase-activity ‘K’ state or low kinase high phosphatase-activity ‘P’ state. The two functional states of the PhoQ sensor in particular map over well to the two-state structural equilibrium of the sensor that have been elucidated by Bayesian inference modeling of disulfide crosslinking efficiencies of the periplasmic and TM2 bundles (39). The autokinase domain of HKs has several distinct structural states and has been elucidated in more than 2 distinct conformations, but many of these states are part of the enzymatic cycle necessary to generate one of two functional outcomes: a phosphorylated response regulator, or a dephosphorylated response regulator. This provides a function based two-state definition, in which ensembles that promotes ATP binding, autophosphorylation and phosphotransfer can be classified as 'K' state, and ensembles that promote the dephosphorylation of the response regulator as 'P' state. Our model is further evaluating the coupling of the sensor and autokinase domains to the signal transduction HAMP domain and how this allows for the modulation of the 'K' to 'P' state transition of PhoQ on an ensemble level. Since the HAMP domain doesn't have a direct experimentally determined output for its signaling states similar to the 'K and 'P' states of the sensor and autokinase, we simply refer to the two signaling states as 'H_1' and 'H_2'.

With these definitions, each of the 3 domains of PhoQ is allowed to sample two signaling states within the full-length protein with an intrinsic equilibrium constant, K_i, and a proportionality of difficulty in this structural transition based on the signaling state of neighboring domains, α_i (**Figure 1.4B**). The sensor interconverts between the high-Mg^{2+} affinity 'P' state, and a low-Mg^{2+} affinity 'K' state with an equilibrium constant K_1, and similarly, the autokinase transitions between the low-kinase/ high phosphatase activity 'P' state and the high kinase-activity 'K' state with an equilibrium constant of K_3. The intervening HAMP domain can interconvert between signaling states H_1 and H_2, with an associated equilibrium constant, K_2. α_1 describes the relative strength in coupling between the sensor and the two states of the HAMP, and α_2 the difference in coupling strengths between the autokinase and the two states of the HAMP. Finally, we allow Mg^{2+} to bind to the two sensor states, 'P' and 'K', with two affinities (K_{dP},

K_{dK}) and alter the phoQ signaling ensemble state. Altogether, this scheme defines an equilibrium population of 8 distinct conformational states, each with a Mg^{2+}-bound and Mg^{2+}-free form, which can be partitioned according to a phenotype of interest (sensor crosslinking competency or PhoP phosphorylation/transcriptional activity) **(Figure 1.4B)**. Four of these states have correlated sensor and autokinase activities (both kinase or phosphatase, K-H_1-K, K-H_2-K, P-H_1-P, P-H_2-P), whereas the remaining four allow for uncoupled responses.

In order to overcome the over-parameterization of our model, which is necessitated by the nature of the question we are answering, we used several single-point mutations in the sensor and HAMP domains to perturb signaling in the protein in a variety of manners while staying away from residues directly involved in Mg^{2+} binding (acidic patch residues) and kinase activity (the entire catalytic domain) **(Figure 1.7)**. We also combined selected HAMP mutants with Gly_7 disconnections to further constrain the parameter space. All together, we provided 55 independent pairs of Mg^{2+}-dependent sensor-crosslinking and kinase-activity curves that were simultaneously determined from the same samples at 5 different [Mg^{2+}] for a total of 55 x 2 x 5 = 550 measurements of kinase activity and sensor crosslinking fractions, representing 35 types of unique sequence perturbations throughout the PhoQ sequence.

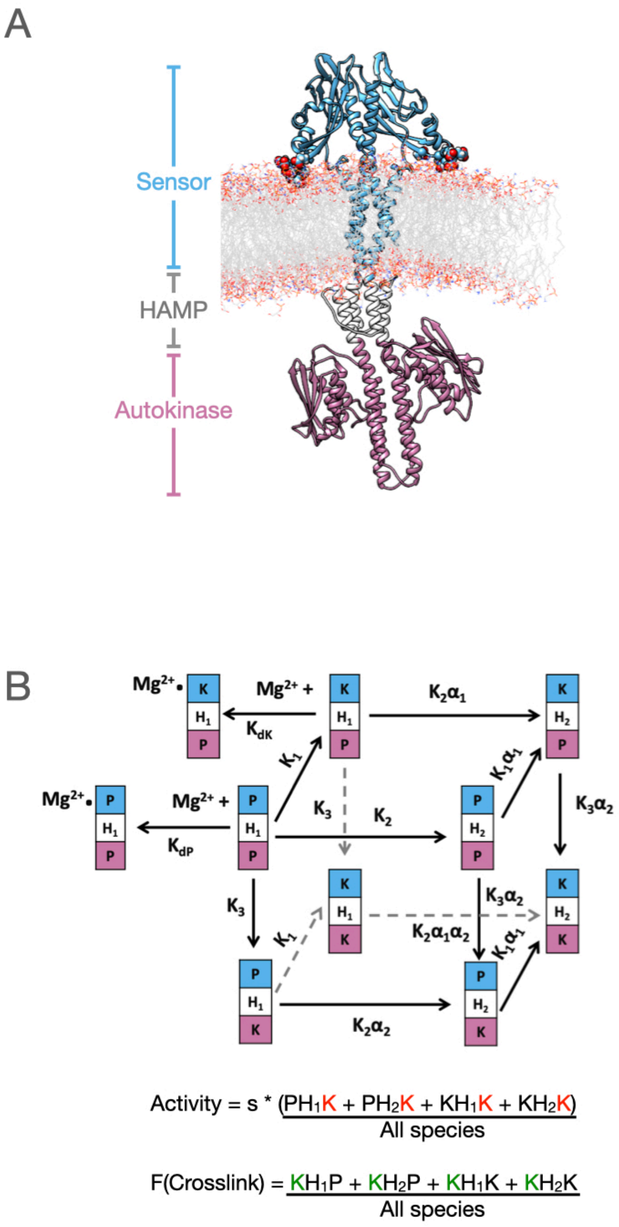


Figure 1.4 - Two-state allosteric coupling model for PhoQ signaling. (A) Model of PhoQ showing domain structure. "Sensor" (PhoQ$_{1\text{-}219}$, blue), "HAMP" (PhoQ$_{220\text{-}260}$, white) and "Autokinase" (PhoQ$_{261\text{-}494}$, purple) domains are allowed to exist in two-state equilibria and couple allosterically **(B)** Allosteric coupling scheme describing all possible transitions for a two-state, three-domain equilibrium. Each domain is allowed to sample a two-state equilibrium with equilibrium constants K_1, K_2, K_3 for sensor, HAMP and autokinase respectively. Neighboring domains are allosterically coupled with efficiencies α_1 (sensor-HAMP) and α_2 (HAMP-Autokinase). The sensor is allowed to bind Mg^{2+} in both signaling states with dissociation constants K_{dK} (low affinity, "kinase on" state) and K_{dP} (high affinity, "kinase off" state). Fraction of PhoQ with kinase in active state, or sensor in cross-linking competent state are calculated as shown and elaborated in the methods section.

Fitting crosslinking and activity of PhoQ and mutants

In order to fit our experimental data to our model, we partition our model populations according to the desired experimental outcomes; sensor crosslinking and autokinase activity. Populations with the sensor in the 'K' state represent high crosslinking efficiency populations, and similarly, populations with the autokinase in the 'K' state represent kinase-active populations (**Figure 1.4B**). These populations are summed up and divided by the total ensemble to yield [Mg^{2+}] dependent expressions of crosslinked and kinase-active fractions, as detailed in the methods section. Experimentally, we provide simultaneously determined sensor-crosslinking fraction and kinase activity data at 5 different concentrations of Mg^{2+} to fit using our model. In cases where replicate experimental data is available, replicates are treated independently in an effort to give proportional representation, and therefore weight, to such data in our global fitting.

The 7 parameters in our model, K_1, K_2, K_3, α_1, α_2, K_{dK} and K_{dP}, as well as a global V_{max} parameter for kinase activity, 'S', are fit to our 550 experimental observations, allowing all parameters to be constrained globally across all datasets, with parameters describing the effect of a point mutation allowed to float locally to account for its effect. Mutations in a domain that are away from domain-domain junctions are allowed to perturb the equilibrium constant of the domain they reside in (i.e. K_1, K_2, K_3). Mutations near domain-domain junctions are additionally allowed to perturb the relevant allosteric coupling constant (α_1, α_2). The remaining parameters are restrained globally across all datasets. The identity of parameters varied for each PhoQ construct is listed in **Figure 1.6**.

Results of allosteric coupling model fitting

Our fitting resulted in non-linear overall R^2 correlation of 0.88, with most datasets showing remarkably good fits, given the constrains of our fitting algorithm (see methods for details of fitting). The most stringent test for the quality of fitting is the fit of the "wild type" (Y60C) PhoQ activity and crosslinking data. Every parameter for these two curves is fit globally across all datasets, and is therefore constrained to satisfy the entire dataset, whereas point mutations/ Gly_7 disconnections allow for one or more parameters to be locally fit to account for the mutation itself. As shown in **Figure 1.5A**, we obtain good global fits for both the activity and crosslinking of wild type PhoQ. The fit for activity shows a shallower transition from high to low activity than the data, perhaps owing to the unknown stoichiometry and cooperativity of Mg^{2+} binding by the sensor, which we have simplified to 1 binding site per monomer in our model. Fits for all 35 point mutants and Gly_7 disconnections are shown in **Figure 1.8**. Our model shows good fits to a vast majority of our experimental inputs. Parameter fit quality and confidence intervals of parameter values is discussed in proceeding sections.

In general, the results of our fit are consistent with the observations made in **Figure 1.3** that a bistable full-length molecule results from strained connections through the HAMP. The sensor domain has a large equilibrium constant, $K_1 = 950$, reflecting the strong propensity of the domain to be in the high crosslinking 'K' state. However, due to the unfavorable coupling ($\alpha_1 = 0.003$) to the thermodynamically preferred 'H_2' state of the HAMP ($K_2 = 22$), the sensor equilibrium is ameliorated to a modest $K_{1.}\alpha_1 = 5.04$. Similarly, the autokinase has a propensity to be in the high-kinase activity 'K' state, with a moderate equilibrium of $S.K_3 = 1.02$. However, when coupled to the preferred 'H_2' state of the HAMP, this propensity is greatly reduced by a factor of

$\alpha_2 \leq 10^{-3}$. Further description of the results of this allosteric fit, including quality of fit, and the interpretation of parameter values is discussed at length in the proceeding sections.

In order to visualize the [Mg^{2+}]-dependent population transition of the population ensemble, we plotted the fraction of each of the 8 signaling states as a function of Mg^{2+}, as colored by fraction of crosslinking or kinase activity arising from each signaling state. All four crosslinking-competent species, $\underline{K}H_1P$, $\underline{K}H_1K$, $\underline{K}H_2P$, $\underline{K}H_2K$ are depleted due to the [Mg^{2+}] dependent transition of the sensor from 'K' to 'P' state, as expected, with almost all of the observed population ensemble crosslinking transition resulting from the transition of $\underline{KH_1}P$ to $\underline{PH_2}P$ states, with minor contributions from $\underline{K}H_2P$ to $\underline{P}H_2P$ transition (**Figure 1.5B**). The activity of the autokinase is also depleted due to Mg^{2+} binding as expected, coming almost entirely from the depletion of the KH_1K to PH_2P state (**Fig 1.5B**). The PH_2P state accumulates in all these transitions with increasing Mg^{2+} concentrations and is the primary species at high Mg^{2+} (93%), with minor accumulation of the PH_1P state (4%). Again, the relative abundance of these PH_1P and PH_2P states is dictated by the stability of the HAMP, $K_2 = 22$. **Figure 1.6** shows how select point mutations which affect HAMP coupling modulate this population ensemble transition. We can see that the basal and maximal activities of PhoQ can be modulated by leveraging the change in HAMP coupling due to mutation to adjust the abundance of population species that are normally minor contributors to observed sensor crosslinking and autokinase activity.

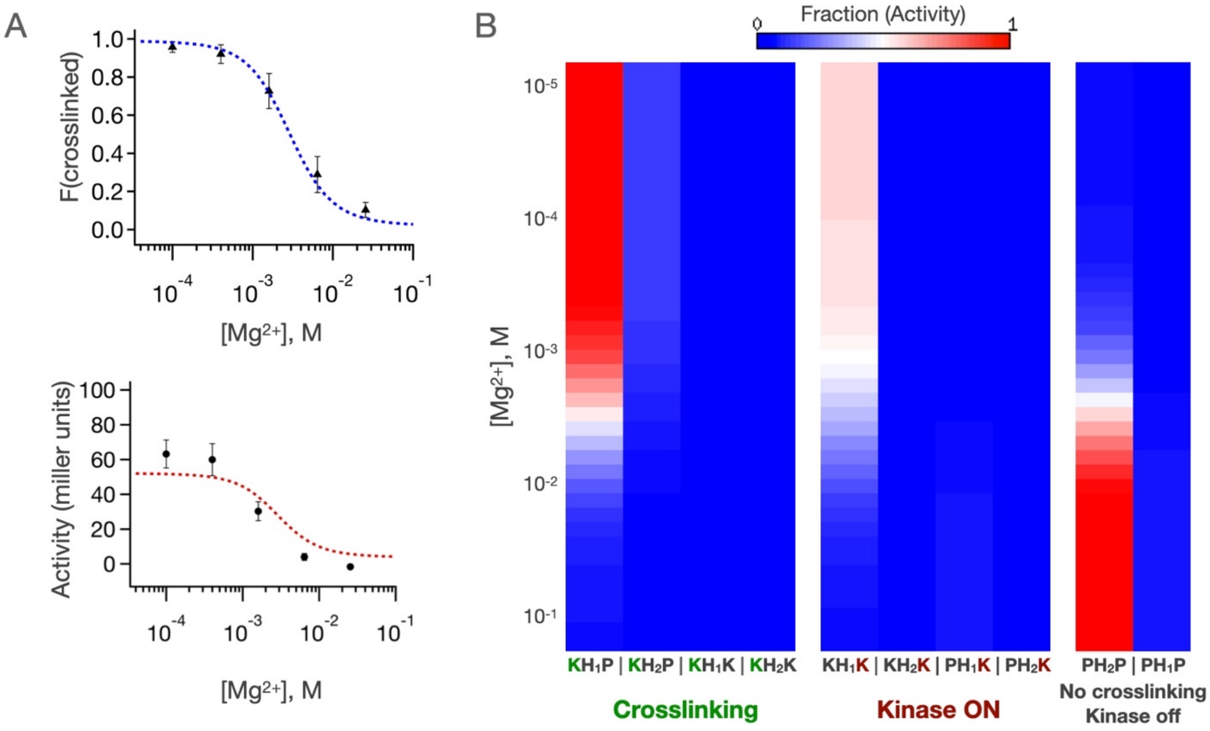


Figure 1.5 - Fitting of WT activity and crosslinking. (A) Total sensor crosslinking fraction (top) and kinase activity (bottom) as a function of [Mg^{2+}]. Fits are shown in dashed lines. Data error bars are +/-SD. **(B)** [Mg^{2+}] dependent contribution to total sensor crosslinking and autokinase activity for all 8 singling states of the PhoQ ensemble. Note that two species, KH_1K and KH_2K have both sensor crosslinking and kinase activity.

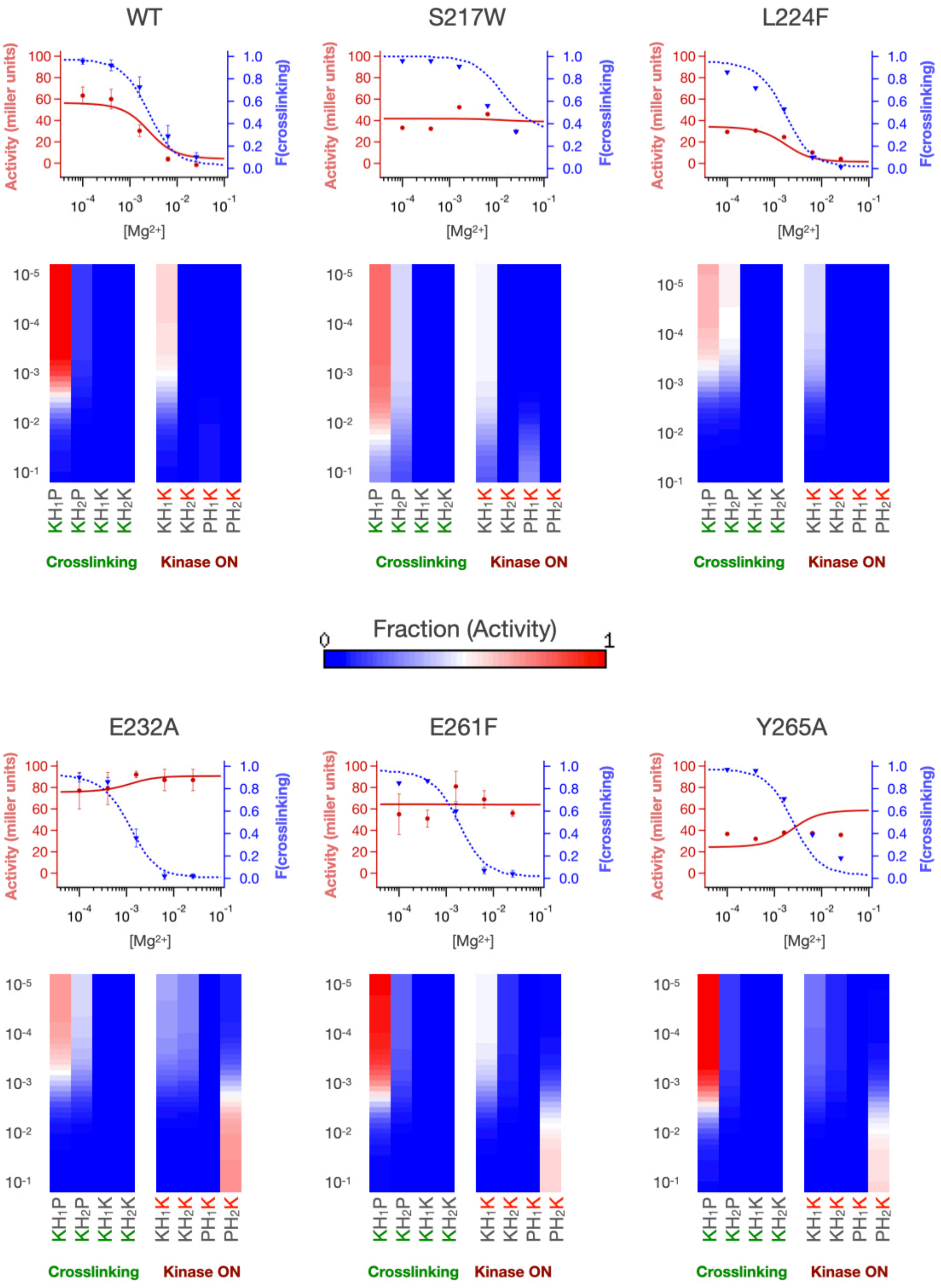


Figure 1.6 - Fitting of activity and crosslinking for select mutants. Autokinase activity and sensor crosslinking efficiency fits for wildtype (Y60C) and 5 point-mutants with a breakdown of the population fractions contributing to observed activity and crosslinking shown below.

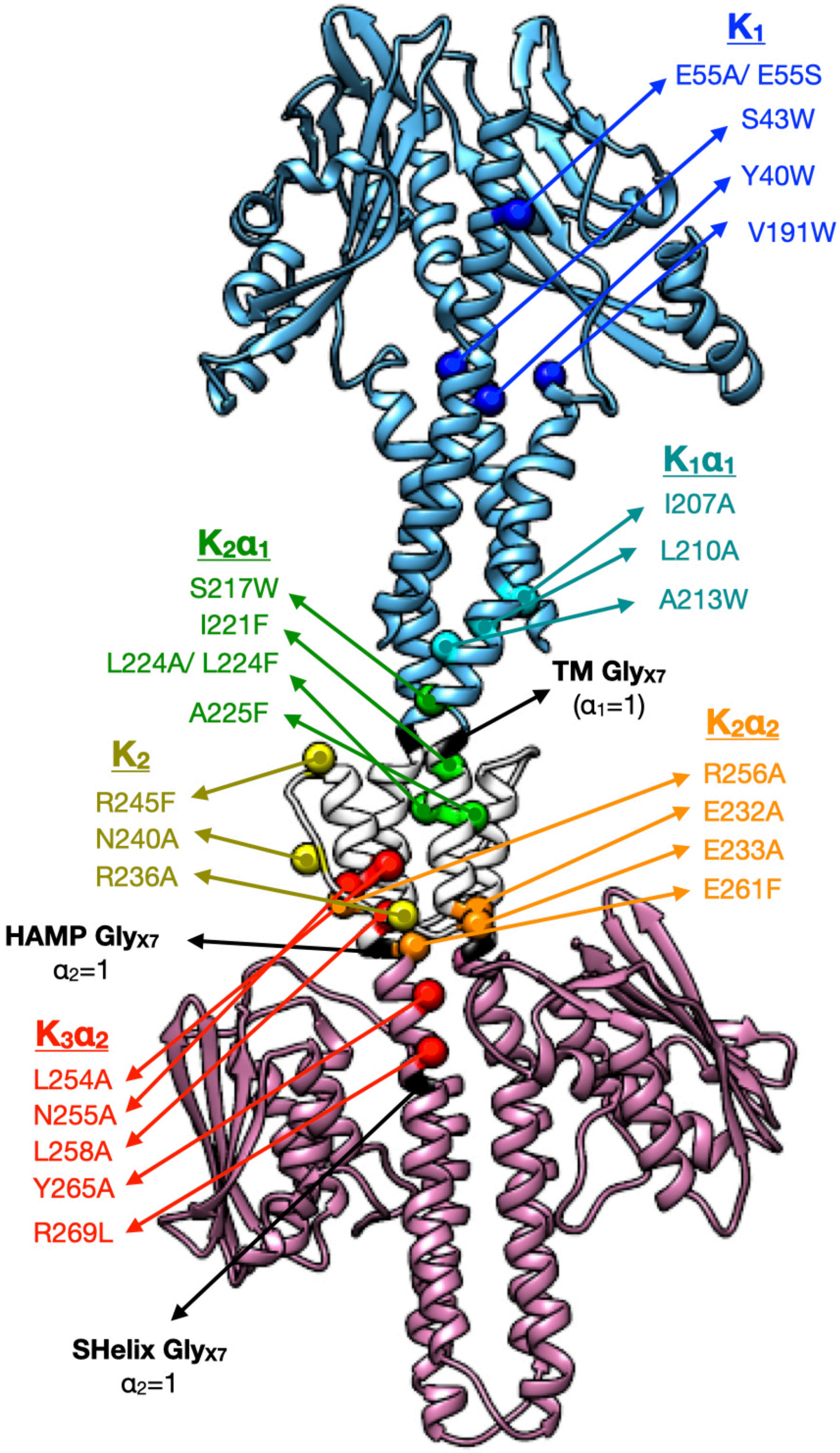


Figure 1.7 – Location of PhoQ mutants used in fitting. Parameters that were varied to fit mutations are indicated above.

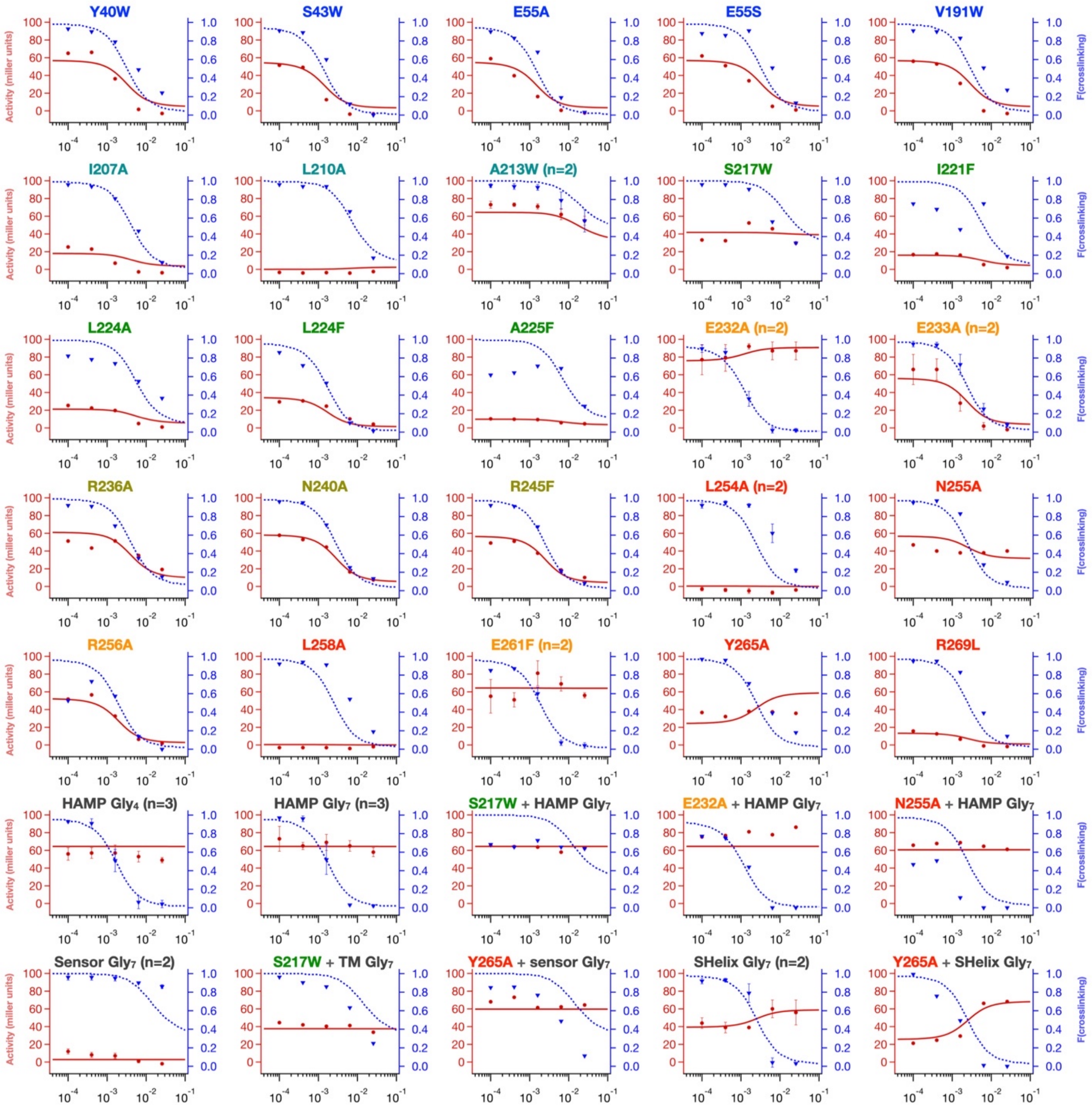


Figure 1.8 - Fitting of mutant and Gly_7 constructs. Measured kinase activity (red circles), kinase activity fit (red line), measured crosslinking efficiency (blue triangles) and crosslinking fit (blue line) are shown for entire PhoQ mutation dataset. Error bars are +/-SD where applicable. Position of mutants is shown on PhoQ model (left) and color coded by locally varied parameters so as to match **Figure 7**.

We were also able to obtain good fits to data from the vast majority of single point mutants and Gly_7 disconnections by locally varying the pertinent parameter (**Figure 1.8**). **Figure 1.7** shows the location of each mutation in a model of PhoQ color-coded by the identity of the locally varied parameter, and values for each parameter fit are described in **Table 1.1.** While the vast majority of our datasets were fit nicely with our allosteric model, some data-sets had poor fits, particularly in the prediction of sensor fraction(crosslinked) and are highlighted in **Figure 1.10**. For I221F, L224A and A225F point mutants, there is a tendency for the model to predict higher crosslinking efficiencies at low [Mg^{2+}] as compared to the data (**Figure 1.10A-C**); we expect this is in part due to potential over-estimation of un-crosslinked PhoQ in our experiments as a result of contamination by cytoplasmically expressed, but not yet membrane-inserted protein, which would be expected to remain un-crosslinked. There are also cases where the mid-point of transition is poorly fit (L254A, L258A), potentially owing to our choices of parameters to locally float for these mutants (**Figure 1.10D-E**). Indeed, significantly better fits were obtained by altering the parameters varied for these mutants from $K_2 + \alpha_2$ to $K_3 + \alpha_2$. This may suggest that these residues are involved in the underlying equilibrium of the autokinase domain itself due to their proximity to the S-helix (PhoQ 261-270). We also observe that both K_3 and α_2 drift to their lower parameter-bounded limits for these two mutants. This may explain why the mid-point of transition is still poorly fit, despite marked improvement compared to $K_2 + \alpha_2$ fits. Some combinations of mutants also had poor fits to crosslinking fraction (S217W + HAMP Gly_7, N255A + HAMP Gly_7, Y265A + Sensor Gly_7, **Figure 1.10F-H**). These may benefit from varying additional parameters that may be affected locally by the Gly_7 insertion, such as K_2 for '+ HAMP Gly_7' constructs and K_1 for '+ Sensor Gly_7' constructs. While these mutationss can possibly benefit from additional local variation of parameters, some mutations can be fit with fewer local parameters than were allowed in the fit, as the values for some of these locally fit parameters remain close to the globally fit value. K_1 (A225F), K_2 (I221F, A225F, E233A, R245F), and K_3 (N255A) are likely dispensable for the fit and are highlighted in **Table 1.1** (green). Despite this, the overall quality of fits is excellent given the global constraining of parameters and shows that it is possible to explain the various activities of PhoQ using a three-domain, two-state allosteric model and localizing the perturbation to a domain, rather than the entire protein. While iterative improvements to overall fit are possible by refining the choice of globally and locally fit parameters, we opted to remain with the classification of parameters that results from the expectation due to position of mutations on PhoQ structure.

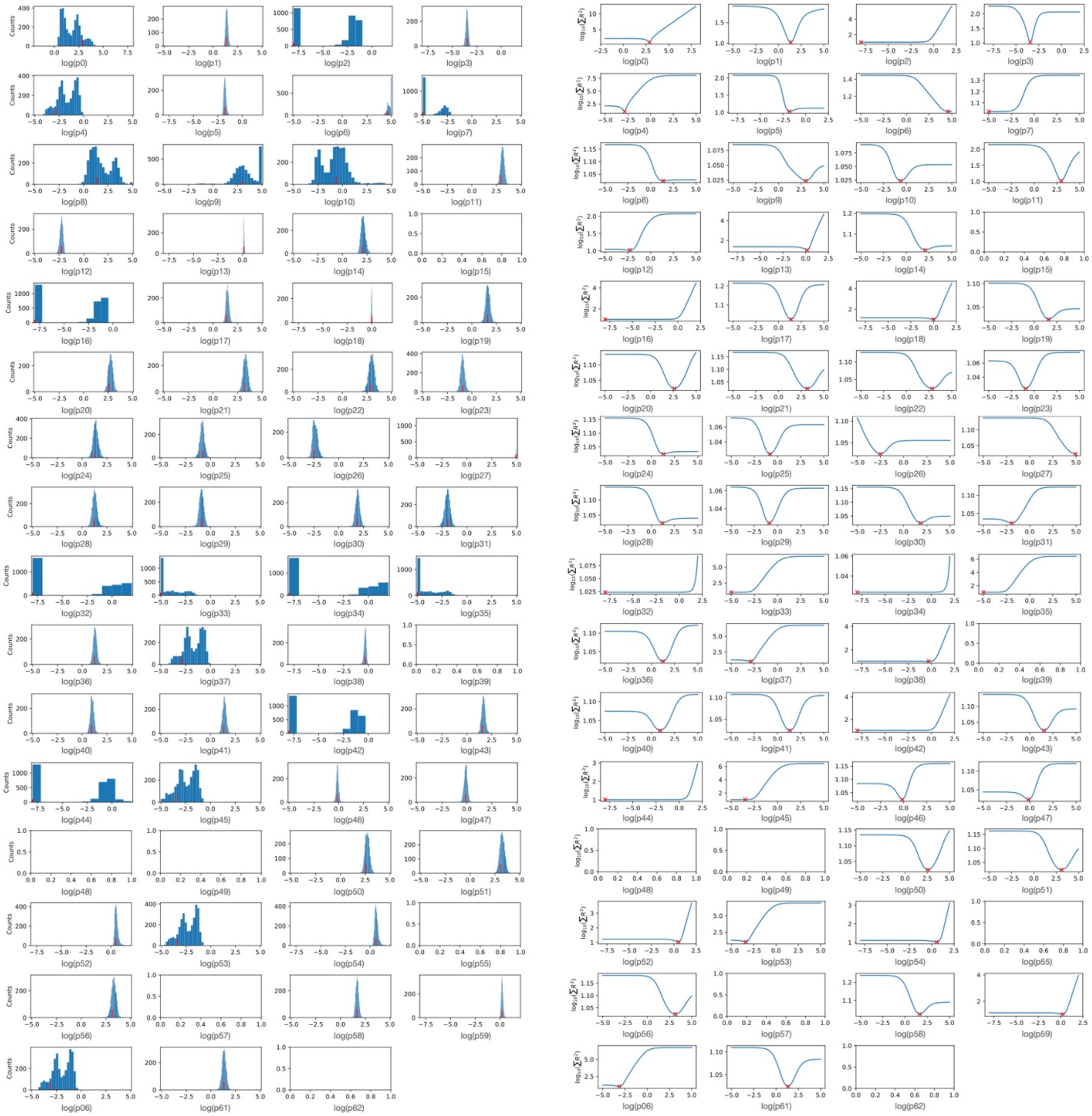


Figure 1.9 - Confidence intervals for fit parameters. Bootstrapped confidence intervals for the 63 parameter sets fit in allosteric model. Parameter values and affected mutations are listed in **Table 1.2**.

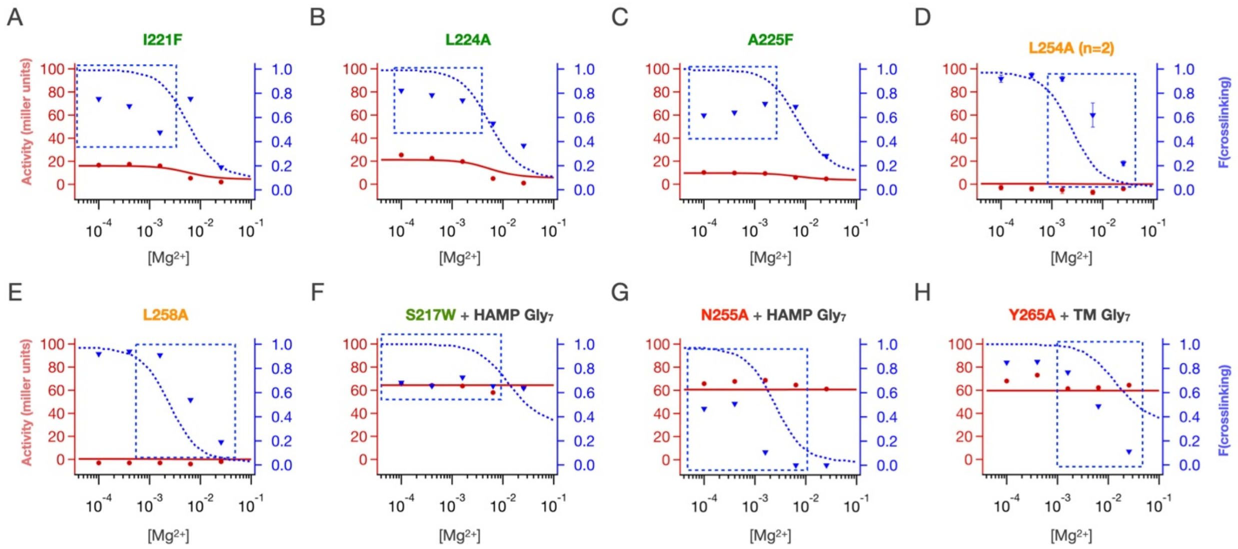


Figure 1.10 - Poor fits in allosteric model fitting. Poor fits were obtained for crosslinking at low [Mg^{2+}] for **(A)** I221F, **(B)** L224A and **(C)** A225F. Poor midpoints of crosslinking transitions were fit for **(D)** L254A and **(E)** L258A. Some combinations of mutations had poor crosslinking fits **(F)** S217W + HAMP Gly$_7$, **(G)** N255A + HAMP Gly$_7$ and activity fits **(H)** Y265A + sensor Gly$_7$

Covariability and sensitivity analysis of parameters in allosteric model

Parameters of an allosteric model are expected to have strong correlations due to the nature of allostery itself. For instance, the parameter for Mg^{2+} affinity to the sensor domain should compensate effectively for the parameter of the intrinsic equilibrium of the sensor domain itself, K_1. To a first approximation, the effect of lowering the affinity of Mg^{2+} for the sensor 'P' state is the same as lowering the equilibrium towards the 'P' state. Similarly, the parameter 'S' can largely compensate for K_3, as higher kinase activities can be a result of increased fraction(active kinase) or higher catalytic efficiency. This covariability is also inherent to the equilibrium and coupling constants themselves. An adjacent domain can have a strong intrinsic equilibrium which couples weakly to a neighboring domain (large K_i, small α_i), or a weak intrinsic equilibrium which couples strongly to a neighboring domain (small K_i, large α_i) and result in similar modulation of the equilibrium of that neighboring domain. In other words, sometimes, the product of variables is better defined than the individual variables themselves. This phenomenon is common in linked equilibria, such as equilibria describing monomer-multimer or folding-unfolding transitions, where often, intermediate equilibrium constants for lowly-populated species are poorly defined, while the overall equilibrium constant is well-defined.

In **Figure 1.11**, we analyze the independence of each parameter in multiple ways. Each panel shows the effect of varying a given parameter as all other values are held constant for the Y60C wildtype kinase activity and sensor crosslinking fits. To the right of these fits, two metrics showing the overall convergence of the parameter fit are displayed. We calculate a bootstrapped confidence interval for each global parameter, as well as the effect of varying the indicated parameter while holding all other parameters constant on the global fit-quality across the entire dataset is shown (see Methods). 7 of the 8 global parameters have well defined confidence intervals and minima in their global fit residuals when varied individually. The exception is the 'α_2' parameter, which drifts to one end of the explored parameter range. α_2 describes the relative abundance 'sensor-H_2-P' and 'sensor-H_2-K' state. Since H_2-P is a strongly coupled state, α_2 must be small so as to prevent uncorrelated activation of the autokinase. As α_2 gets lower, this weakly coupled 'sensor-H_2-K' state population fraction asymptotically approaches 0. Thus, while we can provide an upper bound for α_2 based on experiment sensitivity of autokinase activity as determined by gene-reporter assays, we cannot fit the fraction of this very low-abundance population to arbitrary precision. As can be seen in **Figure 1.11**, the quality of the global fit is insensitive to α_2 lower than ~10^{-3} and does not produce any changes to the calculated Mg^{2+} response activity or crosslinking curves at these values, but nonetheless continues to drift lower due to infinitesimal improvements to overall fit quality.

While some covariability is apparent between K_3 and S parameters, we find that all 8 parameters are truly independent and can only compensate for each other adequately in a small subregion of the overall explored parameter space. As can be seen in the top left panels, both 'K_{dP}' and 'K_1' parameters similarly vary the midpoint of Mg^{2+} response as well as the asymptotic values of activity and crosslinking at high Mg^{2+} when varied. However, they vary in their ability to modulate the steepness of activity and crosslinking transition and are therefore well defined. Similarly, both 'S' and 'K_3' largely raise or lower thresholds of kinase activity (**Figure 1.11**, 3rd row). However, only K_3 can additionally modulate sensor crosslinking activity, as well as change the shape of the Mg^{2+} - responsive curve itself as the 'S' parameter is absent from the crosslinking fitting equation. **Figure 1.12** demonstrates how the covariability between K_3 and S shifts as we explore the parameter space. In regions of parameter space where S is small and K_3 is large, 'K_3' and 'S' adequately compensate for one another, and as such the product, S.K_3 is better defined. But this ability to compensate is compromised as the K_3 parameter gets high enough.

We also generated confidence intervals and parameter sensitivity analyses for each of the 54 other parameters in our fit pertaining to locally fit parameters (**Figure 1.9**). As shown, a vast majority of the local parameters have well defined confidence intervals and residual minima. α_2 parameters for E233A, L254A, R256A, L258A and R269L drift to a minimum value as described above (also highlighted in **Table 1.1**, red). Additionally, K_3 parameter drifts to the minimum range for 2 mutants, L254A and L258A as mentioned previously. We believe the purpose of this exercise, the exploration of possible ligand-dependent behaviors of PhoQ through an allosteric signaling mechanism that involves a signal-transduction domain, and the robustness with which it reproduces observed PhoQ function, is adequately accomplished with the obtained fit quality.

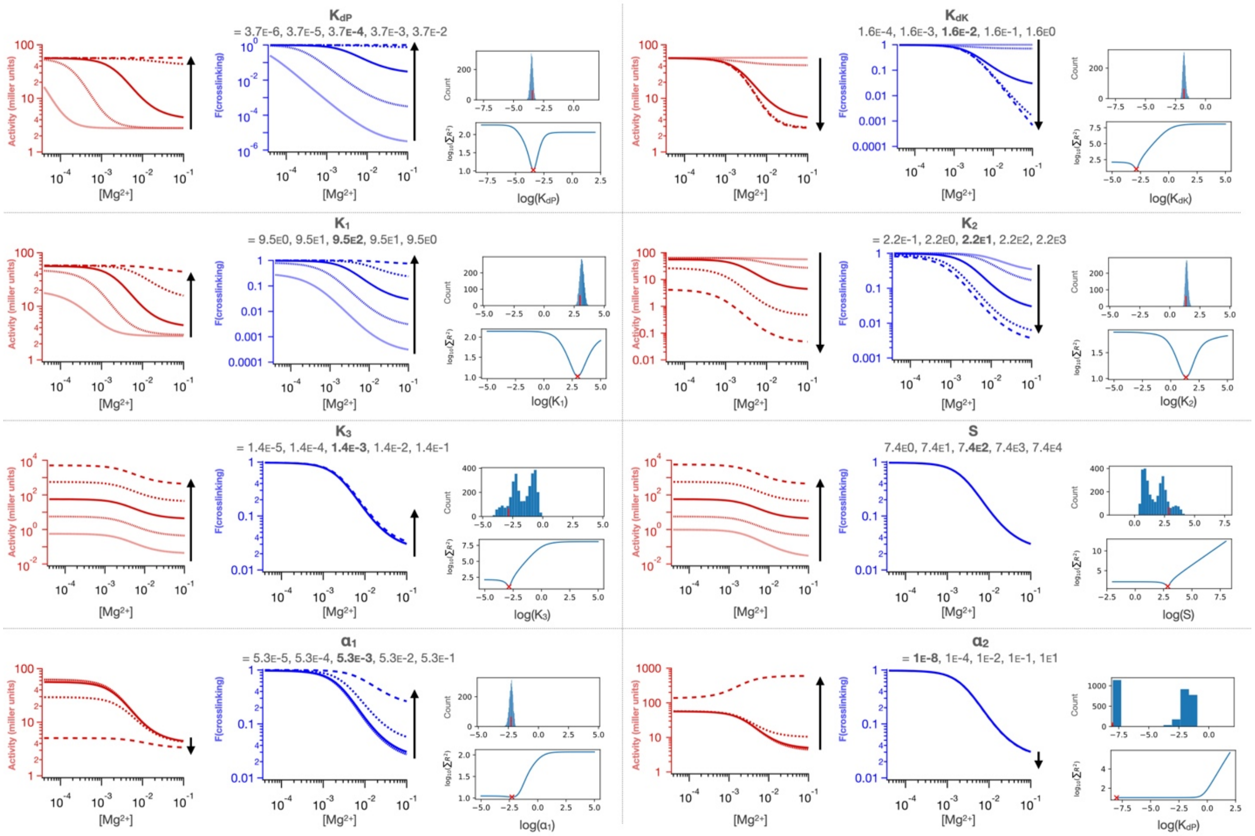


Figure 1.11 - Effect of varying parameters on WT fit. Predicted activity (red) and F(crosslinked) (blue) for wild type (Y60C) PhoQ. Parameters are varied to indicated range while holding all other parameters fixed in the global fit (WT fit value in bold font). Arrows show the behavior of curves as parameter value increases. Bootstrapped confidence intervals (top right) and sum of squared residuals for the entire global fit (bottom right) are plotted as a function of varying the indicated parameter. The parameter fit value is indicated with a red hash (|) on confidence intervals and a red cross (x) on fit sensitivity plots and summarized in **Table 1.1**.

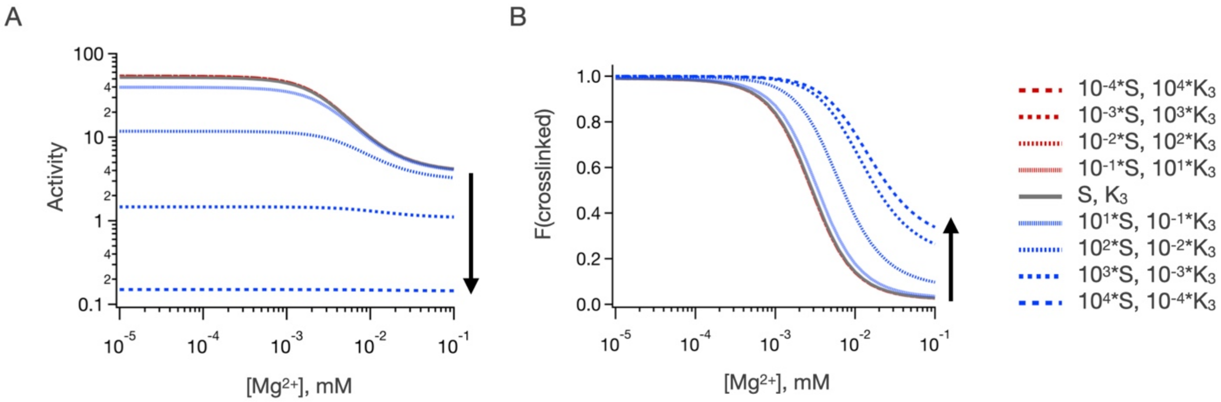


Figure 1.12 – Effect of varying parameters ‘S’ and ‘K$_3$’ while holding S.K$_3$ constant. Both (A) autokinase activity and (B) sensor crosslinking (B) fits for wild type PhoQ do not change as ‘S’ gets smaller and ‘K$_3$’ get larger (red traces) but are sensitive for smaller values of ‘K$_3$’ and larger values of ‘S’ (blue traces).

Two-state allosteric signaling model fails to explain dataset

Given the overparameterization of our allosteric model and the local covariability of variables within the parameter space, we examined whether simpler signaling models could be used to describe the activity of PhoQ and mutants. As mentioned in the previous section, global parameters ‘KdP’, ‘KdX’, and ‘S’ can be adequately folded into the core allosteric parameters K_1 and K_3 respectively. However, that does not simplify the underlying allosteric model of signal transduction itself. Similarly, a 2-state definition is the minimum number of states required for a signaling protein. Therefore, to simplify the model, we must reduce the number of domains that are treated as distinct allosterically coupled entities. A single-domain model in which the entire PhoQ model is treated as one concerted domain is immediately discounted due to its inability to accommodate sensor and autokinase functions that are not perfectly correlated as a function of Mg^{2+} as discussed earlier; many mutants of PhoQ have such uncorrelated activities. However, a two-domain allosteric model could potentially explain our dataset, as it can produce every type of PhoQ sensor crosslinking-activity pair behavior in isolation. In such a model, PhoQ would be divided up into two domains, the “sensor” which encompasses the periplasmic, transmembrane and HAMP domains of PhoQ all together (PhoQ res. 1-260), and the autokinase (**Figure 1.13A**). These 2 domains are allowed to sample two-state equilibria with equilibria constants K_1 and K_3, as previously described, and have a single allosteric coupling constant, α, between them. As such, the model encompasses 3 instead of 5 allosteric parameters (K_1, K_3, α) as well as the global non-allosteric parameters, K_{dP}, K_{dK} and S. When such a model is used to fit our entire dataset, we do not obtain good fits, particularly for the activity of the autokinase. **Figure 1.13B** shows the results of the two-state two-domain model on the Y60C dataset, in which all parameters are globally fit.

Due to the convergence of α = 0.94 globally, we observe effectively Mg^{2+} blind activities across the majority of our fits and it is not possible to simultaneously fit sensor crosslinking and autokinase activities simultaneously. As such, we propose that a minimum of 3 domains is necessary to explain the activity of PhoQ, given the two-state assumption.

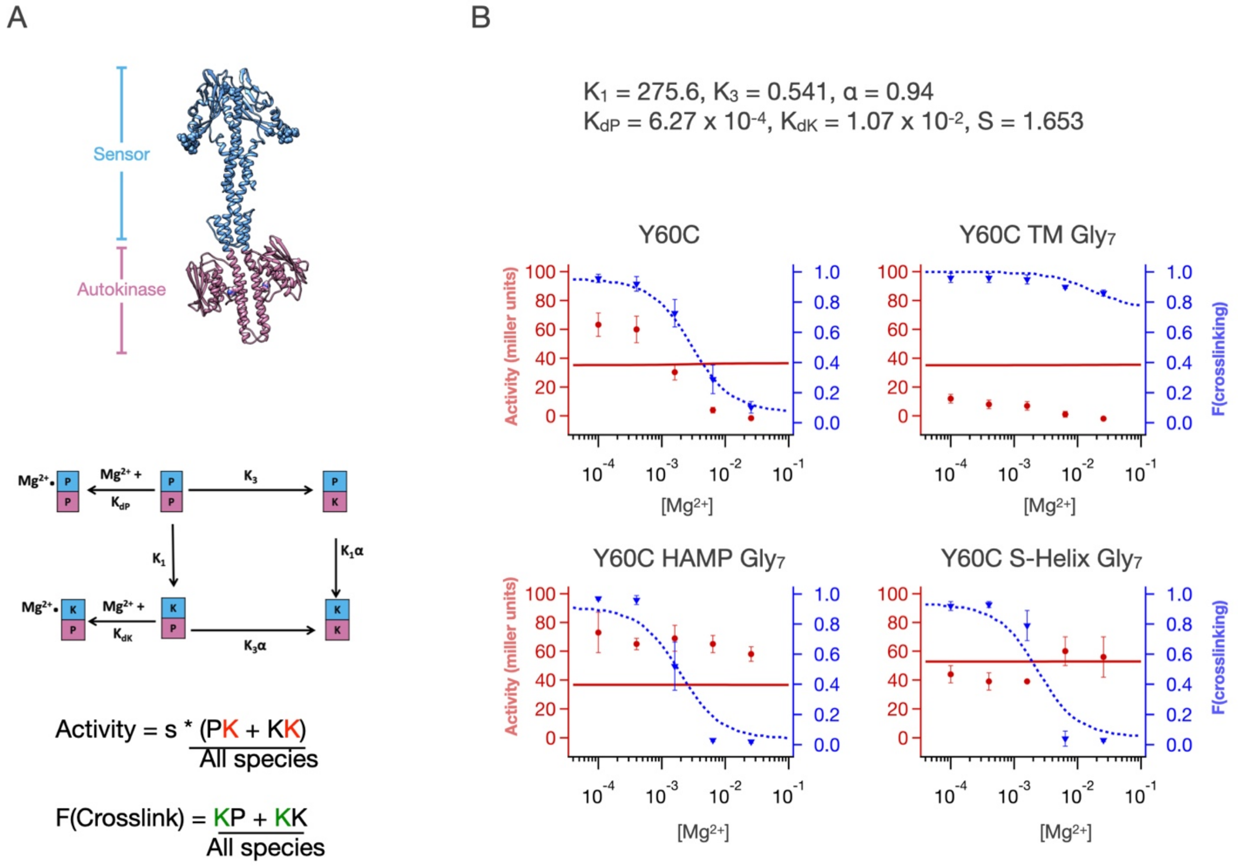


Figure 1.13 - Results of two-state two-domain allosteric model fit. (A) PhoQ is divided into 2 domains, a 'sensor' encompassing the periplasmic, TM and HAMP domains of PhoQ (1-260) and the autokinase (270-494). Activity and Fraction(crosslinking) are computed as shown below model. **(B)** Fit values and results for Y60C set. While adequate fits were obtained for Fraction(crosslinking), the model cannot simultaneously fit kinase activity data.

Glycine disconnections in other kinases

The observation that the HAMP domain of PhoQ had profound impact on the intrinsic activities of its sensor and autokinase led us to probe the generalizability of such behavior. We introduced Gly_7 disconnections in 2 more E. coli HAMP-containing HKs with similar architecture to PhoQ and used gene-reporters to evaluate the activity of these constructs. The HK CpxA responds to periplasmic protein misfolding stress via an accessory protein, CpxP, and upregulates genes to mitigate this stress, such as periplasmic proteases and chaperones and modulation of outer membrane porin expression (52–54). It is similar to PhoQ in that the free HK is kinase-active, and is turned off by the binding of the periplasmic CpxP protein (55). BaeS is a closely related HK, which has significant overlap with CpxA, both in known inducing stimuli and genes regulated, although the exact stimulus and sensing mechanisms are unknown (56). We examined the effect of Gly_7 disconnections in CpxA using a previously established reporter system using a plasmid encoding the CpxA variant of interest, the response regulator CpxR, and a GFP gene-reporter driven by the CpxR-controlled *p*cpxP promoter. We confirm the responsiveness of the cpxP reporter using a sub-inhibitory concentration of a known inducer of CpxA, Brilacidin (**Figure 1.14A**). The insertion of the Gly_7 motifs resulted in the same behavior observed in PhoQ, with the autokinase domain exhibiting a very high basal kinase activity when uncoupled from the HAMP, and the addition of the HAMP domain alone being sufficient for inhibiting this activity.

We also attempted to introduce histidine kinase + response regulator + fluorescent gene-reporter constructs of BaeS into a double CpxA/BaeS knock-out strain. The double KO background was necessary to remove any contributions of CpxA to the gene reporter signal reporting on BaeS activity. Unfortunately, these constructs were poorly tolerated in the double KO strain, with high-signal constructs getting inactivated by an uncharacterized mechanism. Similarly, the use of cytotoxic inducers of BaeS led to a selective depletion of high-signal cells, rather than the expected increase in reporter signal. Nonetheless, we were able to introduce Gly_7 insertions into BaeS and measure the basal activity of the kinase. Similar to both PhoQ and CpxA, the autokinase domain of BaeS showed very high kinase activity when decupled from the HAMP, and this activity was repressed by the re-addition of the HAMP domain alone (**Figure 1.14B**).

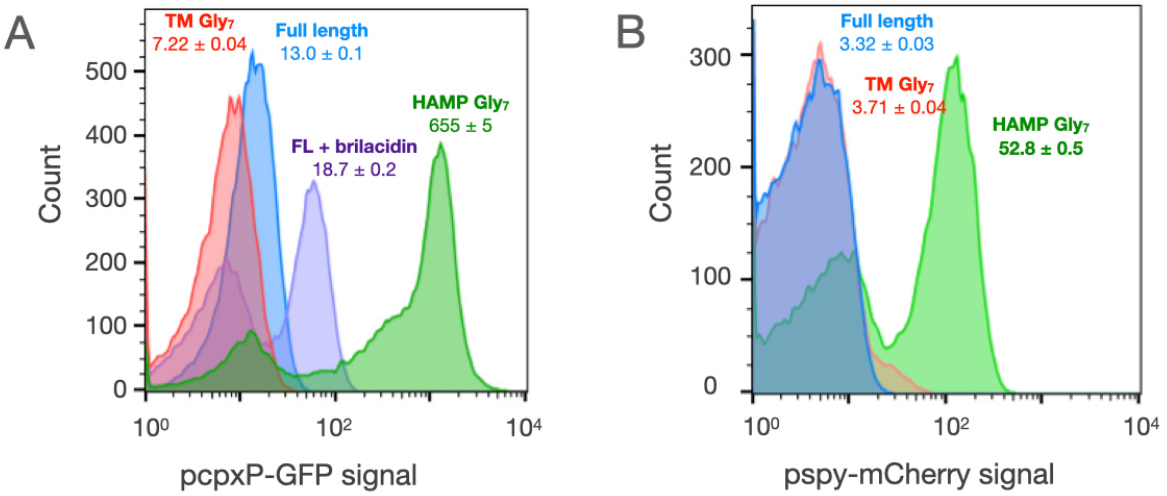


Figure 1.14 - Glycine disconnections in CpxA and BaeS. (A) The activity of CpxA constructs is measured in AFS51 strain (ΔcpxA) using a pcpxP::GFP reporter. Wild type CpxA is responsive to the antimicrobial mimetic, Brilacidin (purple histogram), although continued treatment selects for low signal population over time. The autokinase domain of CpxA in isolation shows very high kinase activity (green), which is repressed to basal levels by the addition of the HAMP domain alone (red). Median reporter fluorescence values are reported below labels. **(B)** The activity of BaeS constructs is measured in a ΔbaeS ΔcpxA strain using a pspy::mCherry reporter. The autokinase domain of BaeS shows high kinase activity (green), which is repressed by the addition of the HAMP domain alone (red). Average fluorescence values for each sample ±STE are reported.

Discussion

In this work, we sought to examine the rationale for the presence of modularly inserted signal transduction domains by examining the role of one such domain, the HAMP, in modulating the signaling population ensemble of a model HK, PhoQ. By using Gly_7 insertions, we show that the HAMP domain is strongly coupled to both sensor and autokinase domains in its thermodynamically favored 'H_2' state and acts in opposition to the intrinsic propensities of these domains. The poor ligand-binding sensor is potentiated for ligand-dependent state transition by the HAMP, and the high intrinsic activity of the autokinase domain is repressed. This creates an overall bistable sensor ideally positioned to respond to ligand from domains that, in isolation, have intrinsic equilibria that are far from ideal for the desired activity of PhoQ – the turning-off of autokinase activity upon ligand binding. This phenomenon is also observed in 2 structurally related E. coli HKs, CpxA, and BaeS, in which the HAMP again significantly alters the activity of the autokinase domain in the absence of sensory events. Therefore, it is attractive to hypothesize that the modular insertion of HAMP, and indeed other signal transduction domains, may serve to adjust innate equilibrium energetics of adjacent domains to physiologically relevant ranges.

We next established and experimentally fit a three-domain, two-state allosteric coupling model to see if we could explain the ligand dependent activity of wild type PhoQ and various functional mutants. We were able to globally fit our allosteric parameters across a large set of simultaneously determined sensor-crosslinking and autokinase activity measurements at 5 different ligand concentrations. Our results show that such a simple model, independent of any structural considerations, can account for the overall bistability of the sensor, and explain how local perturbations made by point mutations can robustly alter the signaling ensemble.

In some ways, our results also address the evolution of functional HK sensors by the apparent modular incorporation of one or more well-folded globular domains. The presence of an autokinase domain defines a histidine kinase, whereas the specific activity of the sensor domain determines its physiological function. These functionalities are subject to many evolutionary constraints, be it the specificity and affinity for ligands in sensor domains, the specificity for membrane homodimerization of HKs (57), or the cognate specificity for response regulator (13, 58–61) and the ability to inhabit and switch between the various conformations required for a full catalytic cycle in the autokinase domain (15). Furthermore, most two-component systems feature multiple accessory protein components involved in sensing, feedback regulation and cross-talk with other signaling systems, which add evolutionary constraints to these domains (62). In the closely related class of chemotaxis proteins, the analogous protein is also subject to extensive post-translational modifications that modulate activity. When all these evolutionary considerations are met, the resulting domain may not be ideally bi-stable in isolation. Indeed, in PhoQ, we found that the sensor strongly prefers to be in the low-affinity state when uncoupled from the HAMP domain, so much so that it remains in this state even at high [Mg^{2+}] (**Figure 1.3**). The favorable coupling of the thermodynamically preferred 'H_2' HAMP signaling state with the high-ligand-affinity phosphatase 'P' sensor state ameliorates this equilibrium into one that still moderately prefers kinase activation in the absence of Mg^{2+}, but can be overcome by Mg^{2+} binding at physiologically relevant ligand concentrations (**Figure 1.15**). Similarly, the autokinase domain has high kinase activity when uncoupled from the HAMP. The preferred 'H_2' HAMP signaling state strongly couples with the phosphatase 'P' autokinase state, strongly correlating a ligand-bound sensor state with a low-kinase activity autokinase state as desired. Therefore, the insertion of signal transduction domains into HK architectures could provide the free-energy currency needed to produce an overall bi-stable HK made up of non-bistable parts. The insertion of multiple signal transduction domains in some HKs may also reflect the robustness of this strategy; in the absence of any fitness costs associated with finding the most optimal solution to creating a bistable protein, many numbers and combinations of signal transduction domains can produce the same desired outcome, so long as one can easily adjust the strength of allosteric coupling between domains.

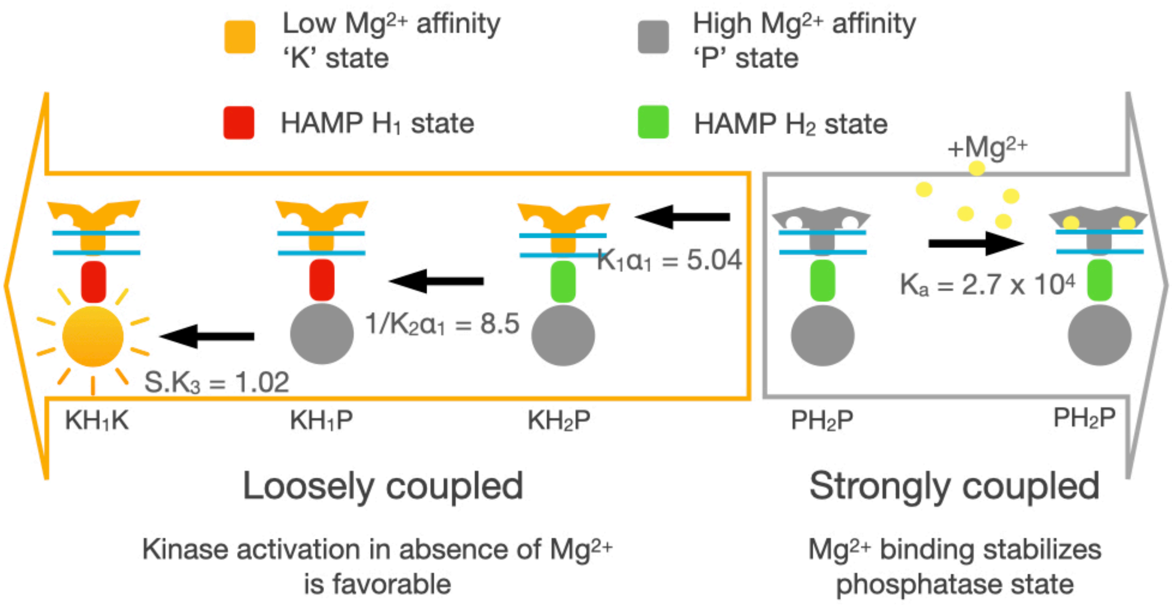

Low Mg2+ affinity 'K' state
High Mg2+ affinity 'P' state
HAMP H1 state
HAMP H2 state
+Mg2+
K1α1 = 5.04
Ka = 2.7 x 10^4
1/K2α1 = 8.5
S.K3 = 1.02
KH1K
KH1P
KH2P
PH2P
PH2P
Loosely coupled
Strongly coupled
Kinase activation in absence of Mg2+ is favorable
Mg2+ binding stabilizes phosphatase state

Figure 1.15 – Coupling to HAMP domain produces desired PhoQ ensemble behavior. The preferred signaling state of the HAMP (green) biases the ensemble to a bistable sensor that in the absence of Mg^{2+} activates with a moderately favorable equilibrium but can be driven to the inactive conformation by Mg^{2+} binding.

We also find in our data that point mutations close to domain-domain interfaces can produce large changes in ligand-dependent behavior, suggesting the alpha-helical connections between domains are easily tunable through simple sequence changes, enabling robust evolution of physiologically relevant HKs. We have shown that the insertion of a few glycine residues is sufficient to completely uncouple domains. On the other extreme, a well folded coiled coil junction can create strong allosteric coupling due to the cooperative folding and stability of such a motif. A range of stabilities can be achieved by various means, including the insertion or deletion of one or more residues to disrupt the canonical heptad pattern of hydrophobic residues of the dimeric core of the protein, as is often observed in the conserved S-helix motif in HKs (51). Schmidt *et. al.* recently (63) showed that crystal structures of cytoplasmic domains in different conformations accommodate the structural deviations of these S-Helix sequence insertions by diffusing the strain over different lengths of the proximal alpha-helical core. These different “accommodation lengths” could reflect the different strengths of allosteric coupling depending on the signaling states of the adjacent domains. We also find conservation of poly-glycine motifs and helix-disrupting proline residues in the juxta-membrane regions of chemotaxis proteins and HKs respectively (46, 47, 64), which could serve as additional modulators of allosteric coupling strength. In some systems, domains are actually segregated to entirely different proteins, in which case the strength of the protein-protein interaction between components can be altered to vary allosteric coupling. These are all evolutionarily accessible solutions to fine-tune the function of a kinase.

Finally, this evolutionary argument may also explain the lack of a parsimonious structural mechanism for signal transduction, even in HKs with a specific domain architecture. Although this problem is largely exacerbated by the dearth of multi-domain structures of HKs in various signaling conformation, several signaling hypotheses have been put forward regarding the structural mechanism for signal transduction in HKs, particularly in HAMP domains. These include the gear-box mechanism (AF1503, Aer2 multi-HAMP)(65), Piston mechanism (Tar) (66, 67), Scissoring mechanism (Tar, BT4663, PhoQ) (39, 68, 69), Orthogonal displacement mechanism (HAMP tandems, Tar) (70–72) and the dynamic HAMP mechanism (Adenylate cyclase HAMP) (73–75). A recently elucidated set of structures of the sensor, TM and signal transduction domains of NarQ remains the only representative of a multidomain transmembrane structure of an HK containing a signal transduction domain, and again shows a rigid-body bending transition of the HAMP domain about the conserved N-term Proline between apo- and holo-states of the sensor(50). It may be that signal transduction mechanisms in HKs are as varied as their modular architecture, and many structural transitions could account for the underlying concern in signaling, which is the allosteric modulation of multi-state equilibria of adjacent domains in response to structural transitions caused by a sensory event. The idea that autokinase domains intrinsically have high-kinase activity and are subsequently inhibited by strong coupling to up-stream domains and the further stabilization of these inhibitory conformations by ligand-binding warrants examination as a generalizable signaling mechanism for histidine kinases. In subsequent chapters, we further examine the sequence and structural basis for one of the established role of the HAMP domain in this chapter, namely, its ability to turn off the intrinsically high autokinase activity, and how this ability is modulated both through changes in HAMP structure, and the frustration of its linkage to the autokinase via the S-Helix motif.

Materials and Methods:

Materials:

BW25113 and HK knockout strains were obtained from the Keio collection.

TIM206 was obtained from Tim Mayashiro (Goulian lab)

Inducers (Iron chloride, Copper sulfate, Sodium Nitrate, Magnesium sulfate)

Brilacidin was obtained from Polymedix

MOPS minimal media

Z-buffer, ONPG, bMe, Na_2CO_3

NEM, Urea, LDS, Tris-SDS gels, buffer

Anti-PentaHis antibody, ECL substrate

Methods:

Cloning: PhoQ mutants were cloned into the pTrc99a plasmid MCS by restriction cloning. Point mutations were made by quick-change mutagenesis and confirmed by sanger sequencing. Hybrid HK-gene reporter plasmids were built in pTrc99a plasmid by introducing a c-terminally 6x His-tagged HK construct into the IPTG inducible MCS, and the mCherry reporter sequence downstream by Gibson cloning. Sequences of the promoters used are listed in the supplement. Gly_7 disconnections and point mutations were introduced by a blunt-end ligation strategy and confirmed by sanger sequencing.

Growth of PhoQ constructs: TIM206 (genotype: ΔPhoQ, mgrB::LacZ) containing various pTrc99a-phoQ were grown overnight at 37 °C in MOPS minimal media + 50 µg/mL AMP and 1 mM $MgSO_4$. These overnights were then diluted 50x into 1 mL MOPS media + 50 µg/mL AMP and 1 mM

$MgSO_4$, and grown at 37°C for 2 hours. These cultures were further diluted 500X into 30 mL MOPS minimal media + 50 µg/mL AMP containing 0.1, 0.4, 1.6, 6.4 and 25.6 mM $MgSO_4$, and grown for at least 5 hours such that the density of the culture reaches log-phase (OD_{600} = 0.2 – 0.8). 500 µL of culture is removed for evaluating beta galactosidase activity, while the remaining culture is used for western analysis.

Beta galactosidase activity: 500 µL of PhoQ culture was combined with 500 µL of 1x Z-buffer, 25 µL of 0.1% SDS in water, and 50 µL of chloroform in a glass test-tube and vortexed for complete lysis. The lysate was then prewarmed to 37°C in a standing incubator before addition of substrate. 0.25 mL of prewarmed 4 mg/mL ONPG in 1x Z-buffer was added to the lysate to initiate hydrolysis, which was then quenched with the addition of 500 µL of 1M Na_2CO_3 after variable incubation periods. The quenched hydrolysis was then centrifuged to remove any cell debris, and absorbance at 420 nm and 550 nm was measured in triplicate using a Biotek synergy2 plate-reader with pathlength correction. Miller units were calculated as follows:

Miller units = 1000*(OD_{420} – 1.75*OD_{550})/(OD_{600}*dilution factor*incubation time)

Membrane prep and western analysis: 30 mL of PhoQ culture was centrifuged at 4°C for 20 minutes to collect a cell-pellet. This cell-pellet was immediately frozen in liquid nitrogen and stored at -80°C until analysis. Frozen pellets were first thawed, suspended and incubated on ice with 500 µg/mL N-Ethylmaleimide (NEM) and 1 mg/mL lysozyme in 50 mM TRIS buffer, pH 8, for 1 hour. Cells were then lysed by 30 seconds of tip-sonication. Lysed cells were then centrifuged at 16000xg for 10 minutes to remove cell debris. Membrane was isolated from the supernatant by centrifugation at 90,000xg for 10 minutes. Membrane pellets were then resuspended in 1X

LDS buffer containing 8M Urea and 500 mM NEM, boiled at 95°C for 10 minutes and analyzed by western blot.

Western blotting: Samples were first separated by TRIS-SDS gel electrophoresis at 200V for 70 min, and then transferred onto nitrocellulose membranes by dry transfer (iBlot2). Membranes were then blocked using 1% BSA in TBS-t buffer (20 mM Tris, 2.5 mM EDTA, 150 mM NaCl, 0.1% Tween-20), probed using an anti-pentaHis HRP antibody, and visualized using luminescent ECL substrate on a BioRad imager. Bands corresponding to PhoQ monomer and dimer were quantified using Image-J software to yield a crosslinking efficiency between 0 and 1.

Measuring activity of CpxA, BaeS: HK constructs were cloned into the MCS of pTrc99a plasmid, and the associated fluorescent reporter gene was cloned downstream. For the CpxA reporter plasmid, the RR, CpxR, was also cloned into the MCS and transformed into AFS51 strain (∆cpxA∆pta::Kan cpxP::GFP) by heat shock transformation. For BaeS, the RR, BaeR, was cloned into an additional plasmid, pSEVA331 under an IPTG inducible promoter and both plasmids were transformed into a ΔBaeSΔCpxA double KO strain by heat shock transformation. Cultures were started by diluting overnights 200-500 folds into fresh LB + AMP media and allowed to grow to mid-log phase (OD_{600} = 0.4 – 0.6) before analysis by flow cytometry. The responsiveness of gene-reporters was confirmed by treating log-phase cultures with reported inducers at various concentrations for 1.5 hours before analysis. Expression of HKs was confirmed by western analysis using the c-terminal 6x His-tag for quantification.

Flow cytometry: LB cultures at mid-log phase were diluted 20x into 1x PBS buffer and 20,000 cells gated by forward and side-scatter were evaluated for GFP fluorescence (pcpxP::GFP; Ex. 488 nm, Em. 515 nm) or mCherry fluorescence (pspy::mCherry, Ex. 488 nm, Em. 620 nm) per sample on a BD FACS caliber instrument. Sample average fluorescence and standard error were determined by standard analysis using Flo-Jo software.

Data Fitting: Kinase active and sensor crosslinking competent states are partitioned to generate expressions dependent on [Mg^{2+}] as the lone variable as shown below. The parameters are then fit globally across all datasets, except for those accounting for the perturbation of a mutant/ Gly$_7$ disconnection, which are fit locally. Locally fit parameters are kept identical between replicates or additive mutations.

$$\text{F active} = \frac{\Sigma\ \textit{Kinase}\!-\!\textit{ON states}}{\Sigma\ \textit{All states}} = \frac{\mathrm{OOX} + \mathrm{OXX} \qquad + \mathrm{XOX} + \mathrm{XXX}}{\mathrm{OOO} + \mathrm{OXO} + \mathrm{OOX} + \mathrm{OXX} \qquad + \mathrm{XOO} + \mathrm{XXO} + \mathrm{XOX} + \mathrm{XXX}}$$

$$= \frac{(OOO + \mathrm{OOO.Mg}) * (\mathrm{K3} + \mathrm{K2.K3a2}) \qquad + (\mathrm{XOO} + \mathrm{XOO.Mg}) * (K3 + K3.K2a1a2)}{(\mathrm{OOO} + \mathrm{OOO.Mg}) * (1 + K2 + K3 + K3a2.K2) \quad + (\mathrm{XOO} + \mathrm{XOO.Mg}) * (1 + K2.a1 + K3 + K3.K2a1a2)}$$

$$= \frac{\left[1 + \frac{[\boldsymbol{Mg}]}{\boldsymbol{Kdo}}\right]^2 * (\mathrm{K3} + \mathrm{K2.K3a2}) + \left[1 + \frac{[\boldsymbol{Mg}]}{\boldsymbol{Kdx}}\right]^2 * (K1.K3 + K1.K3.K2a1a2)}{\left[1 + \frac{[\boldsymbol{Mg}]}{\boldsymbol{Kdo}}\right]^2 * (1 + K2 + K3 + K3a2.K2\Big) \quad + \left[1 + \frac{[\boldsymbol{Mg}]}{\boldsymbol{Kdx}}\right]^2 * (K1 + K1a1.K2 + K1.K3 + K1.K3.K2a1a2)}$$

$$\text{F xlink} = \frac{\Sigma\ \textit{sensor ON states}}{\Sigma\ \textit{All states}} = \frac{\mathrm{XOO} + \mathrm{XXO} + \mathrm{XOX} + \mathrm{XXX}}{\mathrm{OOO} + \mathrm{OXO} + \mathrm{OOX} + \mathrm{OXX} \qquad + \mathrm{XOO} + \mathrm{XXO} + \mathrm{XOX} + \mathrm{XXX}}$$

$$= \frac{\left[1 + \frac{[\boldsymbol{Mg}]}{\boldsymbol{Kdx}}\right]^2 * (K1 + K1.K2a1 + K1.K3 + K1.K2.K3a1a2)}{\left[1 + \frac{[\boldsymbol{Mg}]}{\boldsymbol{Kdo}}\right]^2 * (1 + K2 + K3 + K3a2.K2\Big) \quad + \left[1 + \frac{[\boldsymbol{Mg}]}{\boldsymbol{Kdx}}\right]^2 * (K1 + K1a1.K2 + K1.K3 + K1.K3.K2a1a2)}$$

Activity and crosslinking data were globally fit by first scaling the activity data by a factor of q = (mean of act data) / (mean of xlink data). The refactored data that were globally fit to a 3-state allosteric model was then Data(activity) / q and Data(xlink). Each dataset was fit by a combination of global and local parameters. Global parameters were shared between replicates of the data as well as mutations that were functionally similar were grouped together for fitting grouped parameters. Error analysis was performed through bootstrapping the residuals to calculate confidence intervals, as well as residual sweep analyses. For bootstrapping, we randomly shuffled the residuals from the optimum fit (residuals were chosen at random with replacement) and created new synthetic datasets by adding these residuals to the activity and xlink values of the optimum fit. The parameters were then optimized to fit these new synthetic datasets, with initial parameter values taken from the optimum fit. For residual sweep analysis, all but one of the parameters were fixed to their optimum values, and the variable under analysis was swept across its allowed numerical range, after which the sum of squares of residuals was calculated. The sum of squares was then plotted as a function of the parameter's numerical value.

Table 1.1- List of mutant parameter fits.

Mutant	K_1	K_2	K_3	α_1	α_2	s	K_{dP}	K_{dK}
Y60C	9.5 E +02	2.2 E +01	1.3 E -03	5.3 E -03	1.0 E -08	7.4 E +02	3.7 E -04	1.6 E -02
Y60C HAMP 4		4.5 E +01			1.0 E +00			
Y60C HAMP 7					1.0 E +00			
Y60C SH7		2.1 E +01	7.7 E -04		1.7 E +00			
Y60C TM7				1.0 E +00				
Y40W	1.4 E +03							
S43W	3.8 E +02							
E55A	4.1 E +02							
E55S	1.5 E +03							
V191W	1.2 E +03							
I207A	6.9 E +02			1.1 E -01				
L210A	3.1 E -03			9.1 E +04				
A213W	4.0 E +04			1.0 E -05				
S217W		7.1 E -01		7.6 E -01				
S217W + H7					1.0 E +00			
S217W + TM7				1.0 E +00				
I221F		2.0 E +01		1.5 E -01				
L224A		1.6 E +01		1.3 E -01				
L224F		6.8 E +01		1.2 E -02				
A225F	1.0 E +03	2.4 E +01		2.3 E -01				
E232A		1.1 E +02			1.4 E +00			
E232A + H7					1.0 E +00			
E233A		2.3 E +01			1.0 E -08			
R236A		8.6 E +00						
N240A		1.7 E +01						
R245F		2.2 E +01						
L254A			1.0 E -05		1.0 E -08			
N255A			1.3 E -03		4.9 E -01			
N255A + H7					1.0 E +00			
R256A		3.6 E +01			1.0 E -08			
L258A			1.0 E -05		1.0 E -08			
E261F		3.9 E +01			9.9 E -01			
Y265A			4.1 E -04		3.2 E +00			
Y265A + TM7				1.0 E +00				
Y265A + SH7					3.8 E +00			
R269L			3.3 E -04		1.0 E -08			

Table 1.2 - parameters used in fitting

Par#	Par.	Lower bound	Upper bound	Fit datasets affected
0	S	1.0E-02	1.0E+08	A213W_2, A213W_1, A225F, E232A_1, E232A_2, E232A_HAMP_Gly7, E233A_1, E233A_2, E261F_1, E261F_2, E55A, E55S, I207A, I221F, L210A, L224A, L224F, L254A_1, L254A_2, L258A, N240A, N255A, N255A_HAMP_Gly7, R236A, R245F, R256A, R269L, S217W, S217W_HAMP_Gly7, S217W_Sensor_Gly7, S43W, V191W, Y265A, Y265A_SHelix_Gly7, Y265A_Sensor_Gly7, Y40W, Y60C_1, Y60C_2, Y60C_3, Y60C_4, Y60C_5, Y60C_6, Y60C_7, Y60C_8, Y60C_9, Y60C_HAMP_Gly4_1, Y60C_HAMP_Gly4_2, Y60C_HAMP_Gly4_3, Y60C_HAMP_Gly7_1, Y60C_HAMP_Gly7_2, Y60C_HAMP_Gly7_3, Y60C_SHelix_Gly7_1, Y60C_SHelix_Gly7_2, Y60C_Sensor_Gly7_1, Y60C_Sensor_Gly7_2
1	K_2	1.0E-05	1.0E+05	A213W_2, A213W_1, E55A, E55S, I207A, L210A, L254A_1, L254A_2, L258A, N255A, N255A_HAMP_Gly7, R269L, S43W, V191W, Y265A, Y265A_SHelix_Gly7, Y265A_Sensor_Gly7, Y40W, Y60C_1, Y60C_2, Y60C_3, Y60C_4, Y60C_5, Y60C_6, Y60C_7, Y60C_8, Y60C_9, Y60C_Sensor_Gly7_1, Y60C_Sensor_Gly7_2
2	α2	1.0E-08	1.0E+02	A213W_2, A213W_1, A225F, E55A, E55S, I207A, I221F, L210A, L224A, L224F, N240A, R236A, R245F, S217W, S217W_Sensor_Gly7, S43W, V191W, Y40W, Y60C_1, Y60C_2, Y60C_3, Y60C_4, Y60C_5, Y60C_6, Y60C_7, Y60C_8, Y60C_9, Y60C_Sensor_Gly7_1, Y60C_Sensor_Gly7_2
3	KdP	1.0E-08	1.0E+02	A213W_2, A213W_1, A225F, E232A_1, E232A_2, E232A_HAMP_Gly7, E233A_1, E233A_2, E261F_1, E261F_2, E55A, E55S, I207A, I221F, L210A, L224A, L224F, L254A_1, L254A_2, L258A, N240A, N255A, N255A_HAMP_Gly7, R236A, R245F, R256A, R269L, S217W, S217W_HAMP_Gly7, S217W_Sensor_Gly7, S43W, V191W, Y265A, Y265A_SHelix_Gly7, Y265A_Sensor_Gly7, Y40W, Y60C_1, Y60C_2, Y60C_3, Y60C_4, Y60C_5, Y60C_6, Y60C_7, Y60C_8, Y60C_9, Y60C_HAMP_Gly4_1, Y60C_HAMP_Gly4_2, Y60C_HAMP_Gly4_3, Y60C_HAMP_Gly7_1, Y60C_HAMP_Gly7_2, Y60C_HAMP_Gly7_3, Y60C_SHelix_Gly7_1, Y60C_SHelix_Gly7_2, Y60C_Sensor_Gly7_1, Y60C_Sensor_Gly7_2
4	K3	1.0E-05	1.0E+05	A213W_2, A213W_1, A225F, E232A_1, E232A_2, E232A_HAMP_Gly7, E233A_1, E233A_2, E261F_1, E261F_2, E55A, E55S, I207A, I221F, L210A, L224A, L224F, N240A, R236A, R245F, R256A, S217W, S217W_HAMP_Gly7, S217W_Sensor_Gly7, S43W, V191W, Y40W, Y60C_1, Y60C_2, Y60C_3, Y60C_4, Y60C_5, Y60C_6, Y60C_7, Y60C_8, Y60C_9, Y60C_HAMP_Gly4_1, Y60C_HAMP_Gly4_2, Y60C_HAMP_Gly4_3, Y60C_HAMP_Gly7_1, Y60C_HAMP_Gly7_2, Y60C_HAMP_Gly7_3, Y60C_Sensor_Gly7_1, Y60C_Sensor_Gly7_2

Par#	Par.	lower bound	Upper bound	Fit datasets affected
5	KdK	1.0E-08	1.0E+02	A213W_2, A213W_1, A225F, E232A_1, E232A_2, E232A_HAMP_Gly7, E233A_1, E233A_2, E261F_1, E261F_2, E55A, E55S, I207A, I221F, L210A, L224A, L224F, L254A_1, L254A_2, L258A, N240A, N255A, N255A_HAMP_Gly7, R236A, R245F, R256A, R269L, S217W, S217W_HAMP_Gly7, S217W_Sensor_Gly7, S43W, V191W, Y265A, Y265A_SHelix_Gly7, Y265A_Sensor_Gly7, Y40W, Y60C_1, Y60C_2, Y60C_3, Y60C_4, Y60C_5, Y60C_6, Y60C_7, Y60C_8, Y60C_9, Y60C_HAMP_Gly4_1, Y60C_HAMP_Gly4_2, Y60C_HAMP_Gly4_3, Y60C_HAMP_Gly7_1, Y60C_HAMP_Gly7_2, Y60C_HAMP_Gly7_3, Y60C_SHelix_Gly7_1, Y60C_SHelix_Gly7_2, Y60C_Sensor_Gly7_1, Y60C_Sensor_Gly7_2
6	K1	1.0E-05	1.0E+05	A213W_2, A213W_1
7	α1	1.0E-05	1.0E+05	A213W_2, A213W_1
8	K2	1.0E-05	1.0E+05	A225F
9	K1	1.0E-05	1.0E+05	A225F
10	α1	1.0E-05	1.0E+05	A225F
11	K1	1.0E-05	1.0E+05	E232A_1, E232A_2, E232A_HAMP_Gly7, E233A_1, E233A_2, E261F_1, E261F_2, I221F, L224A, L224F, L254A_1, L254A_2, L258A, N240A, N255A, N255A_HAMP_Gly7, R236A, R245F, R256A, R269L, S217W, S217W_HAMP_Gly7, S217W_Sensor_Gly7, Y265A, Y265A_SHelix_Gly7, Y265A_Sensor_Gly7, Y60C_1, Y60C_2, Y60C_3, Y60C_4, Y60C_5, Y60C_6, Y60C_7, Y60C_8, Y60C_9, Y60C_HAMP_Gly4_1, Y60C_HAMP_Gly4_2, Y60C_HAMP_Gly4_3, Y60C_HAMP_Gly7_1, Y60C_HAMP_Gly7_2, Y60C_HAMP_Gly7_3, Y60C_SHelix_Gly7_1, Y60C_SHelix_Gly7_2, Y60C_Sensor_Gly7_1, Y60C_Sensor_Gly7_2
12	α1	1.0E-05	1.0E+05	E232A_1, E232A_2, E232A_HAMP_Gly7, E233A_1, E233A_2, E261F_1, E261F_2, E55A, E55S, L254A_1, L254A_2, L258A, N240A, N255A, N255A_HAMP_Gly7, R236A, R245F, R256A, R269L, S43W, V191W, Y265A, Y265A_SHelix_Gly7, Y40W, Y60C_1, Y60C_2, Y60C_3, Y60C_4, Y60C_5, Y60C_6, Y60C_7, Y60C_8, Y60C_9, Y60C_HAMP_Gly4_1, Y60C_HAMP_Gly4_2, Y60C_HAMP_Gly4_3, Y60C_HAMP_Gly7_1, Y60C_HAMP_Gly7_2, Y60C_HAMP_Gly7_3, Y60C_SHelix_Gly7_1, Y60C_SHelix_Gly7_2
13	α2	1.0E-08	1.0E+02	E232A_1, E232A_2
14	K2	1.0E-05	1.0E+05	E232A_1, E232A_2, E232A_HAMP_Gly7
15	α2	1.0E+00	1.0E+00	E232A_HAMP_Gly7
16	α2	1.0E-08	1.0E+02	E233A_1, E233A_2
17	K2	1.0E-05	1.0E+05	Y265A_Sensor_Gly7
18	α2	1.0E-08	1.0E+02	Y265A_SHelix_Gly7
19	K2	1.0E-05	1.0E+05	V191W
20	K1	1.0E-05	1.0E+05	S43W
21	K1	1.0E-05	1.0E+05	S217W_Sensor_Gly7, S217W_HAMP_Gly7, S217W
22	K1	1.0E-05	1.0E+05	S217W_Sensor_Gly7
23	α1	1.0E-05	1.0E+05	I207A
24	K2	1.0E-05	1.0E+05	I221F
25	α1	1.0E-05	1.0E+05	I221F
26	K1	1.0E-05	1.0E+05	L210A
27	α1	1.0E-05	1.0E+05	L210A
28	K2	1.0E-05	1.0E+05	L224A

Par#	Par.	lower bound	Upper bound	Fit datasets affected
29	α1	1.0E-05	1.0E+05	L224A
30	K2	1.0E-05	1.0E+05	L224F
31	α1	1.0E-05	1.0E+05	L224F
32	α2	1.0E-08	1.0E+02	L254A_1, L254A_2
33	K3	1.0E-05	1.0E+05	L254A_1, L254A_2
34	α2	1.0E-08	1.0E+02	L258A
35	K3	1.0E-05	1.0E+05	L258A
36	K2	1.0E-05	1.0E+05	N240A
37	K3	1.0E-05	1.0E+05	N255A, N255A_HAMP_Gly7
38	α2	1.0E-08	1.0E+02	N255A
39	α2	1.0E+00	1.0E+00	N255A_HAMP_Gly7
40	K2	1.0E-05	1.0E+05	R236A
41	K2	1.0E-05	1.0E+05	R245F
42	α2	1.0E-08	1.0E+02	R256A
43	K2	1.0E-05	1.0E+05	R256A
44	α2	1.0E-08	1.0E+02	R269L
45	K3	1.0E-05	1.0E+05	R269L
46	K2	1.0E-05	1.0E+05	S217W, S217W_HAMP_Gly7, S217W_Sensor_Gly7
47	α1	1.0E-05	1.0E+05	S217W, S217W_HAMP_Gly7
48	α2	1.0E+00	1.0E+00	S217W_HAMP_Gly7
49	α1	1.0E+00	1.0E+00	S217W_Sensor_Gly7
50	K1	1.0E-05	1.0E+05	S43W
51	K1	1.0E-05	1.0E+05	V191W
52	α2	1.0E-08	1.0E+02	Y265A, Y265A_Sensor_Gly7
53	K3	1.0E-05	1.0E+05	Y265A, Y265A_SHelix_Gly7, Y265A_Sensor_Gly7
54	α2	1.0E-08	1.0E+02	Y265A_SHelix_Gly7
55	α1	1.0E+00	1.0E+00	Y265A_Sensor_Gly7
56	K1	1.0E-05	1.0E+05	Y40W
57	α2	1.0E+00	1.0E+00	Y60C_HAMP_Gly4_1, Y60C_HAMP_Gly4_2, Y60C_HAMP_Gly4_3, Y60C_HAMP_Gly7_1, Y60C_HAMP_Gly7_2, Y60C_HAMP_Gly7_3
58	K2	1.0E-05	1.0E+05	Y60C_HAMP_Gly4_1, Y60C_HAMP_Gly4_2, Y60C_HAMP_Gly4_3, Y60C_HAMP_Gly7_1, Y60C_HAMP_Gly7_2, Y60C_HAMP_Gly7_3
59	α2	1.0E-08	1.0E+02	Y60C_SHelix_Gly7_1, Y60C_SHelix_Gly7_2
60	K3	1.0E-05	1.0E+05	Y60C_SHelix_Gly7_1, Y60C_SHelix_Gly7_2
61	K2	1.0E-05	1.0E+05	Y60C_SHelix_Gly7_1, Y60C_SHelix_Gly7_2
62	α1	1.0E+00	1.0E+00	Y60C_Sensor_Gly7_1, Y60C_Sensor_Gly7_2

References

1. A. M. Stock, V. L. Robinson, P. N. Goudreau, Two-component signal transduction. *Annu. Rev. Biochem.* **69**, 183–215 (2000).

2. E. A. Groisman, C. Mouslim, Sensing by bacterial regulatory systems in host and non-host environments. *Nat. Rev. Microbiol.* **4**, 705–709 (2006).

3. K. Nishino, T. Honda, A. Yamaguchi, Genome-wide analyses of Escherichia coli gene expression responsive to the BaeSR two-component regulatory system. *J. Bacteriol.* **187**, 1763–1772 (2005).

4. H. Hirakawa, K. Nishino, T. Hirata, A. Yamaguchi, Comprehensive studies of drug resistance mediated by overexpression of response regulators of two-component signal transduction systems in Escherichia coli. *J. Bacteriol.* **185**, 1851–1856 (2003).

5. V. Nizet, Antimicrobial peptide resistance mechanisms of human bacterial pathogens. *Curr. Issues Mol. Biol.* **8**, 11–26 (2006).

6. L. Thomas, L. Cook, Two-Component Signal Transduction Systems in the Human Pathogen Streptococcus agalactiae. *Infect. Immun.* **88** (2020).

7. A. Delauné, *et al.*, The WalKR system controls major staphylococcal virulence genes and is involved in triggering the host inflammatory response. *Infect. Immun.* **80**, 3438–3453 (2012).

8. H. U. Ferris, M. Coles, A. N. Lupas, M. D. Hartmann, Crystallographic snapshot of the Escherichia coli EnvZ histidine kinase in an active conformation. *J. Struct. Biol.* **186**, 376–379 (2014).

9. G. Rivera-Cancel, W. Ko, D. R. Tomchick, F. Correa, K. H. Gardner, Full-length structure of a monomeric histidine kinase reveals basis for sensory regulation. *Proc. Natl. Acad. Sci. U. S. A.* **111**, 17839–17844 (2014).

10. C. Wang, *et al.*, Mechanistic insights revealed by the crystal structure of a histidine kinase with signal transducer and sensor domains. *PLoS Biol.* **11**, e1001493 (2013).

11. A. E. Mechaly, N. Sassoon, J.-M. Betton, P. M. Alzari, Segmental helical motions and dynamical asymmetry modulate histidine kinase autophosphorylation. *PLoS Biol.* **12**, e1001776 (2014).

12. A. E. Mechaly, *et al.*, Structural Coupling between Autokinase and Phosphotransferase Reactions in a Bacterial Histidine Kinase. *Struct. Lond. Engl. 1993* **25**, 939-944.e3 (2017).

13. P. Casino, V. Rubio, A. Marina, Structural insight into partner specificity and phosphoryl transfer in two-component signal transduction. *Cell* **139**, 325–336 (2009).

14. D. Albanesi, *et al.*, Structural plasticity and catalysis regulation of a thermosensor histidine kinase. *Proc. Natl. Acad. Sci. U. S. A.* **106**, 16185–16190 (2009).

15. F. Jacob-Dubuisson, A. Mechaly, J.-M. Betton, R. Antoine, Structural insights into the signalling mechanisms of two-component systems. *Nat. Rev. Microbiol.* **16**, 585–593 (2018).

16. T. Krell, *et al.*, Bacterial sensor kinases: diversity in the recognition of environmental signals. *Annu. Rev. Microbiol.* **64**, 539–559 (2010).

17. M. P. Bhate, K. S. Molnar, M. Goulian, W. F. DeGrado, Signal transduction in histidine kinases: insights from new structures. *Struct. Lond. Engl. 1993* **23**, 981–994 (2015).

18. S. I. Miller, A. M. Kukral, J. J. Mekalanos, A two-component regulatory system (phoP phoQ) controls Salmonella typhimurium virulence. *Proc. Natl. Acad. Sci. U. S. A.* **86**, 5054–5058 (1989).

19. E. García Véscovi, F. C. Soncini, E. A. Groisman, Mg2+ as an extracellular signal: environmental regulation of Salmonella virulence. *Cell* **84**, 165–174 (1996).

20. F. C. Soncini, E. García Véscovi, F. Solomon, E. A. Groisman, Molecular basis of the magnesium deprivation response in Salmonella typhimurium: identification of PhoP-regulated genes. *J. Bacteriol.* **178**, 5092–5099 (1996).

21. M. W. Bader, *et al.*, Recognition of antimicrobial peptides by a bacterial sensor kinase. *Cell* **122**, 461–472 (2005).

22. R. E. W. Hancock, J. B. McPhee, Salmonella's sensor for host defense molecules. *Cell* **122**, 320–322 (2005).

23. E. A. Groisman, E. Chiao, C. J. Lipps, F. Heffron, Salmonella typhimurium phoP virulence gene is a transcriptional regulator. *Proc. Natl. Acad. Sci. U. S. A.* **86**, 7077–7081 (1989).

24. P. I. Fields, E. A. Groisman, F. Heffron, A Salmonella locus that controls resistance to microbicidal proteins from phagocytic cells. *Science* **243**, 1059–1062 (1989).

25. I. Behlau, S. I. Miller, A PhoP-repressed gene promotes Salmonella typhimurium invasion of epithelial cells. *J. Bacteriol.* **175**, 4475–4484 (1993).

26. W. J. Belden, S. I. Miller, Further characterization of the PhoP regulon: identification of new PhoP-activated virulence loci. *Infect. Immun.* **62**, 5095–5101 (1994).

27. J. S. Gunn, S. I. Miller, PhoP-PhoQ activates transcription of pmrAB, encoding a two-component regulatory system involved in Salmonella typhimurium antimicrobial peptide resistance. *J. Bacteriol.* **178**, 6857–6864 (1996).

28. L. Guo, *et al.*, Regulation of lipid A modifications by Salmonella typhimurium virulence genes phoP-phoQ. *Science* **276**, 250–253 (1997).

29. B. L. Bearson, L. Wilson, J. W. Foster, A low pH-inducible, PhoPQ-dependent acid tolerance response protects Salmonella typhimurium against inorganic acid stress. *J. Bacteriol.* **180**, 2409–2417 (1998).

30. L. Guo, *et al.*, Lipid A acylation and bacterial resistance against vertebrate antimicrobial peptides. *Cell* **95**, 189–198 (1998).

31. P. Adams, *et al.*, Proteomic detection of PhoPQ- and acid-mediated repression of Salmonella motility. *Proteomics* **1**, 597–607 (2001).

32. M. W. Bader, *et al.*, Regulation of Salmonella typhimurium virulence gene expression by cationic antimicrobial peptides. *Mol. Microbiol.* **50**, 219–230 (2003).

33. Z. D. Dalebroux, S. Matamouros, D. Whittington, R. E. Bishop, S. I. Miller, PhoPQ regulates acidic glycerophospholipid content of the Salmonella Typhimurium outer membrane. *Proc. Natl. Acad. Sci. U. S. A.* **111**, 1963–1968 (2014).

34. U. S. Cho, *et al.*, Metal bridges between the PhoQ sensor domain and the membrane regulate transmembrane signaling. *J. Mol. Biol.* **356**, 1193–1206 (2006).

35. S. Chamnongpol, M. Cromie, E. A. Groisman, Mg2+ sensing by the Mg2+ sensor PhoQ of Salmonella enterica. *J. Mol. Biol.* **325**, 795–807 (2003).

36. L. R. Prost, *et al.*, Activation of the bacterial sensor kinase PhoQ by acidic pH. *Mol. Cell* **26**, 165–174 (2007).

37. J. Choi, E. A. Groisman, Activation of master virulence regulator PhoP in acidic pH requires the Salmonella-specific protein UgtL. *Sci. Signal.* **10** (2017).

38. J. Yuan, F. Jin, T. Glatter, V. Sourjik, Osmosensing by the bacterial PhoQ/PhoP two-component system. *Proc. Natl. Acad. Sci. U. S. A.* **114**, E10792–E10798 (2017).

39. K. S. Molnar, *et al.*, Cys-scanning disulfide crosslinking and bayesian modeling probe the transmembrane signaling mechanism of the histidine kinase, PhoQ. *Struct. Lond. Engl. 1993* **22**, 1239–1251 (2014).

40. L. Aravind, C. P. Ponting, The cytoplasmic helical linker domain of receptor histidine kinase and methyl-accepting proteins is common to many prokaryotic signalling proteins. *FEMS Microbiol. Lett.* **176**, 111–116 (1999).

41. P. Fernández, *et al.*, Transmembrane Prolines Mediate Signal Sensing and Decoding in Bacillus subtilis DesK Histidine Kinase. *mBio* **10** (2019).

42. M. Motz, K. Jung, The role of polyproline motifs in the histidine kinase EnvZ. *PloS One* **13**, e0199782 (2018).

43. N. Akkaladevi, F. Bunyak, D. Stalla, T. A. White, G. L. Hazelbauer, Flexible Hinges in Bacterial Chemoreceptors. *J. Bacteriol.* **200** (2018).

44. S. D. Goldberg, G. D. Clinthorne, M. Goulian, W. F. DeGrado, Transmembrane polar interactions are required for signaling in the Escherichia coli sensor kinase PhoQ. *Proc. Natl. Acad. Sci. U. S. A.* **107**, 8141–8146 (2010).

45. R. Moukhametzianov, *et al.*, Development of the signal in sensory rhodopsin and its transfer to the cognate transducer. *Nature* **440**, 115–119 (2006).

46. P. Fernández, *et al.*, Transmembrane Prolines Mediate Signal Sensing and Decoding in Bacillus subtilis DesK Histidine Kinase. *mBio* **10** (2019).

47. M. Motz, K. Jung, The role of polyproline motifs in the histidine kinase EnvZ. *PloS One* **13**, e0199782 (2018).

48. S. Kitanovic, P. Ames, J. S. Parkinson, A Trigger Residue for Transmembrane Signaling in the Escherichia coli Serine Chemoreceptor. *J. Bacteriol.* **197**, 2568–2579 (2015).

49. C. A. Adase, R. R. Draheim, G. Rueda, R. Desai, M. D. Manson, Residues at the cytoplasmic end of transmembrane helix 2 determine the signal output of the TarEc chemoreceptor. *Biochemistry* **52**, 2729–2738 (2013).

50. I. Gushchin, *et al.*, Mechanism of transmembrane signaling by sensor histidine kinases. *Science* **356** (2017).

51. V. Anantharaman, S. Balaji, L. Aravind, The signaling helix: a common functional theme in diverse signaling proteins. *Biol. Direct* **1**, 25 (2006).

52. R. Keller, *et al.*, The Escherichia coli Envelope Stress Sensor CpxA Responds to Changes in Lipid Bilayer Properties. *Biochemistry* **54**, 3670–3676 (2015).

53. E. Batchelor, D. Walthers, L. J. Kenney, M. Goulian, The Escherichia coli CpxA-CpxR envelope stress response system regulates expression of the porins ompF and ompC. *J. Bacteriol.* **187**, 5723–5731 (2005).

54. P. N. Danese, T. J. Silhavy, The sigma(E) and the Cpx signal transduction systems control the synthesis of periplasmic protein-folding enzymes in Escherichia coli. *Genes Dev.* **11**, 1183–1193 (1997).

55. X. Zhou, *et al.*, Structural basis for two-component system inhibition and pilus sensing by the auxiliary CpxP protein. *J. Biol. Chem.* **286**, 9805–9814 (2011).

56. S. K. D. Leblanc, C. W. Oates, T. L. Raivio, Characterization of the induction and cellular role of the BaeSR two-component envelope stress response of Escherichia coli. *J. Bacteriol.* **193**, 3367–3375 (2011).

57. O. Ashenberg, K. Rozen-Gagnon, M. T. Laub, A. E. Keating, Determinants of homodimerization specificity in histidine kinases. *J. Mol. Biol.* **413**, 222–235 (2011).

58. A. I. Podgornaia, P. Casino, A. Marina, M. T. Laub, Structural basis of a rationally rewired protein-protein interface critical to bacterial signaling. *Struct. Lond. Engl. 1993* **21**, 1636–1647 (2013).

59. J. M. Skerker, *et al.*, Rewiring the specificity of two-component signal transduction systems. *Cell* **133**, 1043–1054 (2008).

60. A. Buschiazzo, F. Trajtenberg, Two-Component Sensing and Regulation: How Do Histidine Kinases Talk with Response Regulators at the Molecular Level? *Annu. Rev. Microbiol.* **73**, 507–528 (2019).

61. N. Ohta, A. Newton, The core dimerization domains of histidine kinases contain recognition specificity for the cognate response regulator. *J. Bacteriol.* **185**, 4424–4431 (2003).

62. M. E. Salazar, M. T. Laub, Temporal and evolutionary dynamics of two-component signaling pathways. *Curr. Opin. Microbiol.* **24**, 7–14 (2015).

63. N. W. Schmidt, G. Grigoryan, W. F. DeGrado, The accommodation index measures the perturbation associated with insertions and deletions in coiled-coils: Application to understand signaling in histidine kinases. *Protein Sci. Publ. Protein Soc.* **26**, 414–435 (2017).

64. N. Akkaladevi, F. Bunyak, D. Stalla, T. A. White, G. L. Hazelbauer, Flexible Hinges in Bacterial Chemoreceptors. *J. Bacteriol.* **200** (2018).

65. M. Inouye, Signaling by transmembrane proteins shifts gears. *Cell* **126**, 829–831 (2006).

66. S. A. Chervitz, J. J. Falke, Molecular mechanism of transmembrane signaling by the aspartate receptor: a model. *Proc. Natl. Acad. Sci. U. S. A.* **93**, 2545–2550 (1996).

67. J. J. Falke, A. H. Erbse, The piston rises again. *Struct. Lond. Engl. 1993* **17**, 1149–1151 (2009).

68. M. V. Milburn, *et al.*, Three-dimensional structures of the ligand-binding domain of the bacterial aspartate receptor with and without a ligand. *Science* **254**, 1342–1347 (1991).

69. E. C. Lowe, A. Baslé, M. Czjzek, S. J. Firbank, D. N. Bolam, A scissor blade-like closing mechanism implicated in transmembrane signaling in a Bacteroides hybrid two-component system. *Proc. Natl. Acad. Sci. U. S. A.* **109**, 7298–7303 (2012).

70. M. V. Airola, K. J. Watts, A. M. Bilwes, B. R. Crane, Structure of concatenated HAMP domains provides a mechanism for signal transduction. *Struct. Lond. Engl. 1993* **18**, 436–448 (2010).

71. K. E. Swain, J. J. Falke, Structure of the conserved HAMP domain in an intact, membrane-bound chemoreceptor: a disulfide mapping study. *Biochemistry* **46**, 13684–13695 (2007).

72. K. E. Swain, M. A. Gonzalez, J. J. Falke, Engineered socket study of signaling through a four-helix bundle: evidence for a yin-yang mechanism in the kinase control module of the aspartate receptor. *Biochemistry* **48**, 9266–9277 (2009).

73. J. S. Parkinson, Signaling mechanisms of HAMP domains in chemoreceptors and sensor kinases. *Annu. Rev. Microbiol.* **64**, 101–122 (2010).

74. V. Stewart, The HAMP signal-conversion domain: static two-state or dynamic three-state? *Mol. Microbiol.* **91**, 853–857 (2014).

75. I. Tews, *et al.*, The structure of a pH-sensing mycobacterial adenylyl cyclase holoenzyme. *Science* **308**, 1020–1023 (2005).

Chapter 2: Exploration of the sequence and signaling landscape of the HAMP domain

Abstract

The HAMP domain of PhoQ has been shown to significantly alter the intrinsic behavior of the adjacent sensor and autokinase domains, potentiating the poor ligand-dependent state transition of the sensor in isolation as well as inhibiting the high kinase activity of the autokinase. In Chapter 1, we suggest that the thermodynamic stability of the four-helix HAMP bundle, as well as the strained coiled-coil connections with adjacent domains is responsible for this modulation. In particular, the thermodynamically favored state of the HAMP was found to be tightly coupled to the phosphatase state of the autokinase domain. We create libraries in which the hydrophobic core of the four-helix HAMP bundle, as well as the junction between the HAMP and autokinase, referred to as the signaling helix (S-Helix) are diversified, and select for sequences that enhance PhoQ kinase activity. We find that destabilizing the four-helix bundle hydrophobic core does indeed lead to higher kinase activity. Furthermore, we find that the wild-type S-Helix sequence is enriched in the high-activity population, along with sequences with high polarity/charge and poor helical propensity. We also detect a coiled-coil interface in the S-Helix that is out of phase with the HAMP coiled-coil register preceding it, and potentially frayed as it approaches the autokinase. Taken together, these observations lend credence to the assertion that the thermodynamically favored HAMP state behaves as a negative allosteric regulator of the autokinase. This repression of autokinase activity is modulated by a frustrated coiled-coil geometry in the S-Helix and alleviated by destabilizing the core HAMP helical bundle structure or

destabilization of the S-Helix itself. Lastly, we employ an unsupervised deep learning method, DeepRT(1), to establish the presence of stronger sequence-activity correlations in the HAMP and explore future directions in machine-learning assisted structure prediction of the HAMP and S-Helix from these functional library screening enrichments.

Introduction

In the previous chapter, we showed that the HAMP domain of PhoQ significantly alters the undesired basal activity of the adjacent sensor and autokinase domains to create a bi-stable full-length protein that is ideally responsive to physiologically relevant ranges of stimuli. The HAMP domain was shown to potentiate the poor ligand binding-dependent state transition of the sensor in isolation, as well as repress the high intrinsic kinase activity of the kinase. This activity was due to the thermodynamically favored HAMP signaling state, 'H_2', being highly coupled to the phosphatase competent 'P' states of the sensor and autokinase domains. Introducing a flexible Glycine linker between the HAMP and the S-Helix of the autokinase relieved this repression of the kinase activity of the autokinase domain.

The sequence of the HAMP suggests that it is capable of multiple parallel four-helix coiled-coil structures, which can be facilitated by consecutive hydrophobic residues in HAMP helix 2. Schmidt *et. al.* (2) showed that aligned HK HAMP sequences have a strong enrichment of hydrophobic residues at positions that can support two core-packing geometries; a canonical 'knobs into holes' packing geometry with hydrophobic residues in *a* and *d* positions form well-packed alternating layers, and a 'complementary x-da' packing geometry, in which the aforementioned consecutive hydrophobic residues (in *g* and *e* positions) swing towards the hydrophobic core through a 26 degree helical rotation (3). This presents a particularly attractive signaling hypothesis in which the conformational states are related through a 'gear-box' helical rotation mechanism (4), and in which the relative stabilities of the two states is finetuned by the identity of the hydrophobic core residues. Other signaling hypotheses, which either involve helical translations ('pistoning' mechanism (5, 6)), changes in helix crossing angles ('scissoring' mechanism (7, 8)), changes in inter-helical distance ('orthogonal displacement' mechanism (9)), or changes in helical and super-helical stability ('dynamic HAMP' mechanism (10–12)) would also depend on the ability of the HAMP hydrophobic core to accommodate these multiple structures. As such, exploring the sequence-function landscape of the HAMP core could help differentiate the strongly coupled 'H_2', and the weakly coupled 'H_1' structural states of the HAMP.

Schmidt *et. al.* (2) further show that the strong hydrophobic heptad repeat found in both the HAMP and the DHp helical bundle of the autokinase is disrupted at the interface between these two four-helix bundles, in an unusually polar linker region known as the Signaling Helix (S-Helix (13)). This region features conserved apparent sequence insertions that put hydrophobic residues out of phase with one another with respect to the heptad-repeat of the coiled-coils on either end. It is suggested that the alternate core packing orders of the HAMP are used to differentially stabilize this disruption in the adjacent S-Helix. What would otherwise be a severe over- or under-twisting of the coiled-coil interface caused by these sequence insertions can be diffusely 'accommodated' by the HAMP domain. Therefore, the strain introduced by the S-Helix is intimately tied to the multi-state equilibrium of the HAMP and should be similarly explored.

In this work, we build three libraries that randomize the hydrophobic core of the HAMP parallel four-helix bundle as well as the S-helix motif that connects this bundle to the conserved autokinase domain. The 11-residue hydrophobic HAMP core of Val, Leu and Ile is varied in two libraries; a conservative substitution library in which these positions are varied to Val, Leu, Ile and Phe (HAMP "ILVF" library), and a more permissive library in which these positions are varied to Val, Leu, Ile, Phe, Ala, Ser, Thr and Pro (HAMP "ILVFSPTA" library). The 10-residue S-helix sequence ('ERERYDKYRT') is allowed to vary to all 20 possible amino acids (and stop codons) in a third library ("S-Helix" library). These libraries are then introduced into a ΔPhoQ strain containing a chromosomal YFP gene reporter for PhoQ activity (mgrB::yfp), sorted for high-PhoQ activity variants, sequenced and compared to the unselected library. We find that high-activity sequence variants are enriched in HAMP core substitutions that destabilize the four-helix bundle, such as the accumulation of smaller residues, and substantial replacements of the native sequence, particularly to Phe, which may produce more molten globule-like conformations. These observations are consistent with the role of the HAMP as a negative regulator of autokinase kinase activity in its native thermodynamically preferred state. In the S-helix, we find that the native sequence is moderately preferred, and sequences of high polarity and low helical propensity are highly preferred. While conservative substitutions based on amino acid physiochemical properties can maintain high kinase activity, a vast majority of substitutions, particularly to apolar residues result in lower kinase activity. We also detect hallmarks of a coiled-coil that is out of phase with the HAMP coiled-coil bundle and frayed closer to the auto-kinase domain enriched for activity, which further supports the assertion that uncoupling from the natively preferred HAMP structure is modulated by the frustrated coiled-coil geometry of the S-Helix and leads to autokinase activation.

Results

Library rationale and construction

We built 3 separate libraries to investigate the energetics of HAMP signaling state equilibria and coupling to autokinase domain. The wild-type HAMP 'core' (x, *a* and *d* positions in HAMP coiled coil heptad) is composed of 8 Leucines, 2 Valines and 1 Isoleucine. The first is a conservative substitution library of the 11 HAMP core hydrophobic positions. We vary all 4 positions to Isoleucine, Leucine, Valine and Phenylalanine, using 'NTT' codons for a theoretical diversity of 4^{11} = 4,194,304 variants. This library will henceforth be referred to as the HAMP 'ILVF' library. The second library also diversifies the same 11 core positions of the HAMP, but includes Serine, Proline, Threonine and Alanine as additional variants using 'NYT' codons, for a theoretical diversity of 8^{11} = 8,589,934,592 variants, and will henceforth be referred to as the HAMP 'ILVFSPTA' library. The final library diversifies the 10-residue wild type sequence of the S-Helix (PhoQ 261-270, 'ERERYDKYRT') to all 20 possible amino acids and stop codon using 'NNK' codons for a theoretical diversity of 20^{10} = 1.024 x 10^{13} variants. Although the theoretical diversities of the ILVFSPTA and S-Helix libraries are well-beyond what can be screened in E. coli, we nonetheless expect a subset of the sequence space (as bottlenecked by library construction) to provide a good representation of the sequence-function relationship of the entire sequence space. After transformation into a cloning-strain intermediate, we had recoveries ranging from 2 x 10^5 – 5 x 10^6 variants, which were deemed appropriately sized for follow-up deep FACS selection and evaluation by next generation sequencing.

Library transformation, growth and selection

It was necessary to passage the constructed library through a cloning-strain (XL10-gold) before introducing it into the competent TIM100 (ΔPhoQ) strain used for sequencing. Many variants in our library are poorly tolerated and benefited from the annealing and super-coiling of the plasmid library as it is passaged through the cloning strain to achieve the desired transformation efficiencies in the KO strain. Transformations were recovered for 90 minutes and plunge frozen until selection experiment. For selection, transformations were thawed on ice, diluted into MOPS minimal media containing 20 mM Mg^{2+} and grown for 4 hours at 37°C. This growth condition should repress the activity of inducible (wild-type like) constructs that have low activity at high Mg^{2+} concentrations, thereby enhancing the sorting of high-activity variants in the selected population by FACS. Due to the relatively small dynamic range of the PhoQ response (5-10 folds), we used a conservate top 1% gate on the mgrB-YFP reporter signal to collect 50K cells with high PhoQ activity from all 3 libraries, henceforth referred to as 'selected' libraries. We additionally collected 5M ungated cells from each population to enable measurement of sequence enrichment under as similar experimental conditions as possible ('unselected' libraries). Collected cells were grown for 4 more hours in LB at 37°C prior to sequencing library building and analysis. Library design and experimental selection is summarized in **Figure 2.1**.

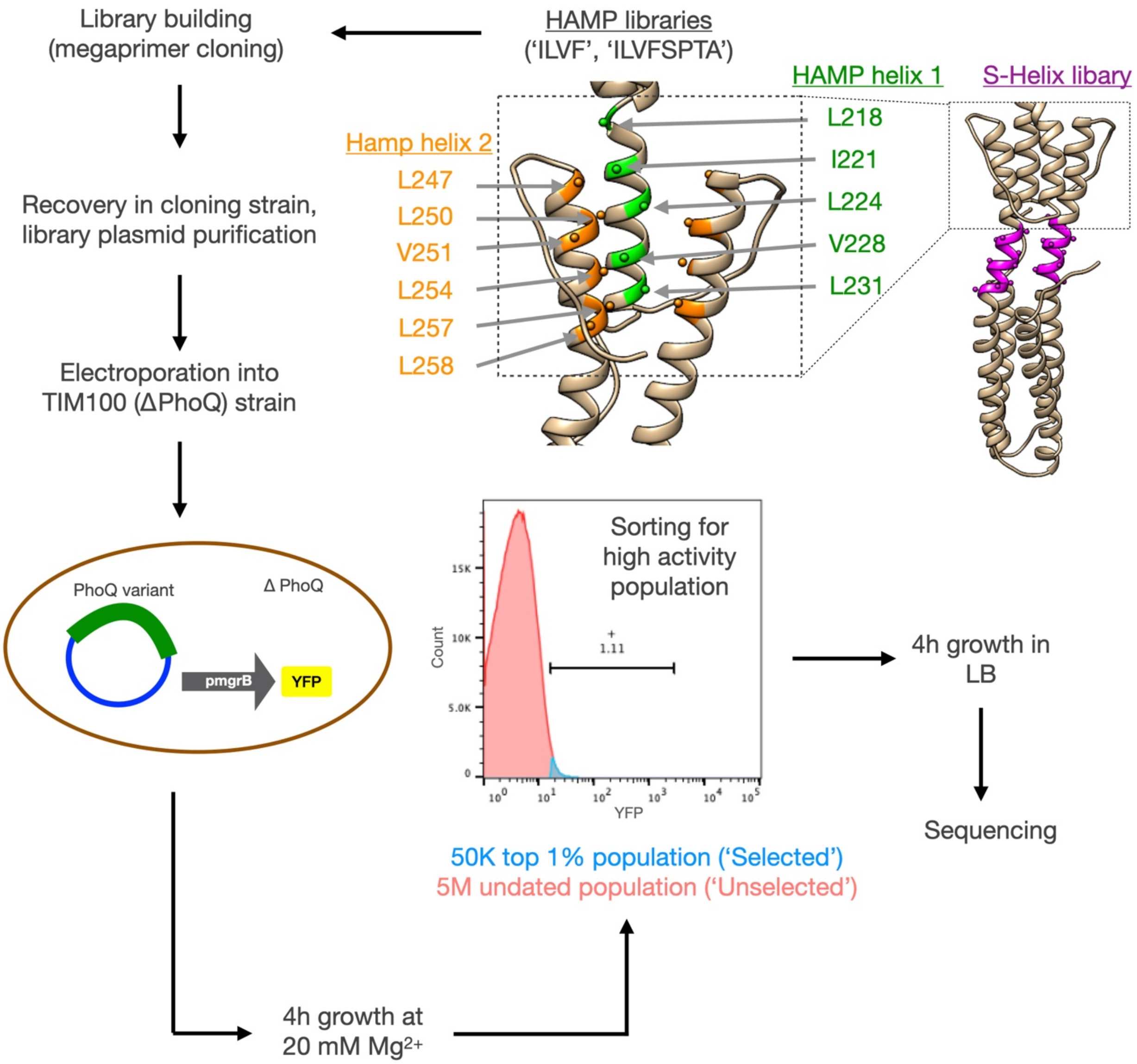


Figure 2.1 – Schematic of PhoQ HAMP and S-Helix library selection experiment. Detailed description of experiment is found in the materials and methods section.

Global outcomes of library sequencing

We built an Illumina paired-end sequencing library of all 3 DNA libraries and both selection criteria, for a total of 6 sequencing libraries, which were then multiplexed into one 100-basepair paired-end HiSeq4000 sequencing run. **Table 2.1** summarizes statistics for the sequencing experiments. Briefly, sequences were paired-end mated, filtered by sequencing quality, translated to protein sequence, and filtered for correct length, and constant PhoQ regions. We recovered 30-48M sequences per sample after all filtering steps. **Figure 2.2** shows the observed counts per unique variants in our library. A minimum sequence count of 10 in selected + unselected populations for each library was used in the analysis moving forward. At that cutoff, we recovered 173,360 unique sequences for ILVF, 134,755 sequences for ILVFSPTA, and 348,417 sequences for S-Helix libraries respectively.

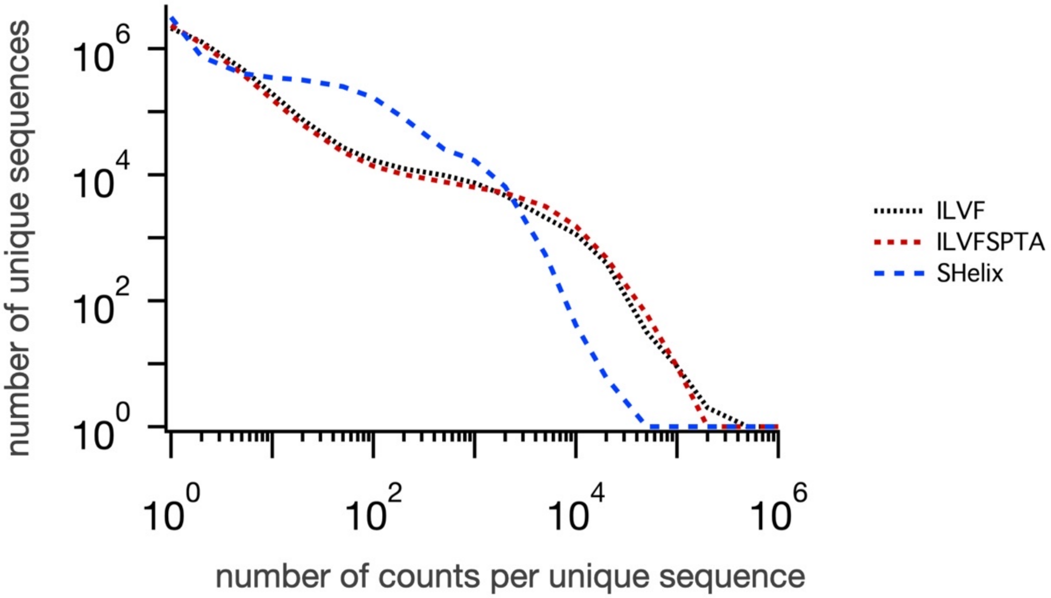


Figure 2.2 – Distribution of counts for unique sequence variants. The number of unique sequences that have a threshold count per sequence (x-axis) in selected + unselected libraries is shown. The S-helix library has more evenly distributed counts per variant than the HAMP libraries (ILVF, ILVFSPTA).

Table 2.2 summarizes the total number of variants observed and total for each substitution level (i.e 1-11 mutations for HAMP libraries, 1-10 mutations for S-Helix libraries), and the associated total sequence counts, whereas **Table 2.3** reports the average number of counts per unique variants for each substitution level. As can be seen, we had a good representation of each number of substitutions possible in each library. **Table 2.3** shows that unique variants have an average count >10, except for a few substitution levels in the S-Helix library. **Table 2.4** compares the coverage of the library parsed by substitution levels. While all substitution levels are adequately covered in the ILVF library (total variant number is 4.2M, which is on par with the maximum library coverage expected given experimental considerations), larger numbers of substitutions are poorly covered in the ILVFSPTA and S-Helix libraries, again owing to the large theoretical diversity of multiply substituted sequences. As such, the few numbers of possible unique variants with low sequence substitutions (≤3 mutations) have good counts per variant and overall coverage of diversity and are more appropriate for comparing to the WT sequence. While we have hundreds of thousands of high sequence substitution variants (≥8 mutations) with good counts per variant for each library, these sequences represent a very small scant sampling of the overall theoretical diversity at these high substitution levels (**Table 2.3**).

We were able to obtain relatively uniform distribution of amino acid variants at all positions in the ILVF and S-Helix libraries. However, the ILVFSPTA library had a significant de-enrichment of small residue substitutions (Ser, Pro, Thr, Ala) relative to large residue substitutions (Ile, Leu, Val, Phe). **Table 2.5** summarizes the linkage-independent sequence counts for the ILVFSPTA library in both selected and unselected populations. As can be observed, number of small variant substitutions was ~5% that of large variant substitutions. This may have to do with the mutagenesis efficiency of changing the wildtype positions (8 Leu, 2 Val, 1 Ile) into the similar codon architectures of I, L, V, F residues (1 codon difference) vs the rather different codon architectures of S, P, T, A residues (2 codons different).

Linkage-independent sequence analysis of HAMP ILVF library

The most direct analysis of sequence enrichment in each library is the linkage-independent variant enrichment calculated for each position of the library, independent of sequence linkage. **Figure 2.3B** shows the enrichment of residues in the ILVF library, with wild-type positions highlighted in grey boxes. The enrichment of WT substitutions per position shows good agreement with that calculated for the exact WT sequence itself (0.96-fold). **Figure 2.3C** shows the clustering of mutation variants by enrichment. The more hydrophobic sidechains, Phenylalanine and Isoleucine cluster together, while Leucine and Valine, with lower hydrophobicities and terminal methyl branches cluster separately. This clustering reflects what one would expect from the physiochemical properties of these variants.

Sequences with large numbers of substitutions may start to diverge from the native structural fold as discussed previously. Therefore, we also repeated this analysis for subsets of the library encoding three or fewer mutations (**Figure 2.6**). The analysis of the ILVF library within this subset largely reflects the conclusions of the linkage-independent sequence analysis itself with respect to the clustering of amino-acids by enrichment. Notably, when we further cluster by variant position, we find that enrichments cluster cleanly by location on either HAMP helices 1 or 2 (**Figure 2.6D**). We also report the linkage independent analysis of single mutants (**Figure 2.4**) and single and double mutant sets (**Figure 2.5**) for completeness.

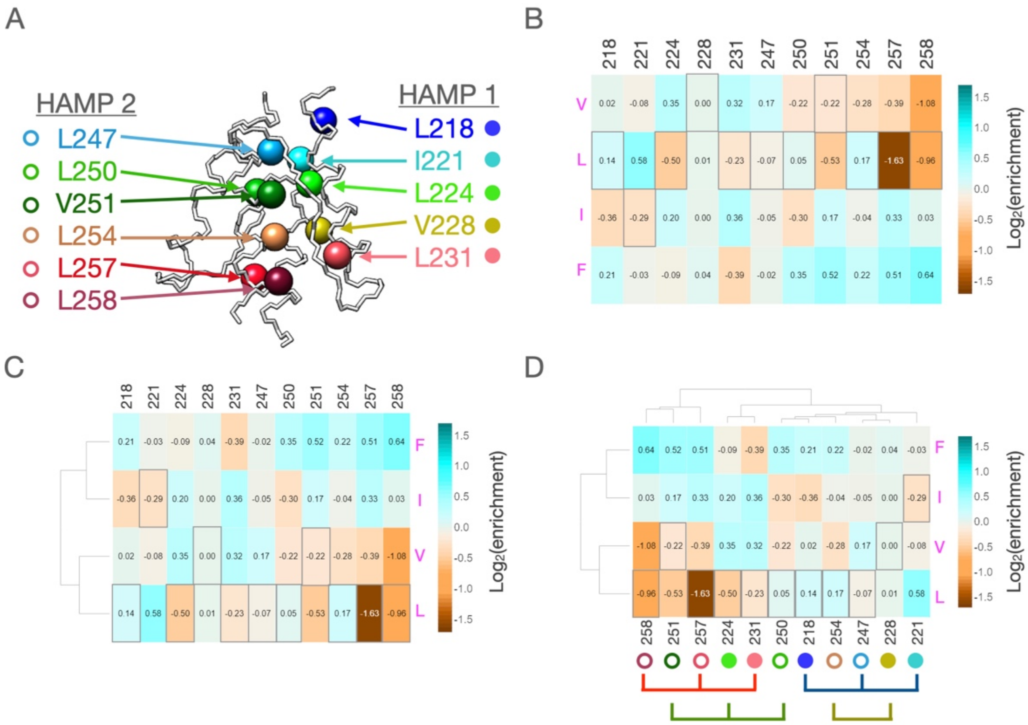


Figure 2.3 – Linkage-independent sequence analysis of HAMP ILVF library. (A) WT sequence and positions of the 'ILVF' library variants are shown on the front monomer of the HAMP domain. Color coding is used in (D) and **Figures 18-25** to indicate position on HAMP helix 1 (solid circles) or HAMP helix 2 (open circles), with colors reflecting depth from N-terminus of bundle **(B)** Linkage independent enrichment was calculated for each position of the HAMP. Values and color scale show log$_2$(enrichment). WT sequence counts were removed to limit bias due to large counts. **(C)** Clustering based on mutation identity shows that amino acids cluster by side-chain size. **(D)** Clustering based on mutation position shows that spatially close mutants generally cluster together.

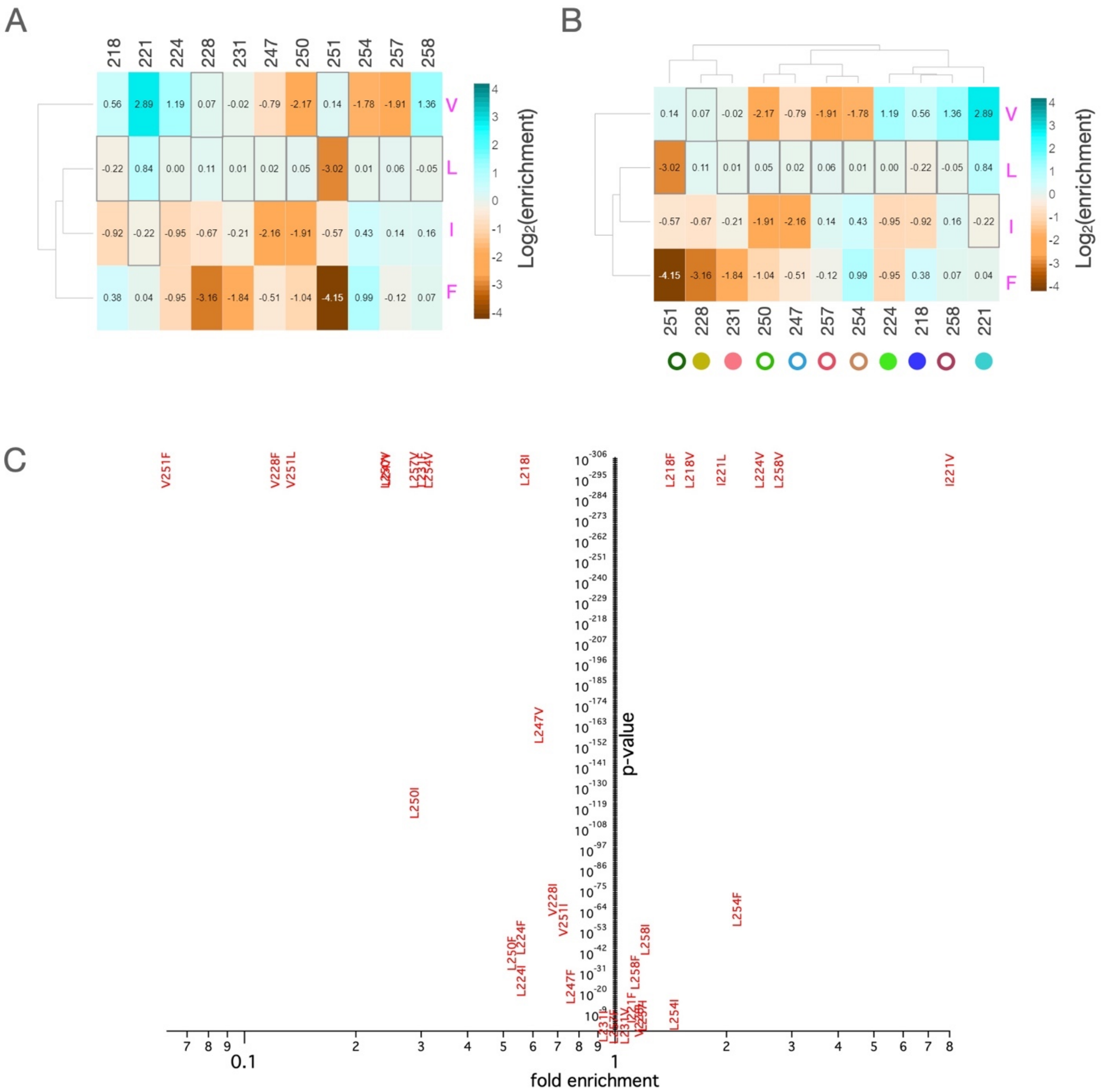


Figure 2.4 – Enrichment of single mutants in HAMP 'ILVF' library. (A) Log$_2$(enrichment) of single mutants is shown. **(B)** Enrichment clustered by mutation identity **(C)** Binomial p-value for enrichment of single mutants is shown. P-values that were too low for calculations were set to 10^{-350}.

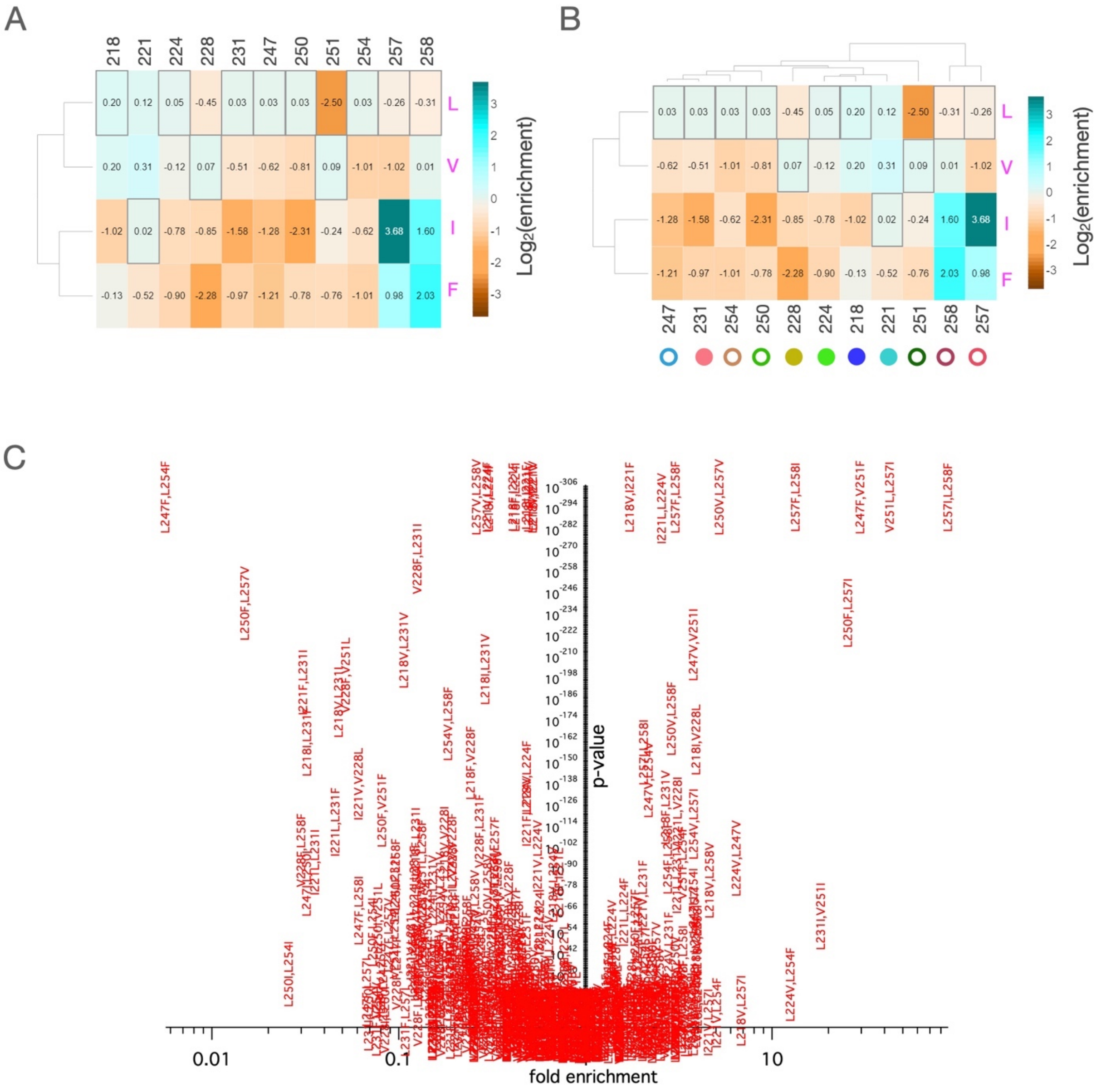
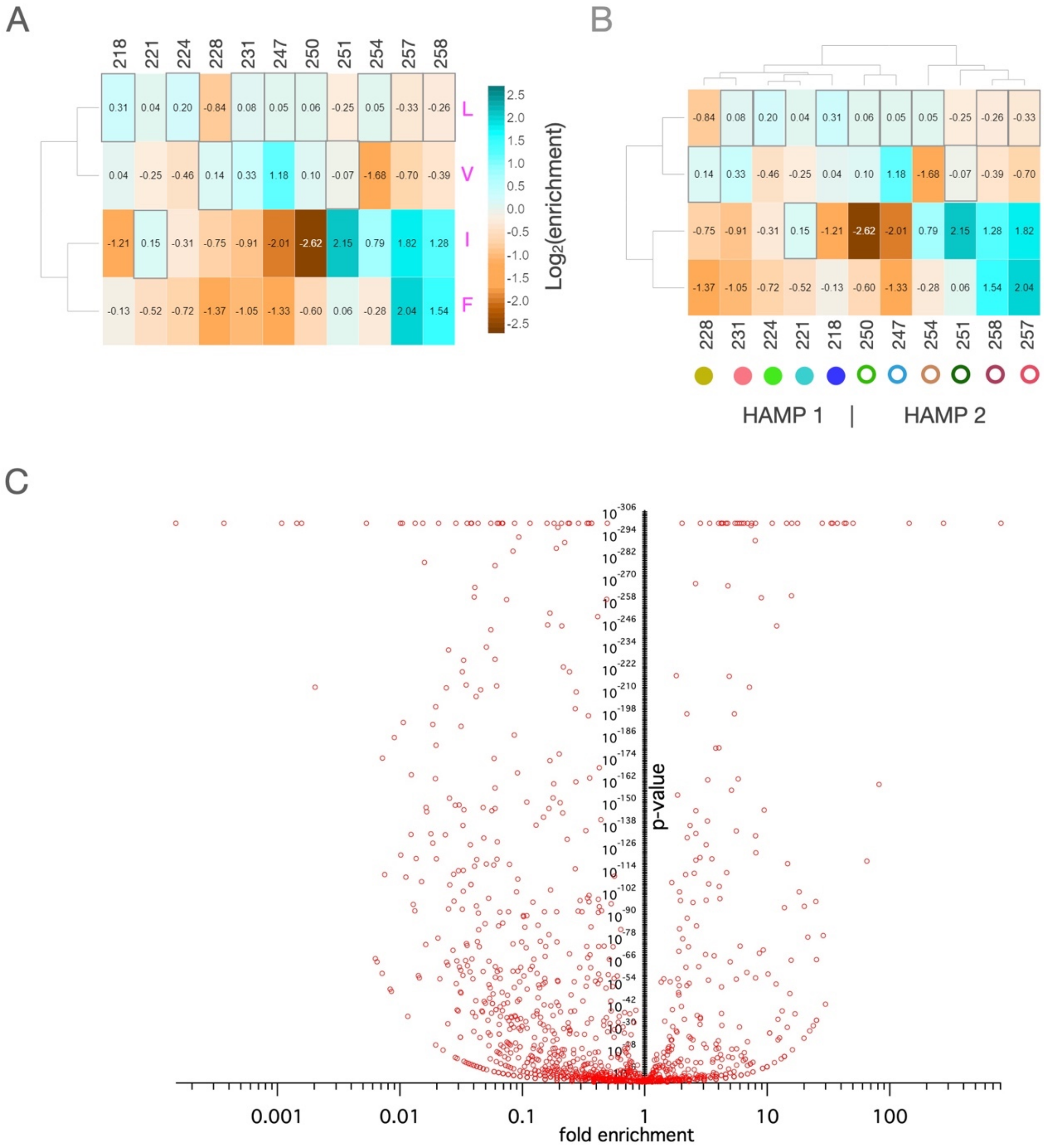


Figure 2.5 – Enrichment of single and double mutants in HAMP 'ILVF' library. (A) Linkage independent log$_2$(enrichment) of single mutant and double mutants is shown for each position. **(B)** Enrichment clustered by mutation identity. **(C)** Binomial p-value for enrichment of double mutants is shown. P-values that were too low for calculations were set to 10^{-350}.



	218	221	224	228	231	247	250	251	254	257	258
L	0.31	0.04	0.20	-0.84	0.08	0.05	0.06	-0.25	0.05	-0.33	-0.26
V	0.04	-0.25	-0.46	0.14	0.33	1.18	0.10	-0.07	-1.68	-0.70	-0.39
I	-1.21	0.15	-0.31	-0.75	-0.91	-2.01	-2.62	2.15	0.79	1.82	1.28
F	-0.13	-0.52	-0.72	-1.37	-1.05	-1.33	-0.60	0.06	-0.28	2.04	1.54

	228	231	224	221	218	250	247	254	251	258	257
L	-0.84	0.08	0.20	0.04	0.31	0.06	0.05	0.05	-0.25	-0.26	-0.33
V	0.14	0.33	-0.46	-0.25	0.04	0.10	1.18	-1.68	-0.07	-0.39	-0.70
I	-0.75	-0.91	-0.31	0.15	-1.21	-2.62	-2.01	0.79	2.15	1.28	1.82
F	-1.37	-1.05	-0.72	-0.52	-0.13	-0.60	-1.33	-0.28	0.06	1.54	2.04

Figure 2.6 - Enrichment of single, double and triple mutants in HAMP 'ILVF' library. (A) Linkage independent $\log_2$(enrichment) is shown for each position for single, double and triple mutants combined. **(B)** Enrichment clustered by mutation identity. **(C)** Binomial p-value for enrichment of double mutants is shown. P-values that were too low for calculations were set to 10^{-350}.

Linkage-independent sequence analysis of the HAMP ILVFSPTA library

The vast majority of sequence substitutions in the ILVFSPTA library were to large hydrophobic residues (Ile, Leu, Val, Phe), as compared to smaller residues (Ser, Pro, Thr, Ala) as discussed above. As such, much of the analysis of this library, particularly as pertaining to large substitutions correlates well with that of the ILVF library. **Figure 2.15A** shows the sequences discovered in common between the ILVF and ILVFSPTA libraries at various total count cut-offs. As shown in **Figure 2.15B**, these common elements show good correlation of enrichment. However, the addition of smaller residue substitutions into this library revealed some interesting observations. **Figure 2.7** shows the linkage-independent sequence enrichment analysis for the ILVFSPTA library. Again, WT sequence substitutions show little to no changes in enrichment, mirroring that observed in the ILVF library. The exact WT sequence itself had an enrichment of 0.85-fold in the ILVFPSTA library, again similar to the 0.96 fold enrichment in the ILVF library. This library however shows a stark contrast between the enrichment of small and large substitutions, with small substitutions clearly being preferred for higher kinase activity across the entire sequence. With the exception of Ile-221, at which small substitutions are strongly de-enriched, small substitutions on both HAMP helices led to higher sequence enrichment and therefore kinase activity. Our clustering analysis was conducted in the presence and absence of the strong de-enrichment of I221T variant in the linkage-independent sequence analysis. When clustered by amino acid variants (in the absence of I221T), we can see a clear clustering of small and large sidechain substitutions (**Figure 2.7C**). This clustering by hydrophobic sidechain size is more or less maintained in the 3-or-fewer substitution subset shown in **Figure 2.10**. Further clustering by variant position in sequence largely shows that positions near the n-terminal and c-terminal halves of the four-helix bundles generally cluster together (**Figure 2.10F**). Position I221 consistently stands out as a position that strongly prefers Leucine, and strongly dis-prefers small residue substitutions, both in the linkage-independent sequence analysis as well as the lower-substitution data subsets. Interestingly, it also often clusters next to L247 position, which is the other position found in the same hydrophobic packing layer.

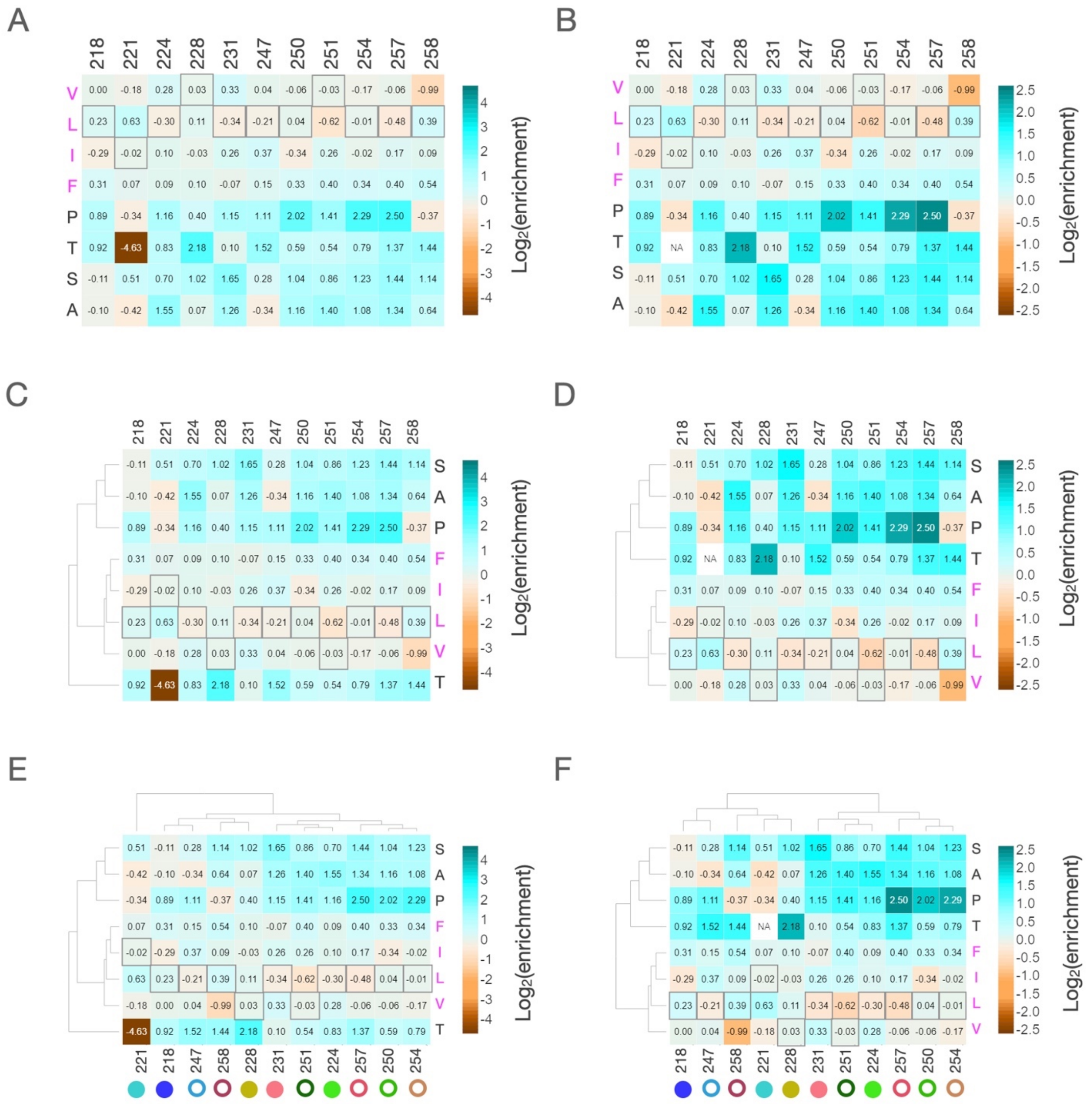


Figure 2.7 - Linkage-independent sequence analysis of HAMP ‘ILVFSPTA’ library. ***(*A)** Linkage independent enrichment was calculated for each position of the HAMP. Values and color scale show log$_2$(enrichment). *WT sequence counts were removed to limit bias due to large counts.* **(B)** I221T position was removed for further clarity. **(C)** Clustering based on mutation identity. **(D)** Clustering enrichment on mutation identity with I221T removed for clarity. Small and large amino acids cluster separately. **(E)** Clustering based on mutation position shows that spatially close mutants cluster together. **(F)** Clustering on mutation position with I221T removed.

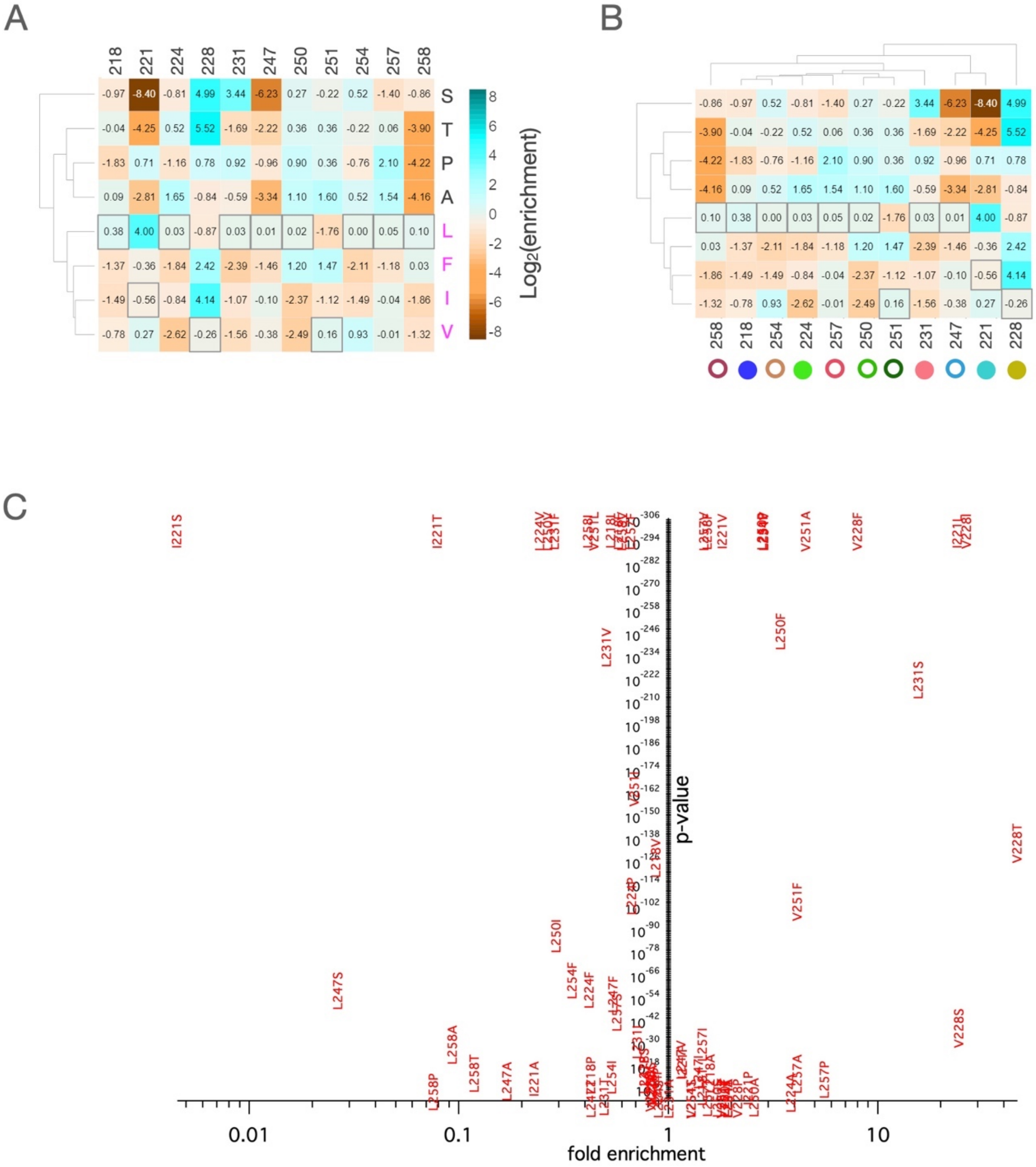


Figure 2.8 - Enrichment of single mutants in HAMP 'ILVFSPTA' library. (A) Log_2(enrichment) of single mutants is shown, with enrichment clustered by mutation identity. **(B)** Enrichment is further clustered by mutation position. **(C)** Binomial p-value for enrichment of single mutants is shown. P-values that were too low for calculations were set to 10^{-350}.

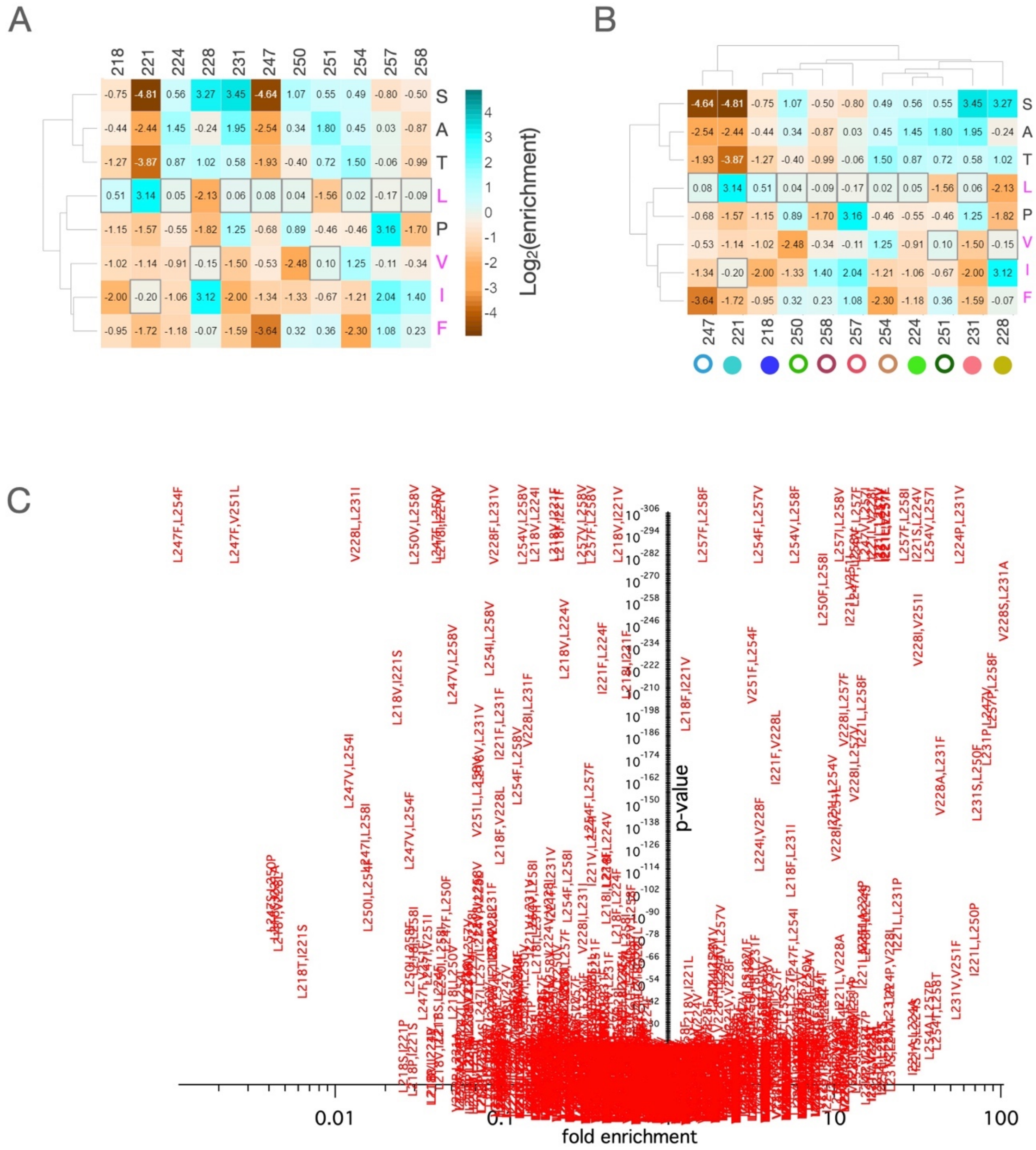


Figure 2.9 - Enrichment of single and double mutants in HAMP 'ILVFSPTA' library. (A) Linkage independent log$_2$(enrichment) of single mutant and double mutants is shown for each position, with enrichment clustered by mutation identity. **(B)** Enrichment is further clustered by mutation position. **(C)** Binomial p-value for enrichment of double mutants is shown. P-values that were too low for calculations were set to 10^{-350}.

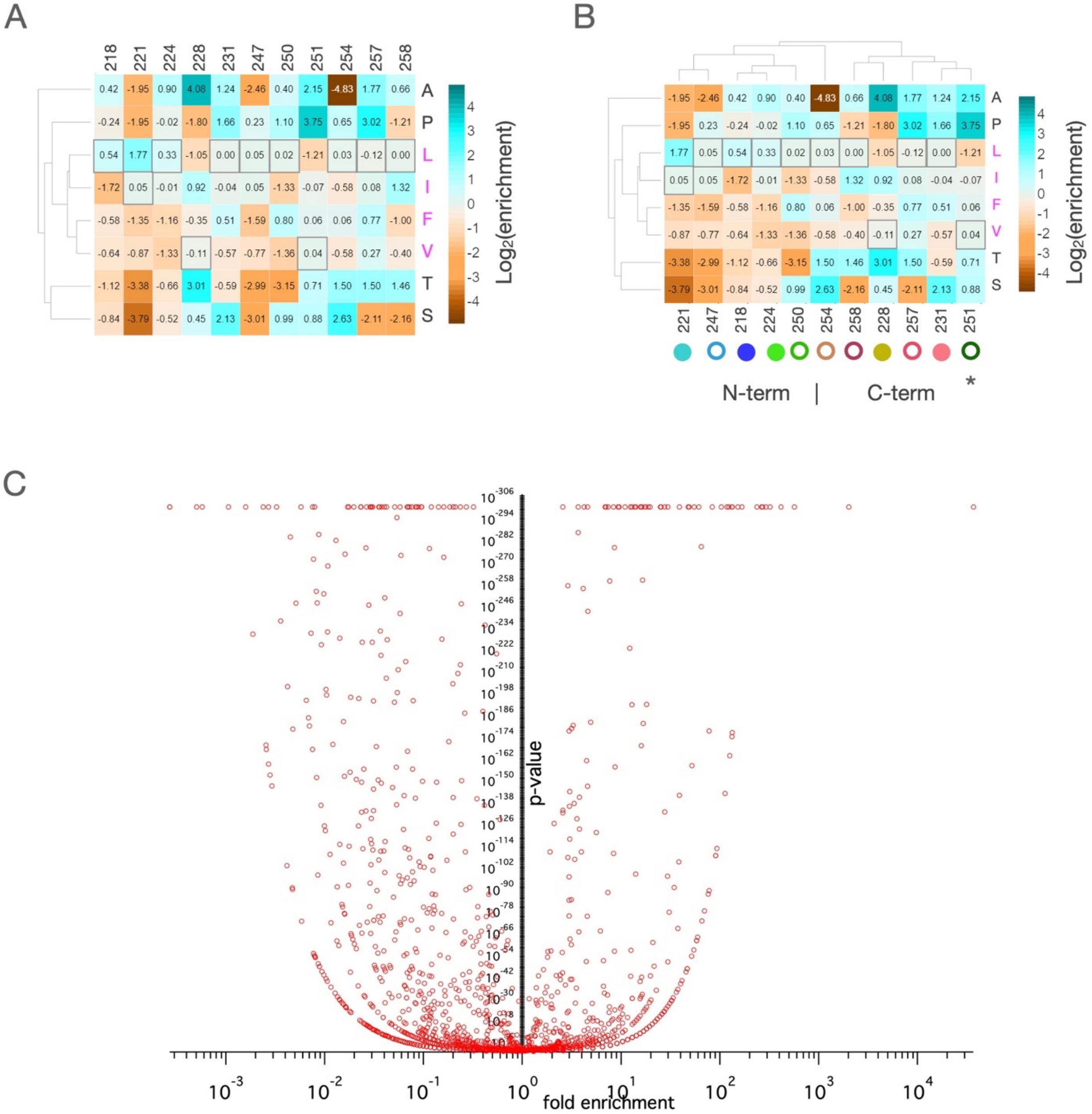


Figure 2.10 - Enrichment of single, double and triple mutants in HAMP 'ILVFSPTA' library. (A) Linkage independent log$_2$(enrichment) is shown for each position for single, double and triple mutants combined, clustered by mutation identity. **(B)** Enrichment is further clustered by mutation position and show clustering by position close to the N-terminal and C-terminal halves of the HAMP bundle with the exception of L251. **C)** Binomial p-value for enrichment of double mutants is shown. P-values that were too low for calculations were set to 10^{-350}.

Linkage-independent sequence analysis of the S-Helix library

Figure 2.11 shows heatmaps for the linkage-independent sequence analysis of the S-Helix library, with colors coding for $\log_2$(enrichment) at each position. Gratifyingly, sequences containing stop codons are de-enriched at all positions, as is expected from constructs that are missing the autokinase domain altogether and therefore cannot produce enhanced kinase-activity. Furthermore, we see a strong pattern of polar/charged substituents resulting in higher enrichment, and large/hydrophobic substituents resulting in de-enrichment. The WT sequence (ERERYDKYRT), which is also notably polar, was found to be enriched by 2.01-fold as well. **Figure 2.11D** shows the clustering of amino-acid variants by enrichment, which results in 2 primary clusters corresponding to charged/polar and apolar substitutions. The apolar cluster is further subclustered into smaller and larger sized sidechain substituents. This indicates that amino acids with similar physiochemical properties can adequately substitute for each other across the S-Helix sequence.

Since the 10 positions of the S-Helix library are contiguous, we also plotted the average enrichment of hydrophobic and polar residues along the sequence length to detect coiled-coil signatures (**Figure 2.12C**). Generally, hydrophobic residue substitutions are de-enriched across the S-Helix (enrichment <1), and Polar residue substitutions are enriched (enrichment >1). We find opposite enrichment trends for these two groups of amino acids, as can be expected for a water-soluble dimer, with hydrophobic residues preferentially enriched in the dimer interface and de-enriched on the surface, and polar residues enriched at the surface and de-enriched at the interface. Interestingly, the peaks of enrichment of hydrophobic residues, positions 262 and 265, do not fit into the hydrophobic heptad repeat pattern of the HAMP domain, and would result in an under-twisting of the HAMP coiled-coil. There is a notable increase in average enrichment of polar residues (and de-enrichment of hydrophobic residues) as we approach the autokinase.

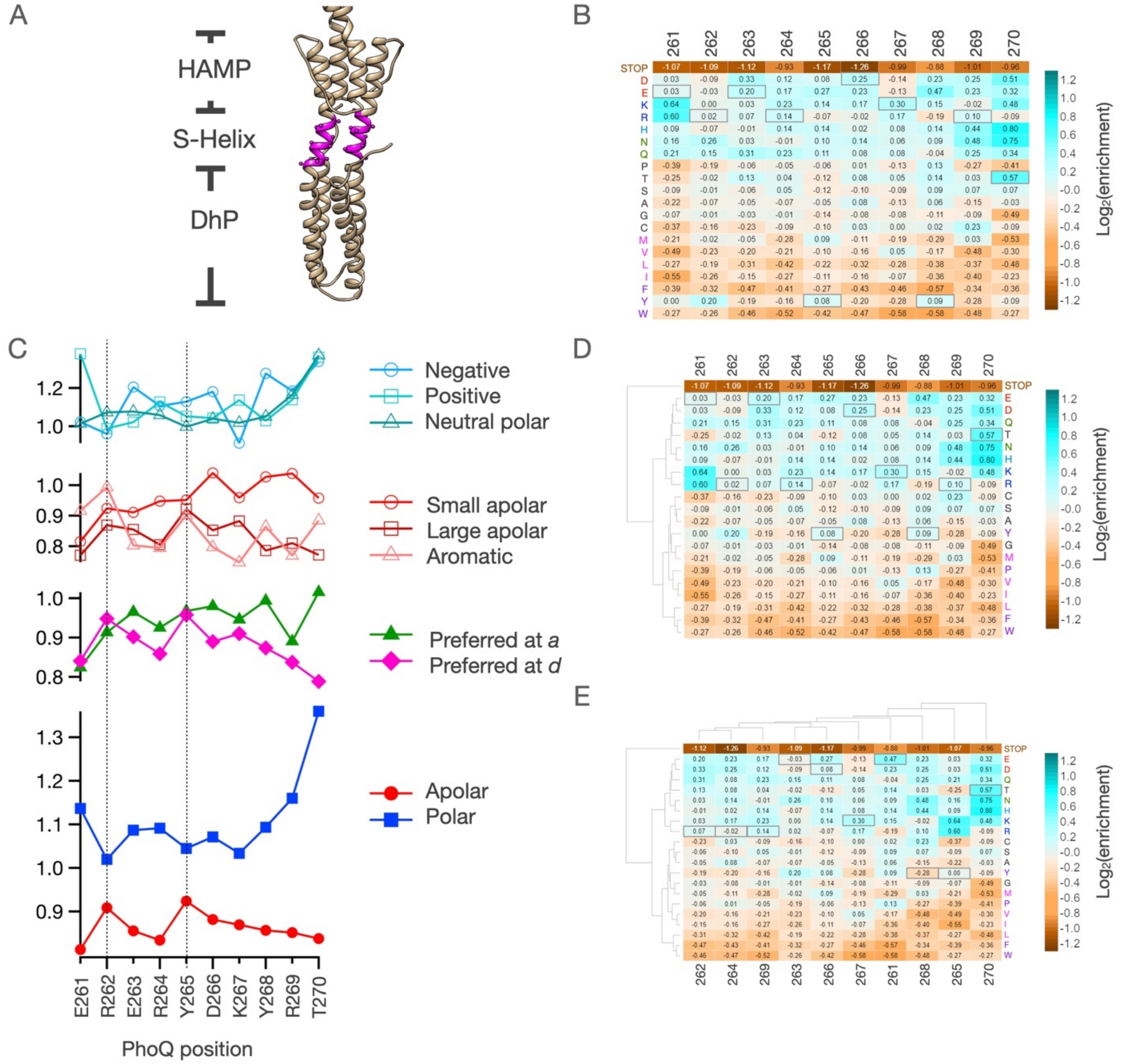


Figure 2.11 - Linkage-independent sequence analysis of S-Helix library. (A) The 10-residue long S-helix is shown in magenta between the HAMP and DhP four-helix bundles of PhoQ. **(B)** Linkage independent enrichment was calculated for each position of the S-Helix. Values and color scale show log$_2$(enrichment). Stop codons were depleted across the dataset as expected. **(C)** Average enrichment scores were calculated for subsets of amino acid types and plotted against S-Helix position. The S-Helix is generally enriched for polar substituents and de-enriched for apolar substituents. Polar and apolar residues show anti-correlated enrichments peaking at positions 262 and 265. **(D)** Clustering based on mutation identity shows that amino acids cluster by side-chain physiochemical properties. **(E)** Clustering based on mutation position show similar WT amino acid type positions cluster together.

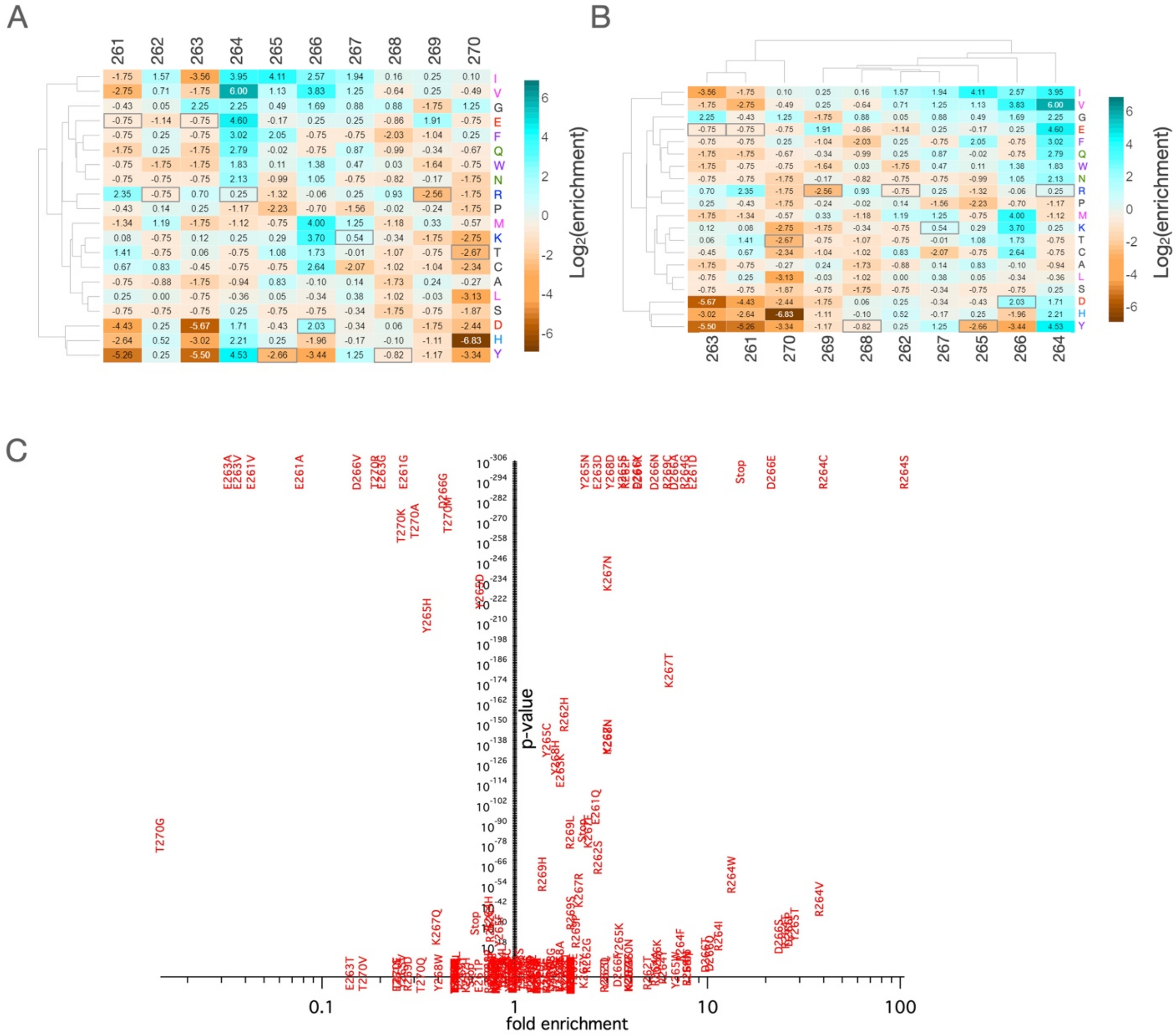


Figure 2.12 - Enrichment of single mutants in S-Helix library. (A) Log_2(enrichment) of single mutants is shown with enrichment clustered by mutation identity. **(B)** Enrichment is further clustered by mutation position. **(C)** Binomial p-value for enrichment of single mutants is shown. P-values that were too low for calculations were set to 10^{-350}.

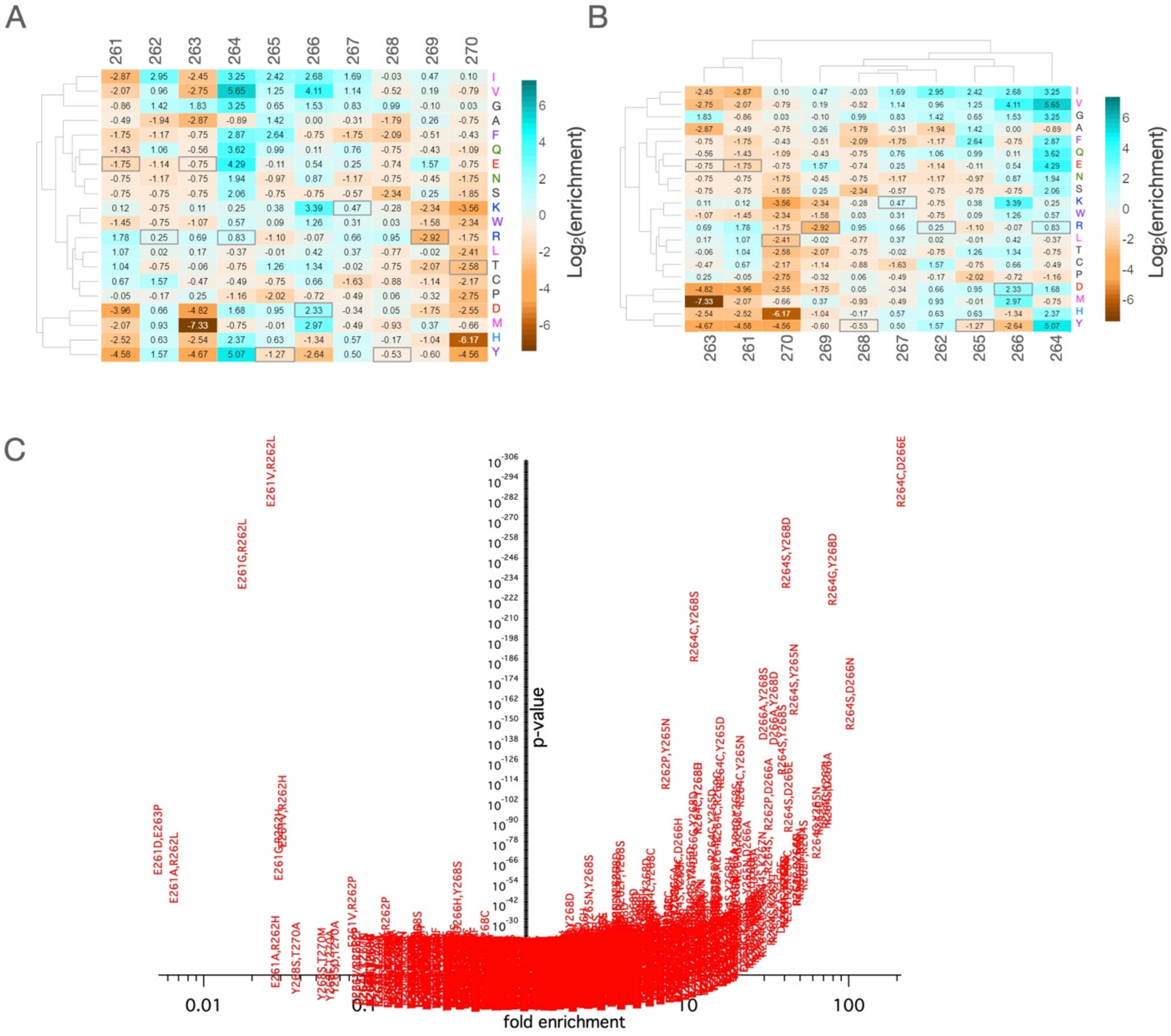


Figure 2.13 - Enrichment of single and double mutants in S-Helix library. (A) Linkage independent $\log_2$(enrichment) of single mutant and double mutants is shown for each position, with enrichment clustered by mutation identity. **(B)** Enrichment is further clustered by mutation position. **(C)** Binomial p-value for enrichment of double mutants is shown. P-values that were too low for calculations were set to 10^{-350}.

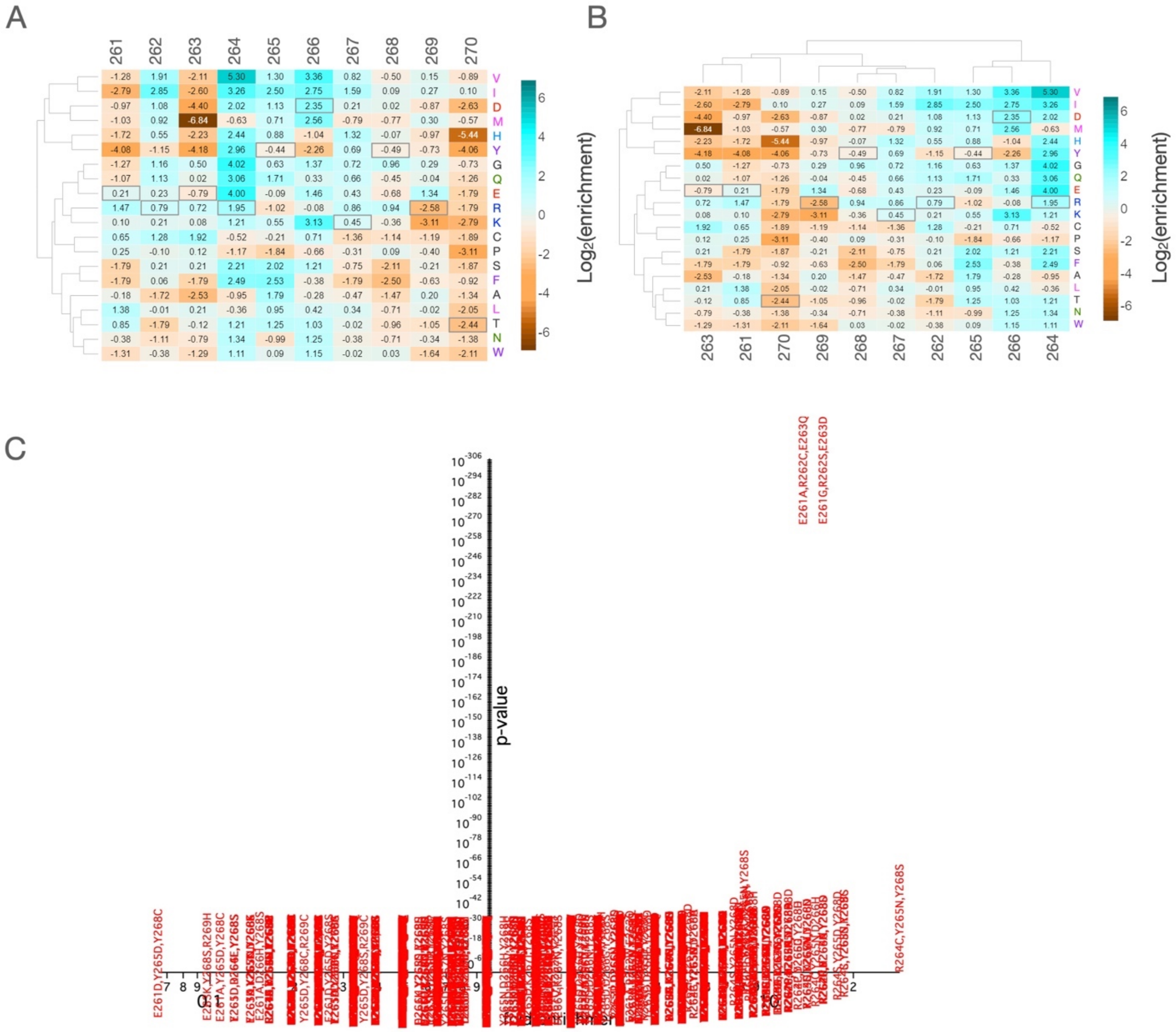


Figure 2.14 - Enrichment of single, double and triple mutants in S-Helix library. (A) Linkage independent $\log_2$(enrichment) is shown for each position for single, double and triple mutants combined, clustered by mutation identity. **(B)** Enrichment is further clustered by mutation position. **(C)** Binomial p-value for enrichment of double mutants is shown. P-values that were too low for calculations were set to 10^{-350}.

Correlation of sequence enrichment with amino acid physiochemical properties

Our global sequence analyses show that the primary determinants of enrichment as well as clustering of substitution profiles by library position seem to reflect fundamental physiochemical properties of the amino acid variants. As such, we examined how much of the variance in our enrichment data is explained by simple properties, such as total number of mutations, sequence polarity, hydrophobicity and alpha-helical propensity. For the 2 HAMP libraries, we also examined both the total and layer-by-layer hydrophobic core volume, whereas for the S-Helix library, we additionally examined the effect of total charge. These analyses were conducted for sequences with total counts ≥10 to reduce overestimation of correlation from low-count sequence variants, with a subset of such sequences with 3 or fewer mutations analyzed separately. Property values for each amino-acid variant were simply totaled across all 20 (S-Helix) or 22 (HAMP libraries) variant positions in the PhoQ dimer.

Table 2.7 summarizes the physiochemical properties and values used for each amino-acid type. **Figures 2.16-2.19** show correlations between physiochemical properties and library enrichment for the ILVF library. Heatmaps are kernel density representations used to better show density of the scatter plots. The spearman correlation of each plot is reported above, as well as listed in **Table 2.6,** along with Pierson correlation coefficients. While we do not see any strong monotonic relationships to isolated properties, as indicated by low Spearman correlations, we can observe that larger numbers of mutations per sequence generally lead to greater effect sizes (**Figure 2.18B**). The lack of strong correlation with gross properties may be unsurprising in the ILVF library given the similarity of these amino acid variants

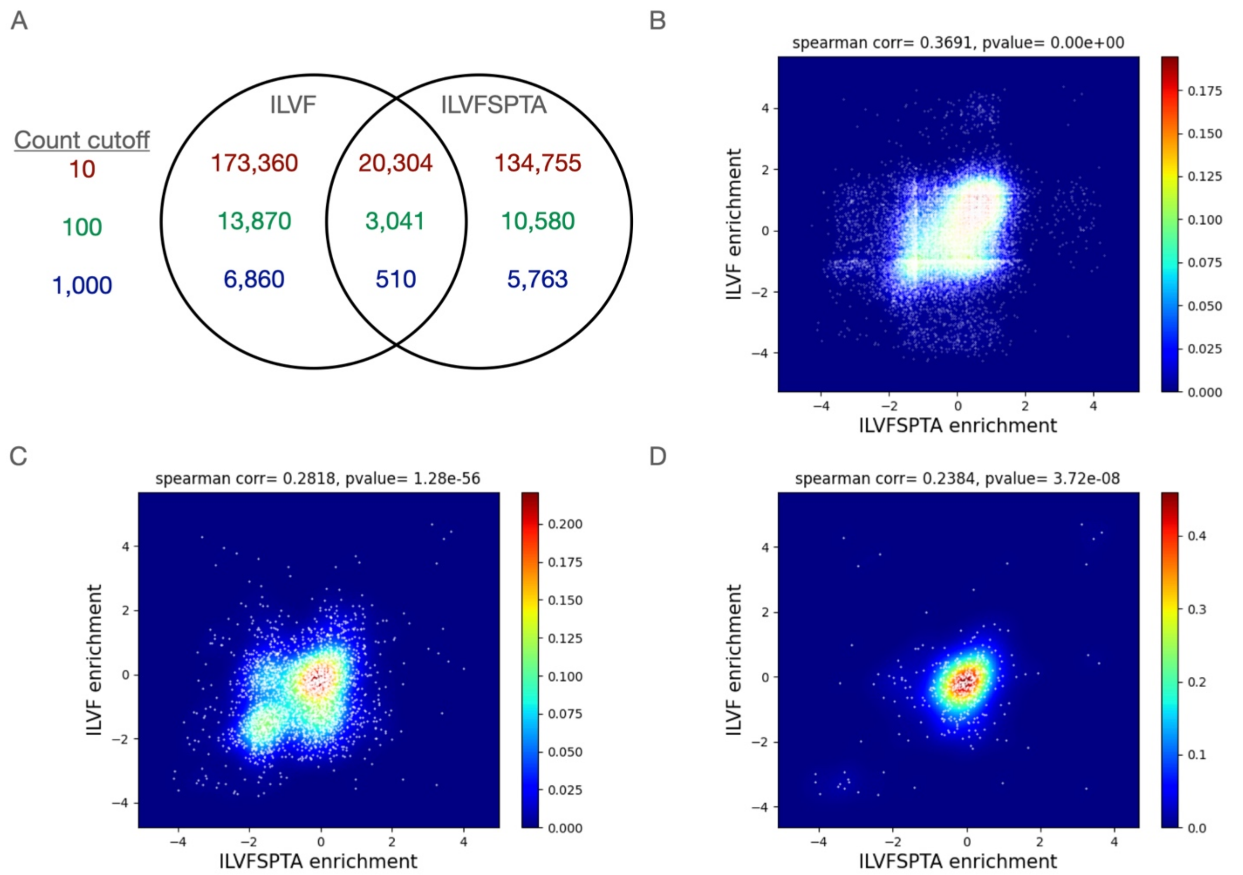


Figure 2.15 – Correlation between HAMP 'ILVF' and 'ILVFSPTA' libraries. (A) Venn diagram showing the number of unique sequences discovered for 'ILVF' and 'ILVFSPTA' libraries with a sequence-per-count threshold of 10, 100 and 1000 counts. **(B)** scatter plot of enrichment for sequences found in common between 'ILVF' and 'ILVFSPTA' libraries with a minimum of 10 counts per sequence in selected + unselected samples (n=20,340). Heat map shows kernel density of scatter plot for clarity. Spearman correlation coefficient and p-value of coefficient are shown above graph. **(C)** same plot as (B) with a minimum of 100 counts per sequence (n=3,041). **(D)** same plot as (B) with a minimum of 1000 counts per sequence (n=510).

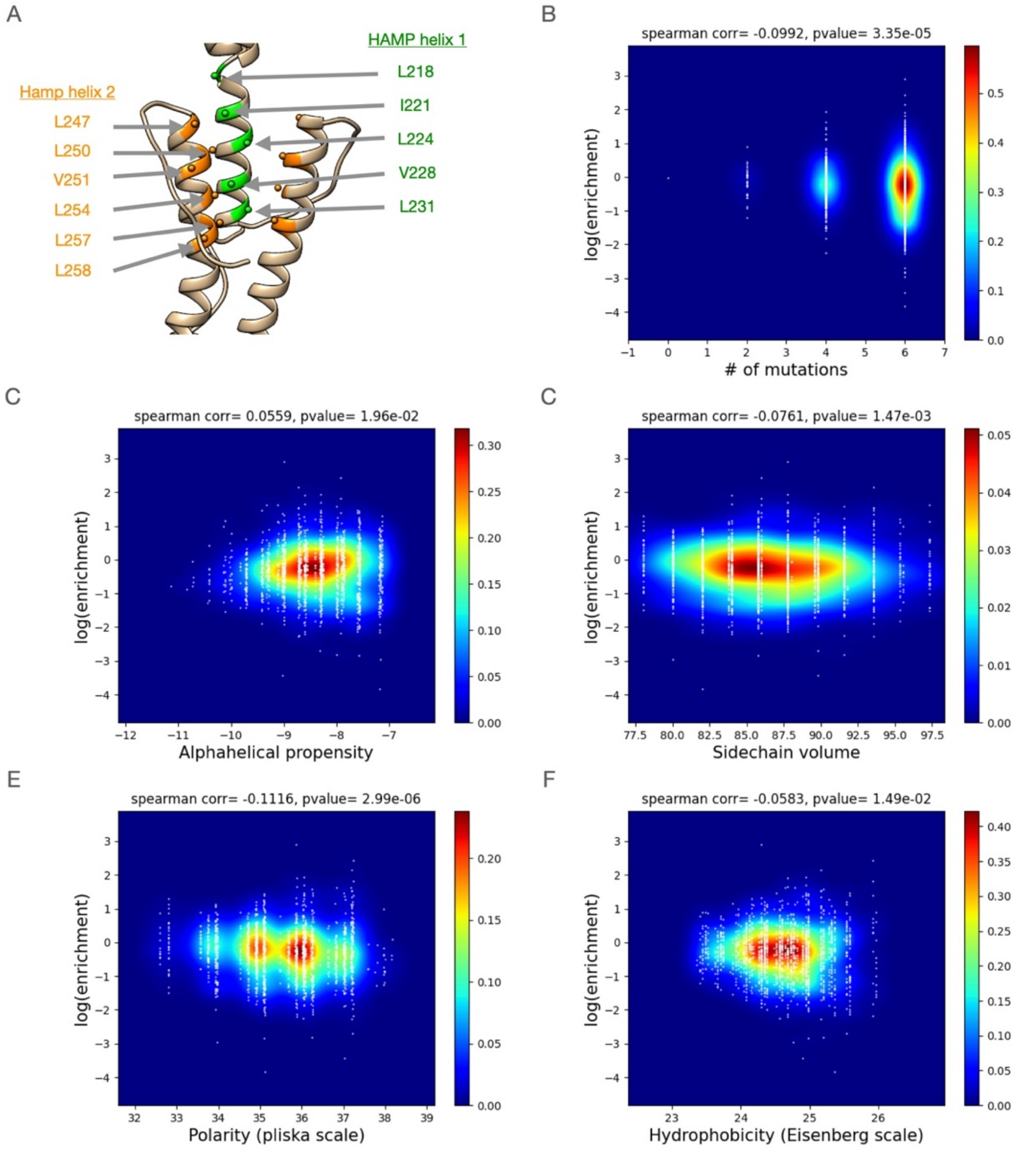


Figure 2.16 – 'ILVF' library sequence enrichment vs. amino acid physiochemical properties (3 mutations or fewer). (A) WT sequence and positions of the 'ILVF' library variants is shown in green (HAMP helix 1) and orange (HAMP helix 2) spheres. The front HAMP helix 1 has been removed for clarity. **(B)** log(enrichment) vs. number of substitutions in sequence for all unique variants with 3 or fewer mutations and a threshold of 10 counts in selected + unselected samples summed across all 11 positions on both monomer chains. Heat map shows kernel density of scatter plot for clarity. Spearman correlation coefficient and p-value of coefficient are shown above graph. **(C)** same plot as (B) for log(enrichment) vs alpha-helical propensity. **(D)** sample plot as (B) for log(enrichment) vs. sidechain volume. **(E)** same plot as (B) for log(enrichment) vs. side-chain polarity (Pliska scale) shows a weak correlation between increasing polarity and increasing enrichment. **(F)** same plot as (B) for log(enrichment) vs side-chain hydrophobicity (Eisenberg scale).

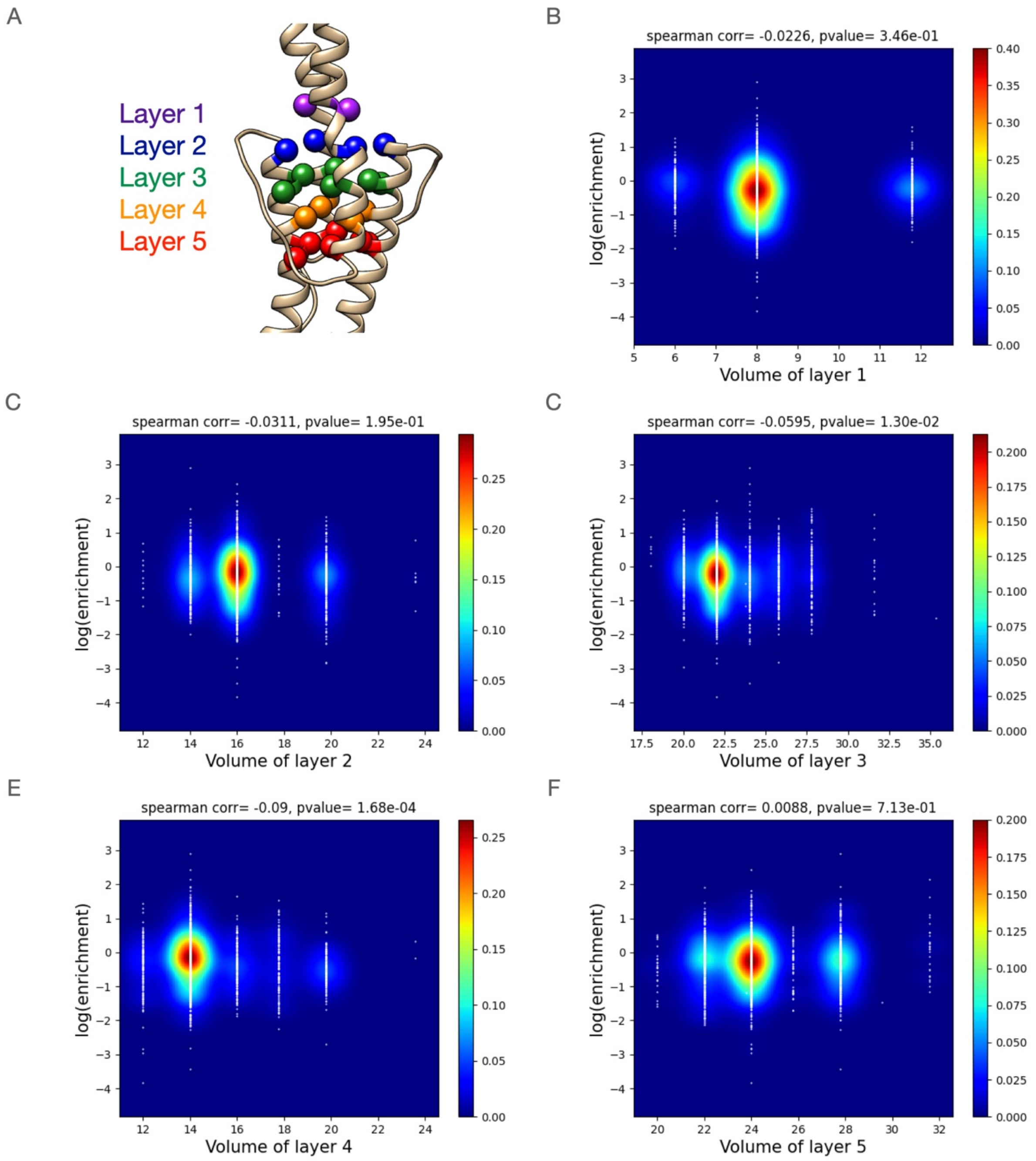


Figure 2.17 - Side-chain volume per hydrophobic packing layer of HAMP 'ILVF' library (3 or fewer mutations). (A) The 5 hydrophobic packing layers of the HAMP domain are shown, including a single layer for the n-terminal 2-helix bundle (layer 1) and four layers of the 4-helix parallel bundle (layers 2-5). **(B)** sum of side-chain volumes for layer 1 vs log(enrichment) for all unique sequences with 3 or fewer mutations and at least 10 counts in selected + unselected samples. Heat map shows kernel density of scatter plot for clarity. Spearman correlation coefficient and p-value of coefficient are shown above graph. **(C)** same plot as (B) for layer 2. **(D)** same plot as (B) for layer 3. **(E)** same plot as (B) for layer 4. **(F)** same plot as (B) for layer 5.

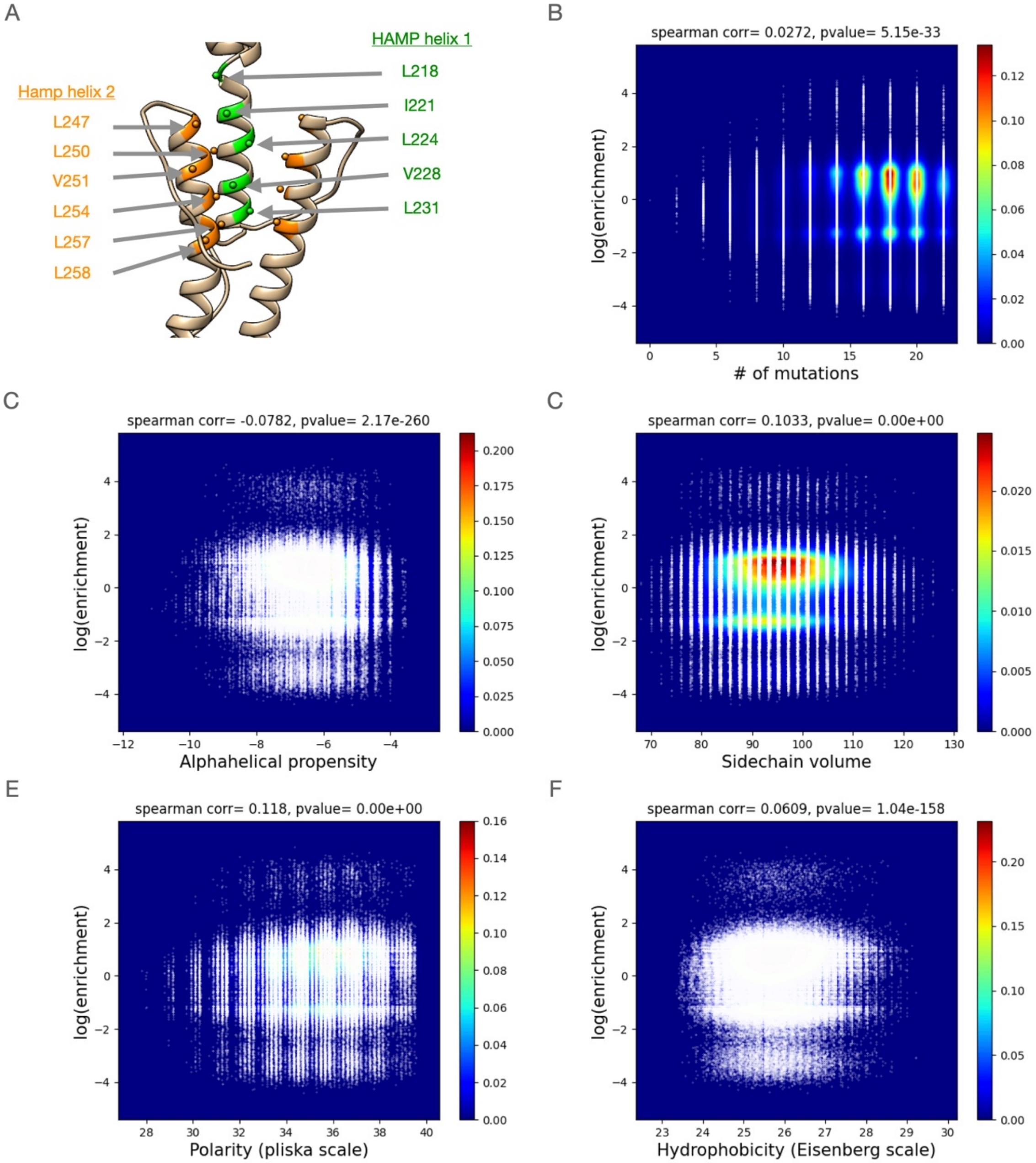


Figure 2.18 - 'ILVF' library sequence enrichment vs. amino acid physiochemical properties (all mutations). (A) WT sequence and positions of the 'ILVF' library variants is shown in green (HAMP helix 1) and orange (HAMP helix 2) spheres. The front HAMP helix 1 has been removed for clarity. **(B)** log(enrichment) vs. number of substitutions in sequence for all unique variants with 3 or fewer mutations and a threshold of 10 counts in selected + unselected samples summed across all 11 positions on both monomer chains. Heat map shows kernel density of scatter plot for clarity. Spearman correlation coefficient and p-value of coefficient are shown above graph. **(C)** same plot as (B) for log(enrichment) vs alpha-helical propensity. **(D)** sample plot as (B) for log(enrichment) vs. sidechain volume. **(E)** same plot as (B) for log(enrichment) vs. side-chain polarity (Pliska scale). **(F)** same plot as (B) for log(enrichment) vs side-chain hydrophobicity (Eisenberg scale).

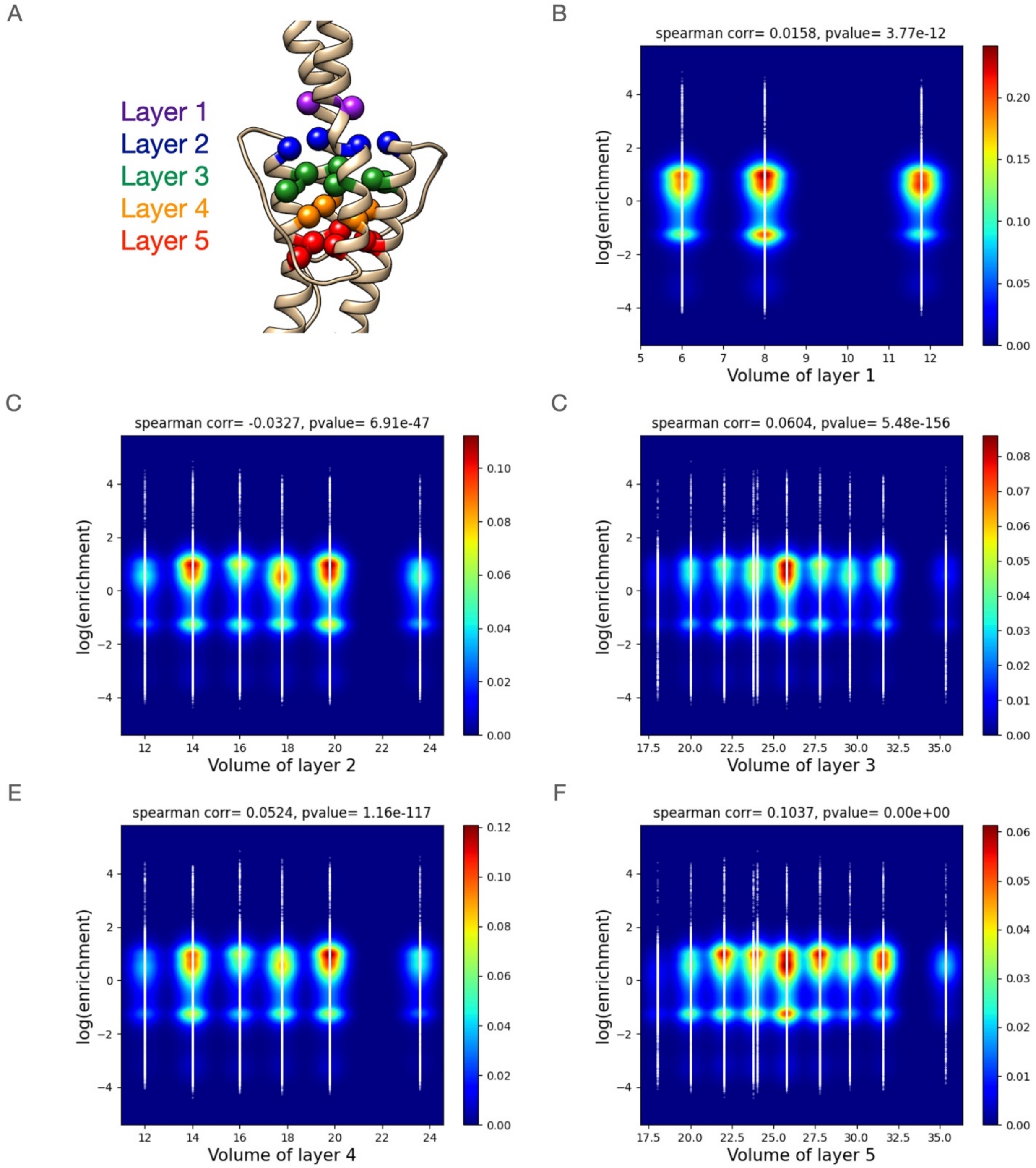


Figure 2.19 - Side-chain volume per hydrophobic packing layer of HAMP 'ILVF' library (all mutations). (A) The 5 hydrophobic packing layers of the HAMP domain are shown, including a single layer for the n-terminal 2-helix bundle (layer 1) and four layers of the 4-helix parallel bundle (layers 2-5). **(B)** sum of side-chain volumes for layer 1 vs log(enrichment) for all unique sequences with 3 or fewer mutations and at least 10 counts in selected + unselected samples. Heat map shows kernel density of scatter plot for clarity. Spearman correlation coefficient and p-value of coefficient are shown above graph. **(C)** same plot as (B) for layer 2. **(D)** same plot as (B) for layer 3. **(E)** same plot as (B) for layer 4. **(F)** same plot as (B) for layer 5.

The ILVFSPTA library shows stronger, albeit still moderate, correlations between physiochemical properties and library enrichment, likely owing to the increased amino acid diversity with respect to the aforementioned properties. Again, larger number of substitutions in the sequence generally lead to larger effect sizes (**Figure 2.22B**). Additionally, we observe moderate correlations in the subset of 3 or fewer mutations. In the subset of ≤3 mutations, there is a moderate correlation between sequence alpha-helical propensity and enrichment, with less alpha-helical sequences resulting in higher sequence enrichment (**Figure 2.20C**). Additionally, sequences that have lower overall hydrophobicities (Eisenberg scale) are moderately correlated with higher sequence enrichment (**Figure 2.20F**). No correlation was observed with other physiochemical properties. It is important to note that the contribution to the computed correlations from sequences with small residue substitutions is muted compared to those with large residue substitutions, again owing to the biased composition of this library (**Table 2.5**). This may generally lead to a lower estimation of the overall variance that is explained by a particular property.

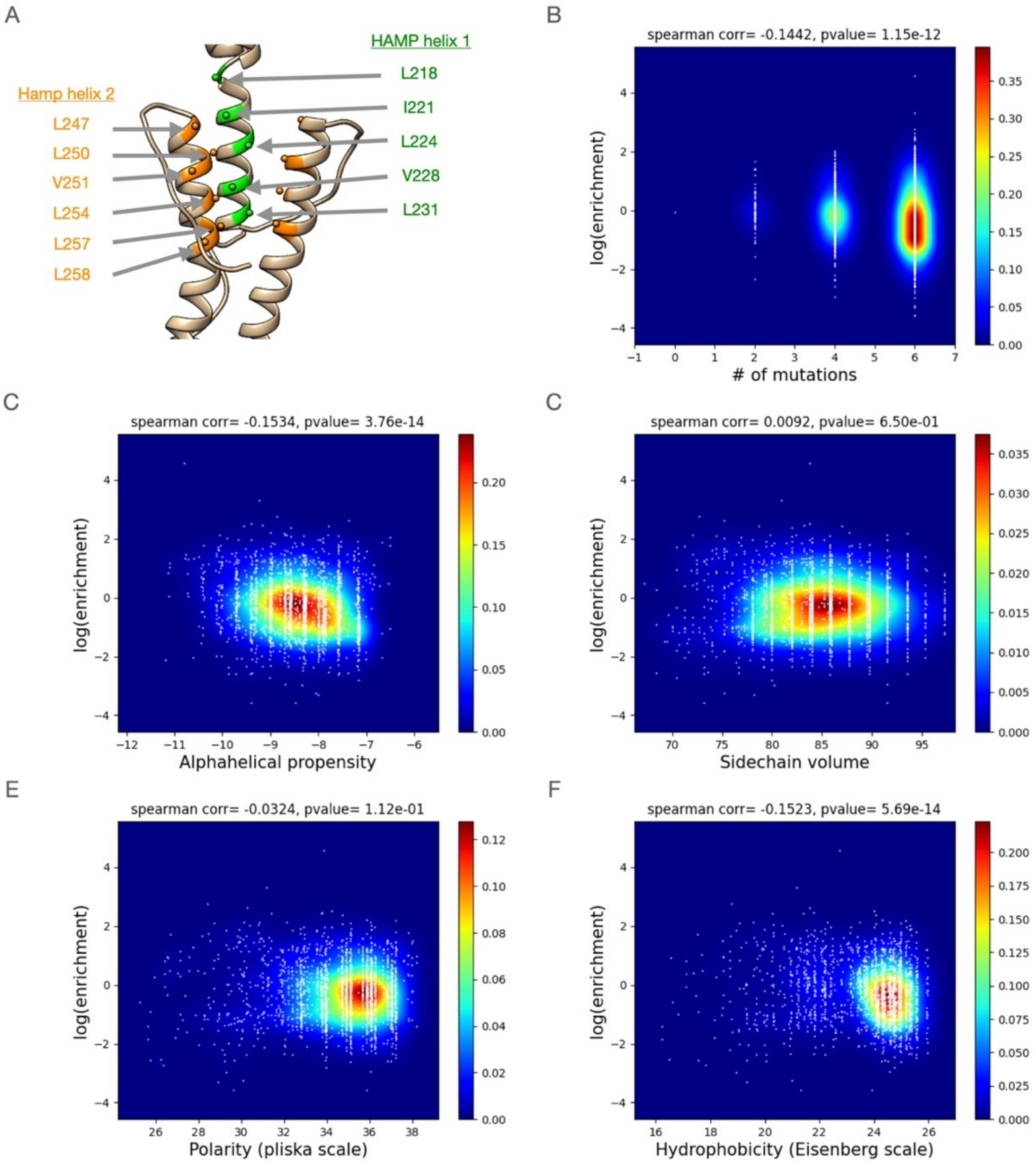


Figure 2.20 - 'ILVFSPTA' library sequence enrichment vs. amino acid physiochemical properties (3 or fewer mutations). (A) WT sequence and positions of the 'ILVFSPTA' library variants is shown in green (HAMP helix 1) and orange (HAMP helix 2) spheres. The front HAMP helix 1 has been removed for clarity. **(B)** log(enrichment) vs. number of substitutions in sequence for all unique variants with 3 or fewer mutations and a threshold of 10 counts in selected + unselected samples summed across all 11 positions on both monomer chains. Heat map shows kernel density of scatter plot for clarity. Spearman correlation coefficient and p-value of coefficient are shown above graph. **(C)** same plot as (B) for log(enrichment) vs alpha-helical propensity. **(D)** sample plot as (B) for log(enrichment) vs. sidechain volume. **(E)** same plot as (B) for log(enrichment) vs. side-chain polarity (Pliska scale) shows a weak correlation between increasing polarity and increasing enrichment. **(F)** same plot as (B) for log(enrichment) vs side-chain hydrophobicity (Eisenberg scale).

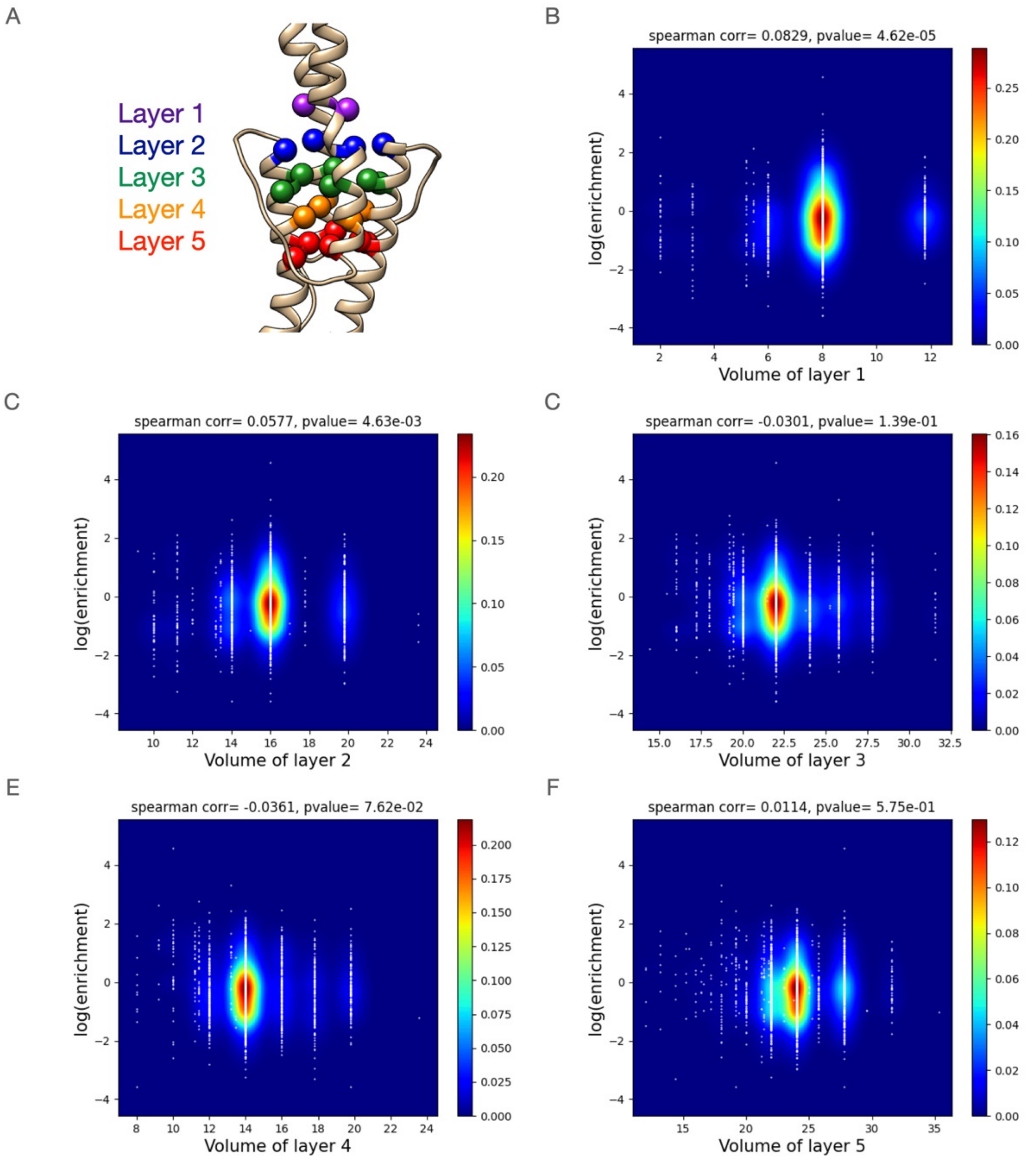


Figure 2.21 - Side-chain volume per hydrophobic packing layer of HAMP 'ILVFSPTA' library (3 or fewer mutations). (A) The 5 hydrophobic packing layers of the HAMP domain are shown, including a single layer for the n-terminal 2-helix bundle (layer 1) and four layers of the 4-helix parallel bundle (layers 2-5). **(B)** sum of side-chain volumes for layer 1 vs log(enrichment) for all unique sequences with 3 or fewer mutations and at least 10 counts in selected + unselected samples. Heat map shows kernel density of scatter plot for clarity. Spearman correlation coefficient and p-value of coefficient are shown above graph. **(C)** same plot as (B) for layer 2. **(D)** same plot as (B) for layer 3. **(E)** same plot as (B) for layer 4. **(F)** same plot as (B) for layer 5.

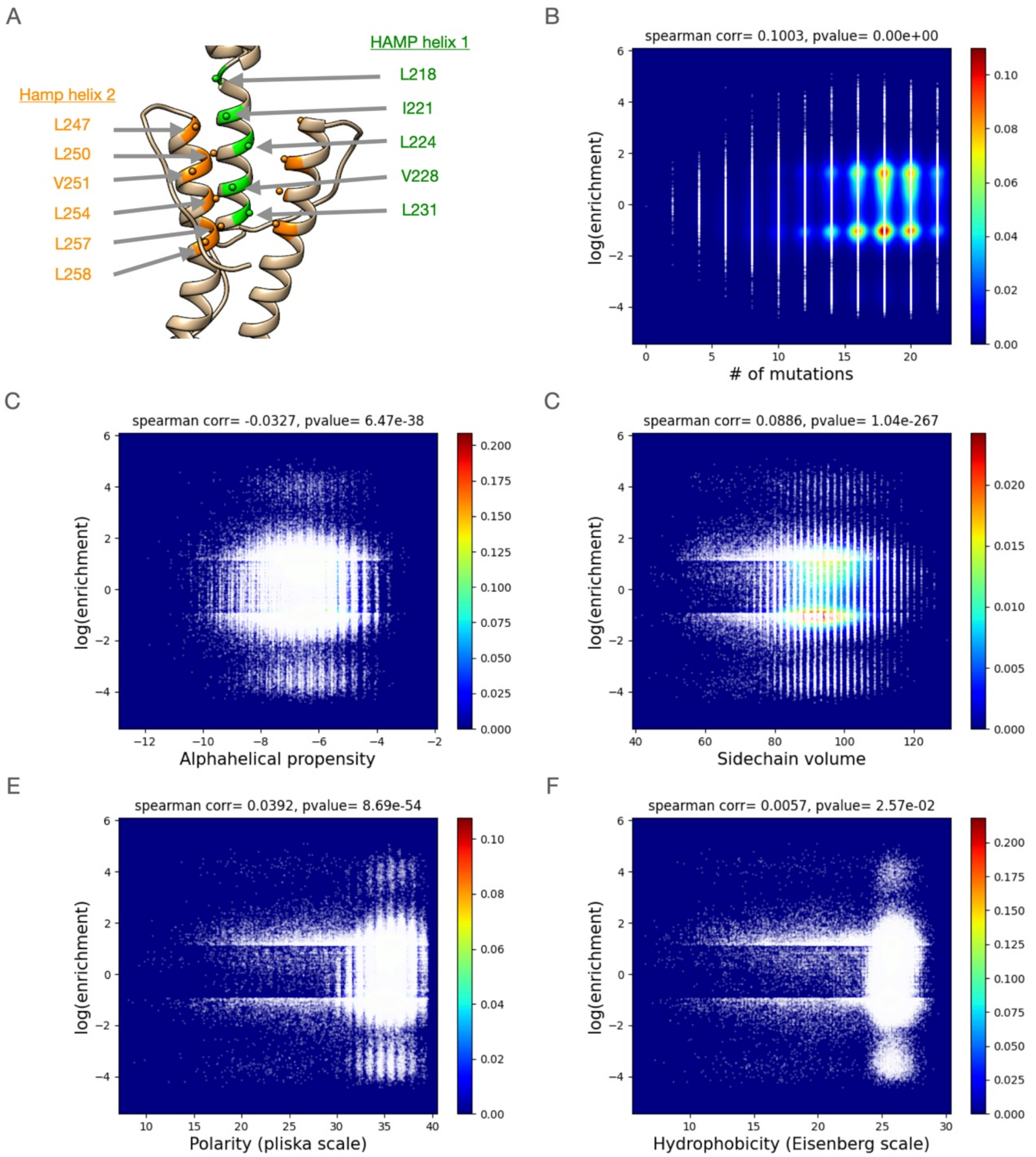


Figure 2.22 - 'ILVFSPTA' library sequence enrichment vs. amino acid physiochemical properties (all mutations). (A) WT sequence and positions of the 'ILVFSPTA' library variants is shown in green (HAMP helix 1) and orange (HAMP helix 2) spheres. The front HAMP helix 1 has been removed for clarity. **(B)** log(enrichment) vs. number of substitutions in sequence for all unique variants with 3 or fewer mutations and a threshold of 10 counts in selected + unselected samples summed across all 11 positions on both monomer chains. Heat map shows kernel density of scatter plot for clarity. Spearman correlation coefficient and p-value of coefficient are shown above graph. **(C)** same plot as (B) for log(enrichment) vs alpha-helical propensity. **(D)** sample plot as (B) for log(enrichment) vs. sidechain volume. **(E)** same plot as (B) for log(enrichment) vs. side-chain polarity (Pliska scale) **(F)** same plot as (B) for log(enrichment) vs side-chain hydrophobicity (Eisenberg scale).

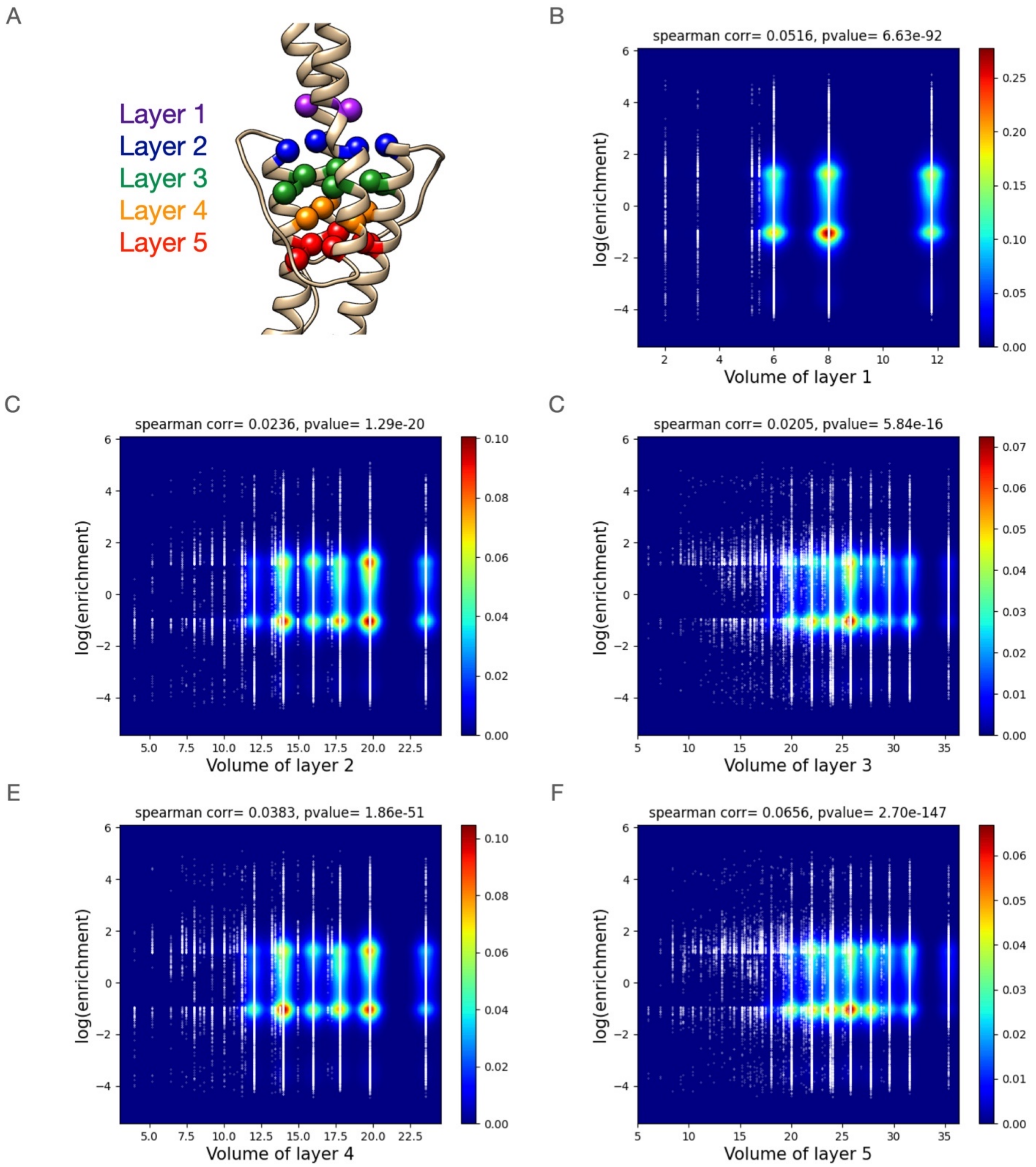


Figure 2.23 - Side-chain volume per hydrophobic packing layer of HAMP 'ILVFSPTA' library (all mutations). (A) The 5 hydrophobic packing layers of the HAMP domain are shown, including a single layer for the n-terminal 2-helix bundle (layer 1) and four layers of the 4-helix parallel bundle (layers 2-5). **(B)** sum of side-chain volumes for layer 1 vs log(enrichment) for all unique sequences with 3 or fewer mutations and at least 10 counts in selected + unselected samples. Heat map shows kernel density of scatter plot for clarity. Spearman correlation coefficient and p-value of coefficient are shown above graph. **(C)** same plot as (B) for layer 2. **(D)** same plot as (B) for layer 3. **(E)** same plot as (B) for layer 4. **(F)** same plot as (B) for layer 5.

The S-Helix library had the largest correlations between enrichment and sequence properties overall. Like in both HAMP libraries, greater number of sequence substitutions generally led to larger effect sizes (**Figure 2.25B**). When we examine sequences with 3 or fewer mutations, we find a strong correlation between alpha-helical propensity and library enrichment, with less helical sequences leading to higher autokinase activities (**Figure 2.25C**). This relationship was hypothesized in chapter 1, with less helical S-Helix sequences representing a lower coupling constant between the HAMP and the autokinase, thereby lessening the inhibition of the autokinase activity by the HAMP. We also find moderate correlations between library enrichment and overall charge (**Figure 2.25D**), with more negative sequences having higher enrichments in general, and hydrophobicity (**Figure 2.25F**), with more hydrophobic sequences having greater enrichment values.

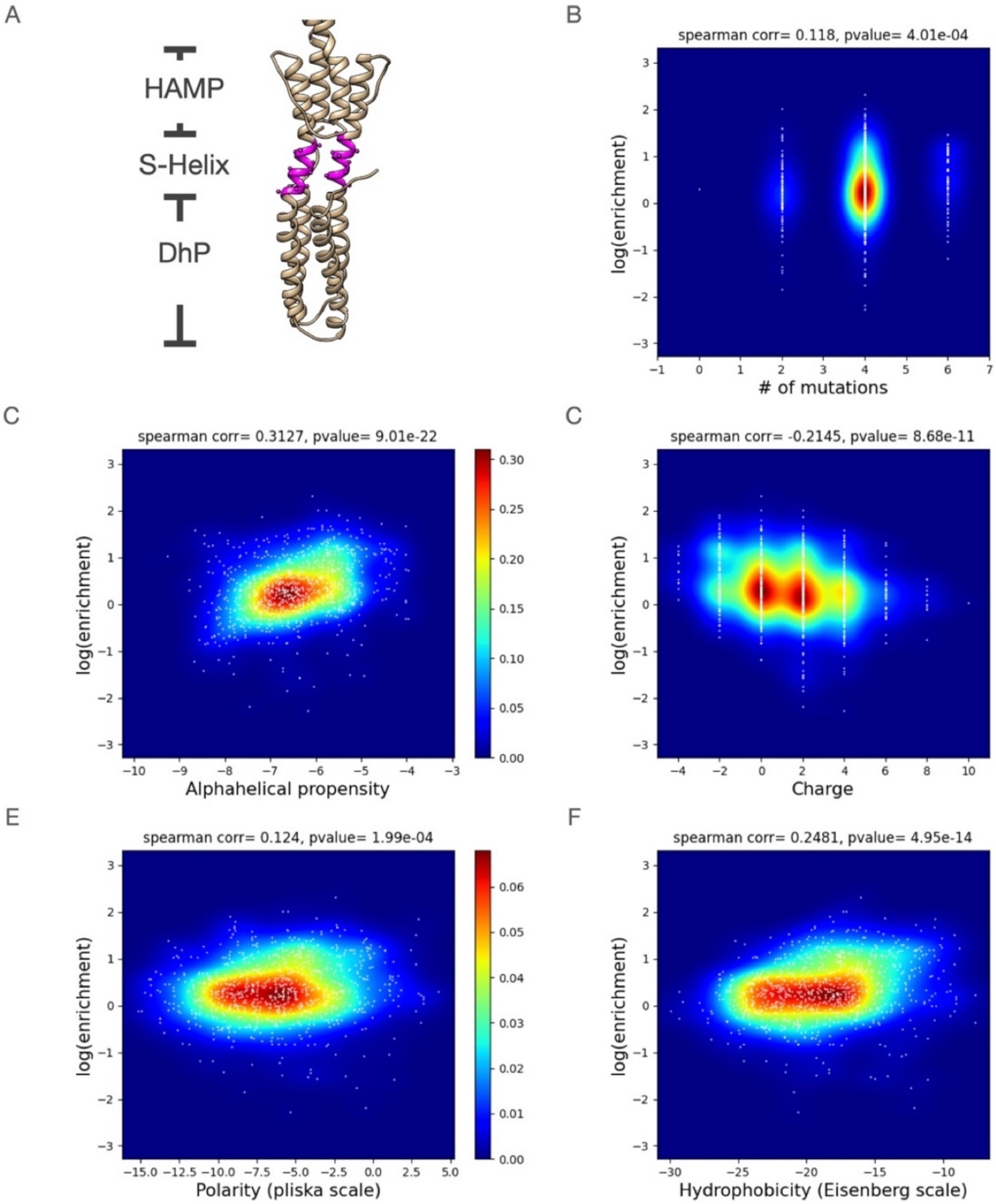


Figure 2.24 – S-Helix library sequence enrichment vs. amino acid physiochemical properties (3 or fewer mutations). (A) The 10-residue long S-helix (magenta) bridges the HAMP and DHp four-helix bundles. **(B)** log(enrichment) vs. number of substitutions in sequence for all unique variants with 3 or fewer mutations and a threshold of 10 counts in selected + unselected samples summed across all 10 positions on both monomer chains. Heat map shows kernel density of scatter plot for clarity. Spearman correlation coefficient and p-value of coefficient are shown above graph. **(C)** same plot as (B) for log(enrichment) vs alpha-helical propensity. **(D)** sample plot as (B) for log(enrichment) vs. total charge. **(E)** same plot as (B) for log(enrichment) vs. side-chain polarity (Pliska scale) shows a weak correlation between increasing polarity and increasing enrichment. **(F)** same plot as (B) for log(enrichment) vs side-chain hydrophobicity (Eisenberg scale).

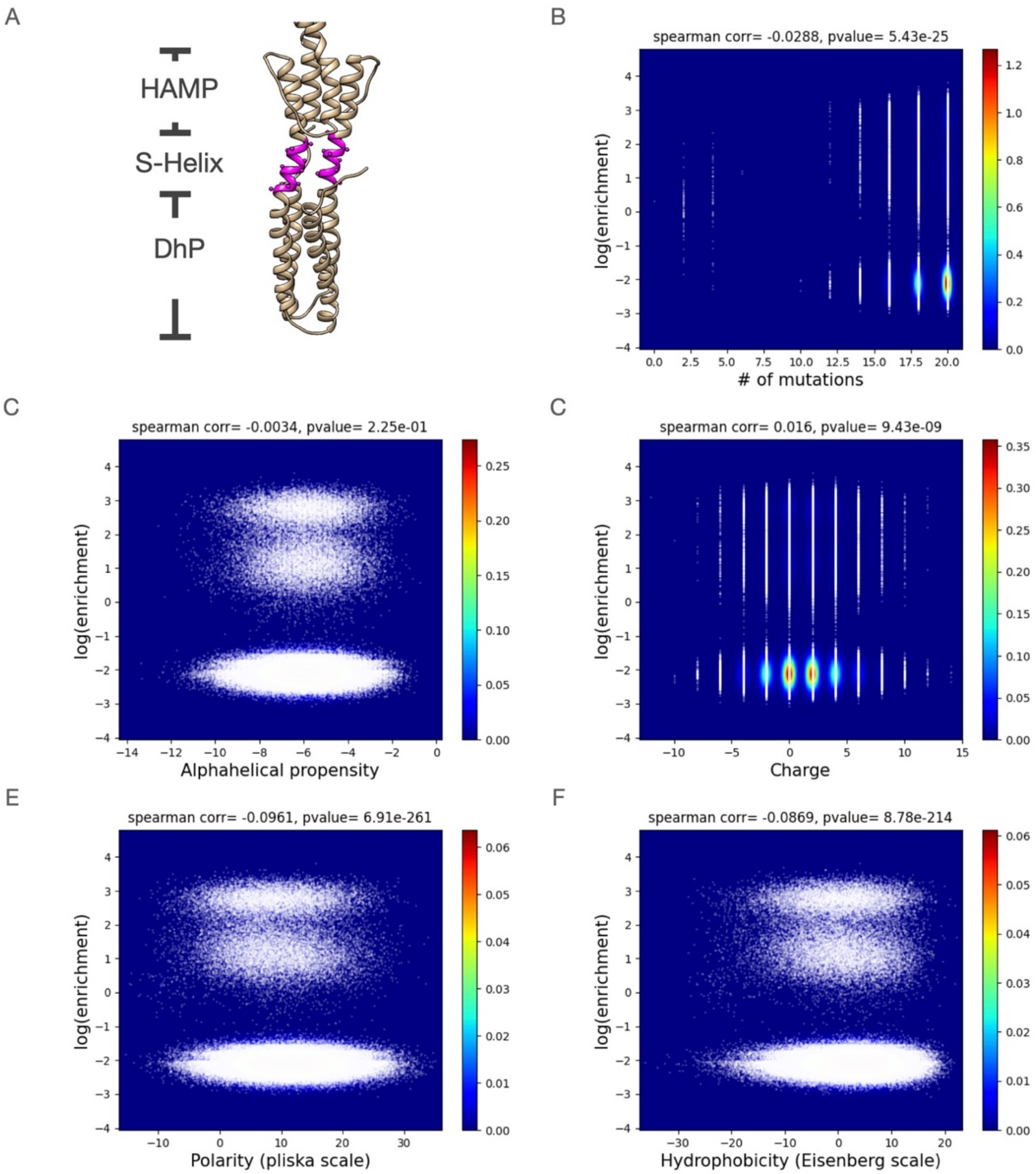


Figure 2.25 – S-Helix library sequence enrichment vs. amino acid physiochemical properties (all mutations, 100 count cut-off). (A) The 10-residue long S-helix (magenta) bridges the HAMP and DhP four-helix bundles. **(B)** log(enrichment) vs. number of substitutions in sequence for all unique variants with a threshold of 100 counts in selected + unselected samples summed across all 11 positions on both monomer chains. Heat map shows kernel density of scatter plot for clarity. Spearman correlation coefficient and p-value of coefficient are shown above graph. **(C)** same plot as (B) for log(enrichment) vs alpha-helical propensity. **(D)** sample plot as (B) for log(enrichment) vs. total charge. **(E)** same plot as (B) for log(enrichment) vs. side-chain polarity (Pliska scale) shows a weak correlation between increasing polarity and increasing enrichment. **(F)** same plot as (B) for log(enrichment) vs side-chain hydrophobicity (Eisenberg scale).

Predicting phoQ activity from HAMP and S-Helix sequences: A deep learning approach

In order to conclusively determine the predictive relationship between PhoQ library sequences and associated function enrichment scores, we turned to an unsupervised machine learning software, DeepRT (1). DeepRT uses convolutional neural networks (CNN) and recursive neural networks (RNN) to eliminate the use of handcrafted rules in creating associations between peptide sequences and a desired functional outcome. Trained on various chromatography elutions of peptide sequences, DeepRT was able to predict retention times of peptides in various chromatography applications with an astoundingly high correlation (>0.99). We sought to use this method to establish the predictability of PhoQ activity (using library enrichment as proxy) from no other information than the primary sequence itself. Trimmed sequences encompassing library regions and their associated normalized enrichment scores (normalized to the min-max range, see **Methods**) were used for training, with 90% data used in training sets, and 10% used for evaluation. **Figure 2.26** shows the correlation between DeepRT-predicted enrichment scores and experimentally determined library enrichment scores. With a spearman correlation of >0.8 for the ILVFSPTA library, and >0.6 for the S-Helix library, we are able to show the presence of significant predictive information in our sequences without the incorporation of any features related to amino-acid or peptide physiochemical properties and structural input regarding PhoQ domains. While these correlation coefficients pale in comparison to those achieved for the predication of sequence retention times by DeepRT, it is important to note that the associative strength of the data used in training and evaluation are markedly different. Properties like peptide retention times can be precisely measured experimentally, and do not have large variabilities between measurements, whereas sequence enrichment by functional selection of a library, particularly based on the muted dynamic range of PhoQ response, are expected to have significant noise. In future work, we aim to evaluate the feature representations of the input sequences, which is simultaneously learned on along with the predictive model, for their ability to reproduce the clustering of amino acids by their physiochemical properties and any possible co-evolution of sequences as it relates to PhoQ activity. We are also exploring supervised machine learning methods, such as AlphaFold2.0 to attempt HAMP structure prediction from functional sequencing data.

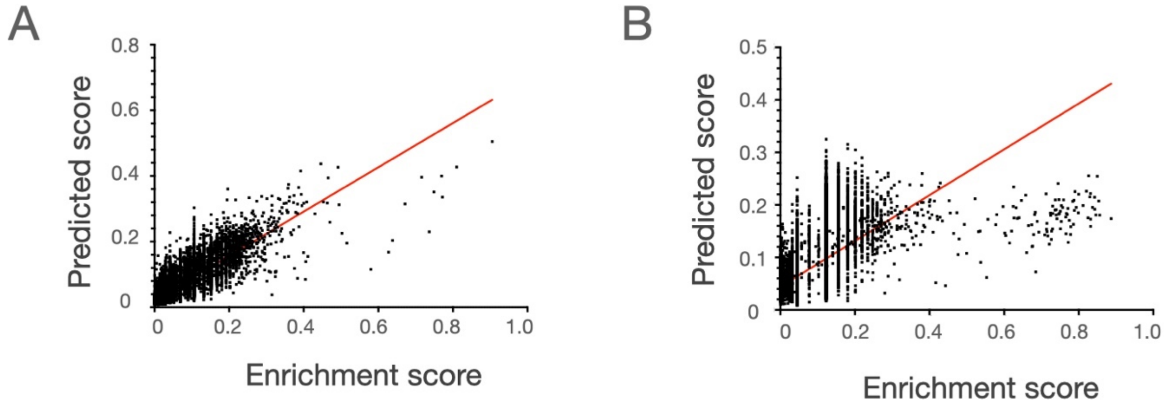


Figure 2.26 - DeepRT method and results of sequence enrichment prediction. (A) correlation of library enrichment scores to DeepRT model predicted scores for ILVFSPTA library. 2,109,998 sequences were used in the training set and model was evaluated against 244,897 test sequences. Correlation of prediction was 0.819|0.755 (Spearman|Pierson). **(B)** correlation of library enrichment scores to DeepRT model predicted scores for S-Helix library. 2,204,069 sequences were used in the training set and model was evaluated against 244,897 test sequences. Correlation of prediction was 0.605|0.607 (Spearman|Pierson).

Discussion

In this work, we evaluate the sequence and structural landscapes of the HAMP domain and the S-Helix motif that links it to the autokinase for its ability to modulate the activity of the latter domain. We created libraries that vary the 11-residue hydrophobic core of the HAMP domain to conservative and permissive hydrophobic substitutions, as well as the 10-residue S-Helix linker region to all possible amino acid variants (and stop codons). By sorting for library variants that retain high PhoQ activity at inhibitory high [Mg^{2+}] conditions, we are able to assess how the structure of the parallel four-helix bundle of the HAMP, and the off-canonical geometry coiled-coil of the S-helix alter the repression of kinase activity by the HAMP. We find that, as hypothesized, variants that are hypothesized to destabilize the thermodynamically favored state of the HAMP bundle, such as the accumulation of small residues in the core, activate the autokinase. While no single gross physiochemical property (hydrophobicity, alpha-helical propensity etc) of the HAMP hydrophobic core can adequately explain the observed enrichments, at least not monotonically, we can still predict phoQ activity from the HAMP core sequence with a high correlation using a deep-learning method developed model. Similarly, variants in the S-Helix that lower structural coupling between the HAMP and the autokinase are found to activate the kinase. We observed stronger correlations between gross physiochemical properties of the S-Helix and library enrichment, with enhanced polarity and decreased alpha-helical propensity being moderate predictors of library enrichment. We also detect the presence of a potential off-register helical interface in the S-Helix, based on the relative enrichments of polar and apolar residues along the S-Helix length. This may correspond to the hypothesized lowly-coupled kinase state of PhoQ, in which the HAMP and S-Helix are bridged by an underwound coiled-coil. DeepRT, once again, establishes stronger sequence-enrichment correlations for the S-Helix library, again demonstrating the multi-factorial nature of these activity determining properties.

Given the utility of such a method, one can explore several more targeted libraries and stronger sorting methods for functional variants to further examine the concept of autokinase activity modulation by the HAMP domain. One big impediment to such an experiment is the relatively low dynamic range of PhoQ activity, which has a 5-10 fold change in activity between stimulatory and inhibitory conditions. Using a HK with a higher dynamic range could allow for the more stringent selection of active variants. Furthermore, the relative depletion of small residues in the ILVFSPTA libraries at the library building step limited our ability to evaluate any generalizable rules regarding hydrophobicity and side-chain volume in this library. A library with a more balanced composition, or in a screening background that is more tolerant of these hypothetically high-activity variants would yield more interpretable results. Finally, it would be useful to complement this library screening strategy by sorting for populations that remain inactive at low [Mg^{2+}] concentrations, to further differentiate between sequences that have WT-like responsiveness to ligand, as opposed to sequences that are incapable of producing high kinase output altogether.

With respect to analyses of such libraries, unsupervised machine learning is uniquely useful in elucidating sequence rules as they relate to function, without the bias of any structural models of HKs, signaling mechanisms, or even the knowledge of fundamental amino-acid properties. Having established the predictive value of sequences in an unbiased manner via

DeepRT, we can examine more nuanced and HK-specific questions, given various available structures, homology models and functional data in future work.

Materials and Methods

Materials

BW25113 and HK knockout strains were obtained from the Keio collection.

TIM100 (ΔPhoQ) was obtained from Tim Mayashiro (Goulian lab)

DNA megaprimers were purchased from IDT

Genemorph II EZ clone mutagenesis kit

Magnesium sulfate

MOPS minimal media

PCR instrument

Electroporation cuvettes

Electroporator

Methods

Library construction: C-terminally his-tagged wild type E. coli PhoQ was cloned into the MCS of pTrc99a by restriction cloning. This plasmid was used as the template for library construction. To clone libraries, single-stranded DNA primers spanning the HAMP domain and S-helix domain and containing library variants of interest were purchased from IDT (sequences below). A double stranded megaprimer was synthesized using 1 round of PCR and a complementary primer (**primer 1, 2**). 250 ng of double-stranded megaprimer was mixed with 50 ng of pTrc99a-PhoQ plasmid in 1X enzyme mix and 25 rounds of PCR conducted, as per manufacturer instructions.

The resulting library was checked for proper diversification by Sanger sequencing and concentrated down to 10 µL of DNA in MilliQ water by a mini-PCR clean-up kit. 5 µL of this DNA was transformed into SURE electrocompetent cells (Agilent), recovered in SOC media and grown overnight. Library plasmid was then extracted using a miniprep kit.

HAMP_ILVF lib:

GTGGGTCGCCGCCTGGTGGAGTNTTCGCCCCNTTGAAGCCNTTGCAAAAGAANTTCGCGAANTTGAAGAACATAACCGCGAATTGCTCAATCCAGCCACAACGCGAGAANTTACCAGTNTTNTTCGAAACNTTAACCGANTTNTTAAAAGTGAACGCGAACGTTACGACAAATAC

HAMP_ILVFSPTA lib:

GTGGGTCGCCGCCTGGTGGAGTNYTCGCCCCNYTGAAGCCNYTGCAAAAGAANYTCGCGAANYTGAAGAACATAACCGCGAATTGCTCAATCCAGCCACAACGCGAGAANYTACCAGTNYTNYTCGAAACNYTAACCGANYTNYTAAAAGTGAACGCGAACGTTACGACAAATAC

S-helix library:

CAGTCTGGTACGAAACCTGAACCGATTGTTAAAAAGTNNKNNKNNKNNKNNKNNKNNKNNKNNKNNKACGCTCACCGACCTGACCCATAGTCTGAAA

Primer 1 (HAMP_lib_**R):** GTATTTGTCGTAACGTTCGCGTTCACTTTT

Primer 2 (SHelix_lib_R): TTTCAGACTATGGGTCAGGTCGGTGAGCGT

Electroporation of library into TIM100: To make electrocompetent TIM100 cells, a culture of TIM100 was grown in LB media to OD600 ~ 0.4, centrifuged at 4C, and washed 3X with ice-cold 10% glycerol in H2O. The final resuspension was tested for competency using pTrc99a-PhoQ WT plasmid. To introduce constructed libraries into TIM100, 2 µL of library plasmid was combined with 100 µL of electrocompetent TIM100 cells in a 2 mm gap electroporation cuvette, incubated on ice and electroporated (1.8 KV, 200 Ω, 25 µFd). Electroporated cells were immediately recovered in 1 mL of SOC broth, transferred into culture tubes and incubated at 37°C with shaking for 90 mins. Cultures were then frozen and stored at -80C in 20% glycerol until sorting.

Sorting for active PhoQ variants: Frozen stocks of TIM100 containing library were thawed on Ice, and diluted into MOPS minimal media + 20 mM $MgSO_4$ and grown with shaking at 37°C for 4 hours. The culture was then sorted by FACS (MoFlo Astrios) using a 100µm nozzle**;** $5x10^6$ cells were collected gated by forward and side-scatter alone for the "unselected" library, and $1x10^5$ cells in the top 1% GFP signal were collected for the "selected" library. Recovered cells were diluted into 1X LB media and grown for an additional 4 hours at 37°C with shaking, flash frozen and stored at -80°C until library sequencing.

Illumina sequencing library construction and sequencing: Sorted library cultures were thawed on ice, and centrifuged at 90,000xg for 10 minutes to collect cells. Cell pellet was then resuspended in 1X KAPA PCR mix containing **primers 3-14**, and heated at 95°C for 10 minutes to lyse cells and release library. After addition of polymerase, 20 rounds of PCR were conducted to amplify library containing regions. In order to overcome the potential issue of base-calling redundancy between the 6 sequencing libraries, a staggered priming approach was used for each library. Resulting product was purified by gel-extraction and extended with illumina sequencing and barcoding sequences. Libraries were barcoded using custom built 7 bp barcode Illumina sequencing primers (DeRisi lab, **primer 15, 16**) using Kapa Hi-fi polymerase (Kapa Biosystems) for 2 amplification cycles, and barcoded library amplified in the same reaction using hot-start primers (**primer 17, 18**) for an additional 12 cycles. Hot-start primers were activated by heating reaction at 94°C for 10 min. Resulting barcoded libraries were checked for quality by gel electrophoresis, concentration quantified by nano-drop and multiplexed (1 sample with 6 libraries). This sample as then paired-end sequenced (100 bp reads per end) using a HiSeq Illumina sequencer with a 7 bp barcode sequencing using standard illumina sequencing primers per manufacturer's instruction at the Center for Advanced Technology at UCSF.

ILVF "selected":

Primer 3 (For): 5' - CTACACGACGCTCTTCCGATCT CGCCTGGTGGAGT - 3'

Primer 4 (Rev): 5' - CAGACGTGTGCTCTTCCGATCT GGGTCAGGTCGGTGAGCG - 3'

ILVF "unselected":

Primer 5 (For): 5' - CTACACGACGCTCTTCCGATCT CCGCCTGGTGGAGT - 3'

Primer 6 (Rev): 5' - CAGACGTGTGCTCTTCCGATCT GGTCAGGTCGGTGAGCG - 3'

ILVFSPTA "selected":

Primer 7 (For): 5' - CTACACGACGCTCTTCCGATCT GCCGCCTGGTGGAGT - 3'

Primer 8 (Rev): 5' - CAGACGTGTGCTCTTCCGATCT GTCAGGTCGGTGAGCG - 3'

ILVFSPTA "unselected":

Primer 9 (For): 5' - CTACACGACGCTCTTCCGATCT CGCCGCCTGGTGGAGT - 3'

Primer 10 (Rev): 5' - CAGACGTGTGCTCTTCCGATCT TCAGGTCGGTGAGCG - 3

SHelix "selected":

Primer 11 (For): 5' - CTACACGACGCTCTTCCGATCT TCGCCGCCTGGTGGAGT - 3'

Primer 12 (Rev): 5' - CAGACGTGTGCTCTTCCGATCT CAGGTCGGTGAGCG - 3'

SHelix "unselected":

Primer 13 (For): 5' - CTACACGACGCTCTTCCGATCT GTCGCCGCCTGGTGGAGT - 3'

Primer 14 (Rev): 5’ - CAGACGTGTGCTCTTCCGATCT AGGTCGGTGAGCG - 3’

PCR protocol

1. 95°C, 3 min

2. 98°C, 20 sec

3. 65°C, 15 sec

4. 72°C, 30 sec

6. Repeat 2-4 for 20 cycles

7. 72°C, 2 min

8. 4°C, forever

Illumina library primers:

Primer 15 (For): 5’ - AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGAC

GCTCTTCCGATCT - 3’

Primer 16 (Rev): 5’ -

CAAGCAGAAGACGGCATACGAGAT***XXXXXXX***GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT - 3’

Hot-start primers:

Primer 17 (SoLM1_18): 5’ - AATGATACGGCGACCACC - 3’

Primer 18 (SoLM2_19): 5’ - CAAGCAGAAGACGGCATAC - 3’

PCR protocol:

1. 94°C, 1 min

2. 94°C, 30 sec

3. 40°C, 30 sec

4. 72°C, 1 min

5. Repeat steps 2-4 1X

6. 94°C, 10 min (hot start primer activation)

7. 94°C, 30 sec

8. 55°C, 30 sec

9. 72°C, 1 min

10. Repeat steps 7-9 11X

11. 72°C, 5 min

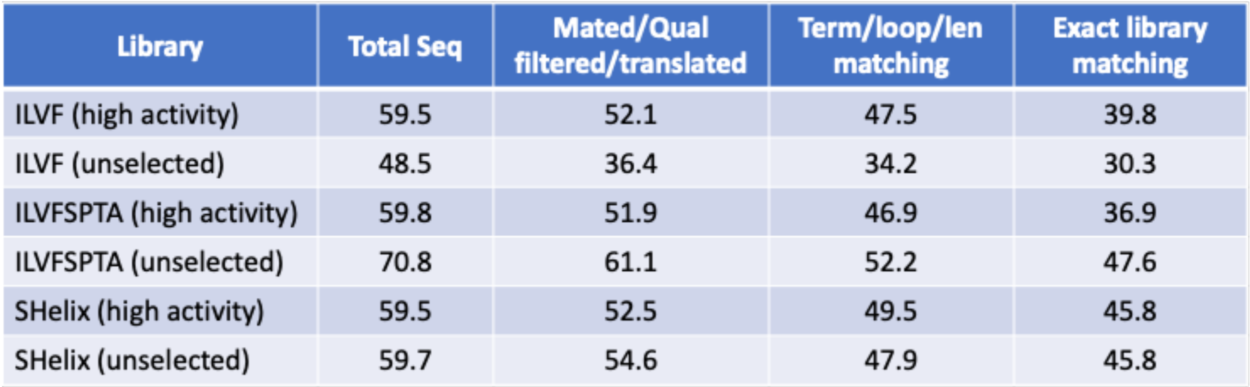
Data analysis: Paired-end mating, quality filtering and translation of sequences was carried out using galaxy project tools (14). Briefly, paired-end sequences were mated using a 9 basepair perfect match criteria using fastq-join, then filtered for sequencing quality (<1% error for >90% of sequence), converted to FASTA format and translated to amino-acid sequences. Sequences were then filtered by length and exact matching to constant sequences of amplified region of PhoQ. Sequences that had a non-coded variant in library positions were also discarded. Retention of sequences after each step is summarized in Table **3**. Sequences were then counted, and enrichment of variants between "selected" and "unselected" libraries was calculated as shown.

Enrichment ratio = (count in "selected" +1)/ (count in "unselected").

For linkage-independent sequence analysis, each amino acid variant at each position was counted for both selected and unselected libraries without preserving sequence connectivity, and an enrichment ratio was calculated. The averaged $\log_2$(enrichment) was then used to hierarchically cluster by amino acid variant and sequence variant position. Heatmaps were generated using R script d3heatmap tool. For S-Helix library, average enrichments were calculated for apolar (M, I, L, V, F, Y, W, C, A), polar (D, E, K, R, H, N, Q, S, T), large apolar (M, I, L, V, F), small apolar (C, A), aromatic (Y, W), positive (K, R, H), negative (D, E), uncharged polar residues (N, Q, S, T), preferred at *a* position in soluble dimers (A, C, E, L, I, T, V) and preferred at *d* position in soluble dimers (G, L, I, M, V, Y) (15) and plotted against residue number.

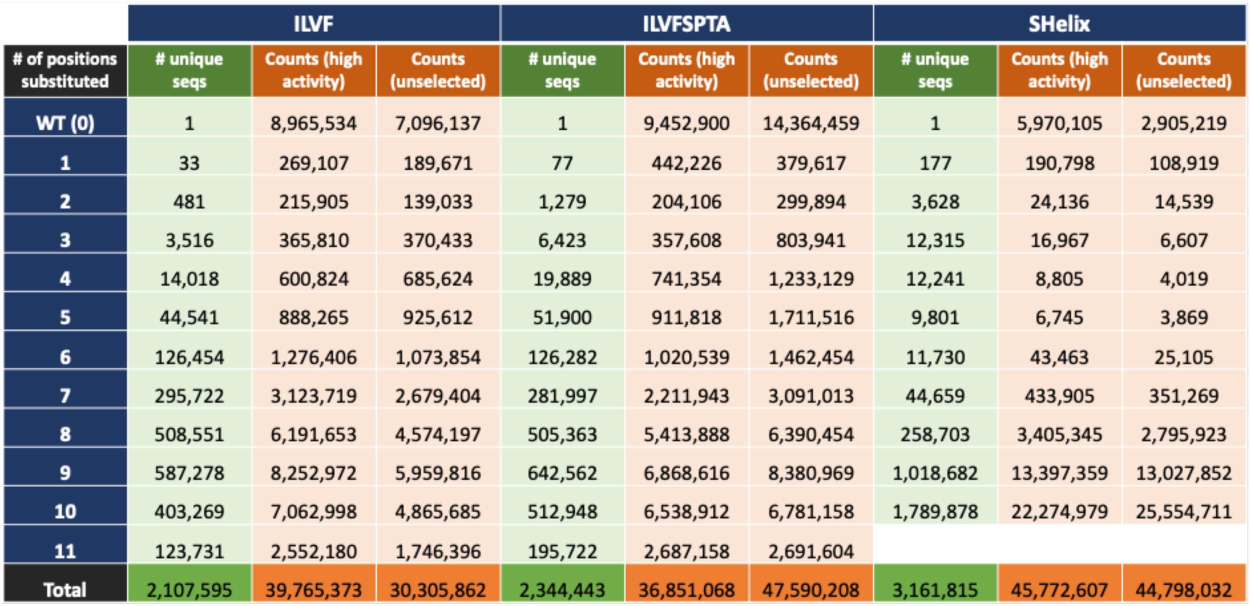
To calculate gross sequence physiochemical properties, hydrophobicity (Eisenberg scale (16)), polarity (Pliska scale (17)), side-chain volume (18), alpha-helical propensity (19) and charge values (listed in **Table 2.7**) were summed across all HAMP-core or S-Helix positions for sequence with total count ≥10. Kernel density plots representing data density were generated for the resulting enrichment vs. property scatter plots using python Scipy.stats kde function, and Spearman and Pierson correlation coefficients calculated using Scipy.stats spearmanr and Numpy corrcoef functions respectively.

Table 2.1 – sequence recovery after various quality filtering steps. Number of sequences in millions that were discovered experimentally and retained through paired-end mating, sequencing quality filters, length and constant regions matching.



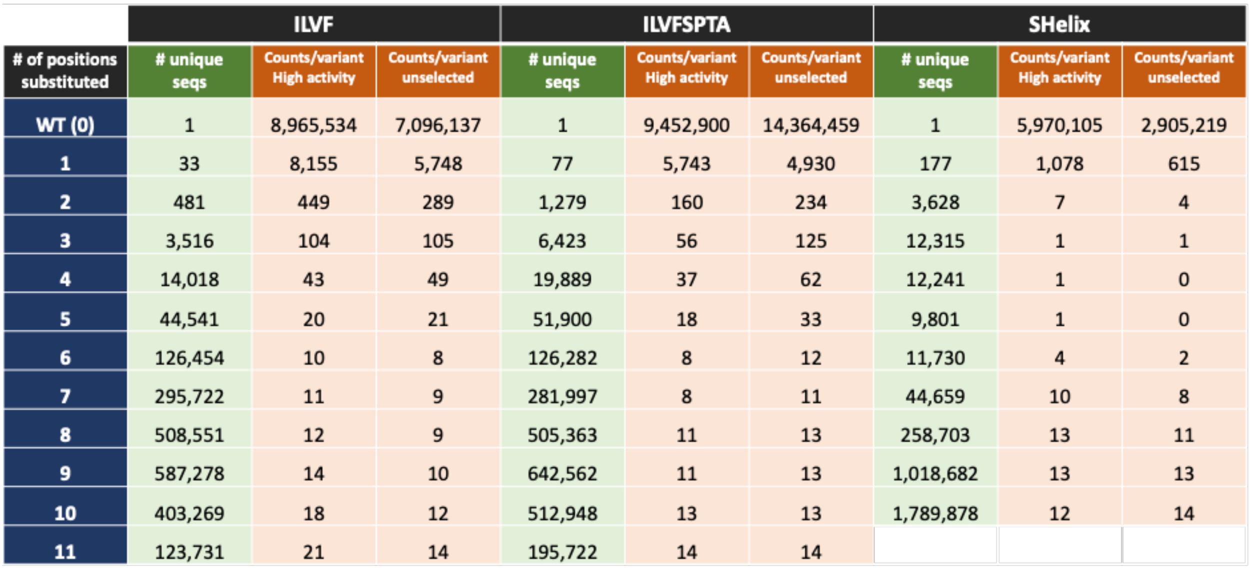
Library	Total Seq	Mated/Qual filtered/translated	Term/loop/len matching	Exact library matching
ILVF (high activity)	59.5	52.1	47.5	39.8
ILVF (unselected)	48.5	36.4	34.2	30.3
ILVFSPTA (high activity)	59.8	51.9	46.9	36.9
ILVFSPTA (unselected)	70.8	61.1	52.2	47.6
SHelix (high activity)	59.5	52.5	49.5	45.8
SHelix (unselected)	59.7	54.6	47.9	45.8

Table 2.2 – Total counts for each number of mutations in HAMP and S-Helix library.



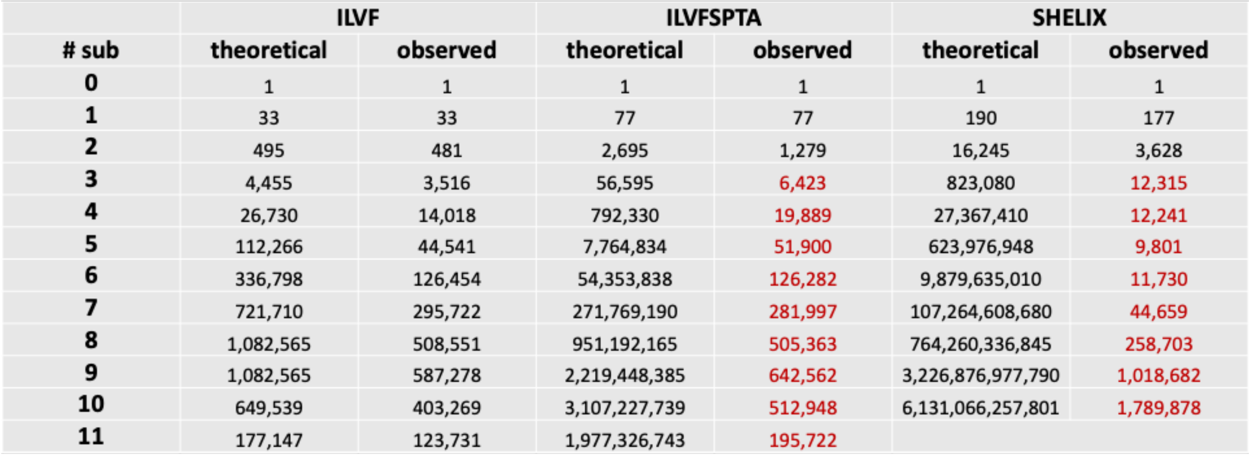
	ILVF			ILVFSPTA			SHelix		
# of positions substituted	# unique seqs	Counts (high activity)	Counts (unselected)	# unique seqs	Counts (high activity)	Counts (unselected)	# unique seqs	Counts (high activity)	Counts (unselected)
WT (0)	1	8,965,534	7,096,137	1	9,452,900	14,364,459	1	5,970,105	2,905,219
1	33	269,107	189,671	77	442,226	379,617	177	190,798	108,919
2	481	215,905	139,033	1,279	204,106	299,894	3,628	24,136	14,539
3	3,516	365,810	370,433	6,423	357,608	803,941	12,315	16,967	6,607
4	14,018	600,824	685,624	19,889	741,354	1,233,129	12,241	8,805	4,019
5	44,541	888,265	925,612	51,900	911,818	1,711,516	9,801	6,745	3,869
6	126,454	1,276,406	1,073,854	126,282	1,020,539	1,462,454	11,730	43,463	25,105
7	295,722	3,123,719	2,679,404	281,997	2,211,943	3,091,013	44,659	433,905	351,269
8	508,551	6,191,653	4,574,197	505,363	5,413,888	6,390,454	258,703	3,405,345	2,795,923
9	587,278	8,252,972	5,959,816	642,562	6,868,616	8,380,969	1,018,682	13,397,359	13,027,852
10	403,269	7,062,998	4,865,685	512,948	6,538,912	6,781,158	1,789,878	22,274,979	25,554,711
11	123,731	2,552,180	1,746,396	195,722	2,687,158	2,691,604			
Total	2,107,595	39,765,373	30,305,862	2,344,443	36,851,068	47,590,208	3,161,815	45,772,607	44,798,032

Table 2.3 – Average number of counts per unique variant for each number of mutations. Green columns show the number of observed unique sequence variants. Orange columns show the average count per unique variant for each mutation number.



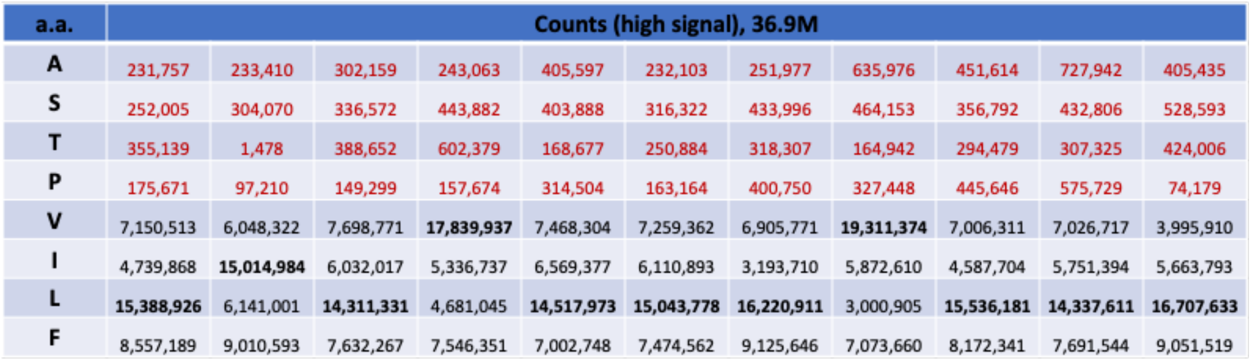
	ILVF			ILVFSPTA			SHelix		
# of positions substituted	# unique seqs	Counts/variant High activity	Counts/variant unselected	# unique seqs	Counts/variant High activity	Counts/variant unselected	# unique seqs	Counts/variant High activity	Counts/variant unselected
WT (0)	1	8,965,534	7,096,137	1	9,452,900	14,364,459	1	5,970,105	2,905,219
1	33	8,155	5,748	77	5,743	4,930	177	1,078	615
2	481	449	289	1,279	160	234	3,628	7	4
3	3,516	104	105	6,423	56	125	12,315	1	1
4	14,018	43	49	19,889	37	62	12,241	1	0
5	44,541	20	21	51,900	18	33	9,801	1	0
6	126,454	10	8	126,282	8	12	11,730	4	2
7	295,722	11	9	281,997	8	11	44,659	10	8
8	508,551	12	9	505,363	11	13	258,703	13	11
9	587,278	14	10	642,562	11	13	1,018,682	13	13
10	403,269	18	12	512,948	13	13	1,789,878	12	14
11	123,731	21	14	195,722	14	14			

Table 2.4 – Completeness of HAMP and S-Helix libraries. The theoretical and observed unique variants counts for each number of mutations is shown. Low representations are highlighted in red text.

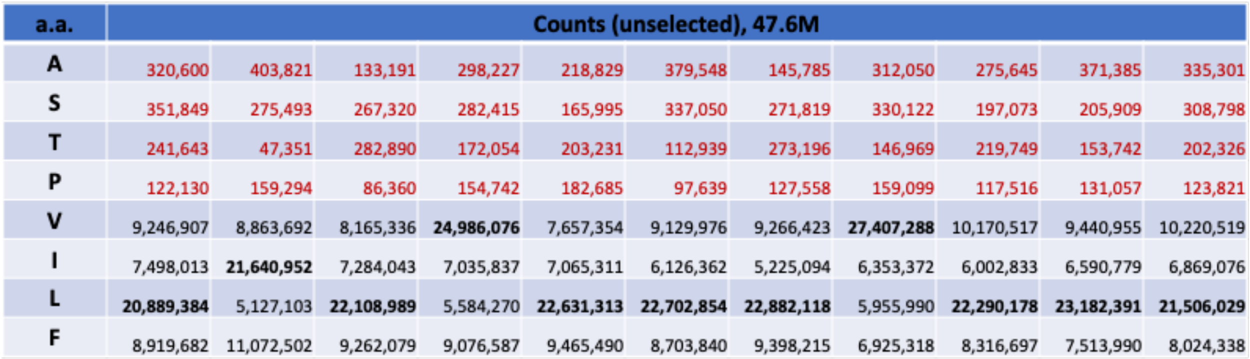


	ILVF		ILVFSPTA		SHELIX	
# sub	**theoretical**	**observed**	**theoretical**	**observed**	**theoretical**	**observed**
0	1	1	1	1	1	1
1	33	33	77	77	190	177
2	495	481	2,695	1,279	16,245	3,628
3	4,455	3,516	56,595	6,423	823,080	12,315
4	26,730	14,018	792,330	19,889	27,367,410	12,241
5	112,266	44,541	7,764,834	51,900	623,976,948	9,801
6	336,798	126,454	54,353,838	126,282	9,879,635,010	11,730
7	721,710	295,722	271,769,190	281,997	107,264,608,680	44,659
8	1,082,565	508,551	951,192,165	505,363	764,260,336,845	258,703
9	1,082,565	587,278	2,219,448,385	642,562	3,226,876,977,790	1,018,682
10	649,539	403,269	3,107,227,739	512,948	6,131,066,257,801	1,789,878
11	177,147	123,731	1,977,326,743	195,722		

Table 2.5 – Number of variants observed at each position of the HAMP ‘ILVFSPTA’ library. Counts at each position are shown for the high signal (selected) and unselected populations. There was a notable depletion of small residue variants in both samples.



a.a.	Counts (high signal), 36.9M										
A	231,757	233,410	302,159	243,063	405,597	232,103	251,977	635,976	451,614	727,942	405,435
S	252,005	304,070	336,572	443,882	403,888	316,322	433,996	464,153	356,792	432,806	528,593
T	355,139	1,478	388,652	602,379	168,677	250,884	318,307	164,942	294,479	307,325	424,006
P	175,671	97,210	149,299	157,674	314,504	163,164	400,750	327,448	445,646	575,729	74,179
V	7,150,513	6,048,322	7,698,771	**17,839,937**	7,468,304	7,259,362	6,905,771	**19,311,374**	7,006,311	7,026,717	3,995,910
I	4,739,868	**15,014,984**	6,032,017	5,336,737	6,569,377	6,110,893	3,193,710	5,872,610	4,587,704	5,751,394	5,663,793
L	**15,388,926**	6,141,001	**14,311,331**	4,681,045	**14,517,973**	**15,043,778**	**16,220,911**	3,000,905	**15,536,181**	**14,337,611**	**16,707,633**
F	8,557,189	9,010,593	7,632,267	7,546,351	7,002,748	7,474,562	9,125,646	7,073,660	8,172,341	7,691,544	9,051,519



a.a.	Counts (unselected), 47.6M										
A	320,600	403,821	133,191	298,227	218,829	379,548	145,785	312,050	275,645	371,385	335,301
S	351,849	275,493	267,320	282,415	165,995	337,050	271,819	330,122	197,073	205,909	308,798
T	241,643	47,351	282,890	172,054	203,231	112,939	273,196	146,969	219,749	153,742	202,326
P	122,130	159,294	86,360	154,742	182,685	97,639	127,558	159,099	117,516	131,057	123,821
V	9,246,907	8,863,692	8,165,336	**24,986,076**	7,657,354	9,129,976	9,266,423	**27,407,288**	10,170,517	9,440,955	10,220,519
I	7,498,013	**21,640,952**	7,284,043	7,035,837	7,065,311	6,126,362	5,225,094	6,353,372	6,002,833	6,590,779	6,869,076
L	**20,889,384**	5,127,103	**22,108,989**	5,584,270	**22,631,313**	**22,702,854**	**22,882,118**	5,955,990	**22,290,178**	**23,182,391**	**21,506,029**
F	8,919,682	11,072,502	9,262,079	9,076,587	9,465,490	8,703,840	9,398,215	6,925,318	8,316,697	7,513,990	8,024,338

Table 2.6 - Summary of Spearman and Pearson correlations for enrichment vs. properties. Spearman correlation, associated p-value and Pearson correlations of log(enrichment) vs various helical properties were calculated for all variants with at least 10 counts in selected + unselected samples for either ≤3 mutation or any number of mutations.

	ILVF			ILVF		
	≤3 mut, ≥10 counts			≥10 counts		
property	**Spearman**	**p-value**	**Pearson**	**Spearman**	**p-value**	**Pearson**
# of mutations	-0.0992	3.35E-05	-0.0925	0.0272	5.15E-33	0.0566
Polarity	-0.1116	2.99E-06	-0.0843	0.118	0.00E+00	0.0972
Hydrophobicity	-0.0583	1.49E-02	-0.0664	0.0609	1.04E-158	0.0463
Alpha-helical propensity	0.0559	1.96E-02	0.0244	-0.0782	2.17E-260	-0.058
Sidechain volume	-0.0761	1.47E-03	-0.0453	0.1033	0.00E+00	0.1042
Volume of layer 1	-0.0226	3.46E-01	0.013	0.0158	3.77E-12	0.0275
Volume of layer 2	-0.0311	1.95E-01	-0.0578	-0.0327	6.91E-47	-0.0273
Volume of layer 3	-0.0595	1.30E-02	-0.0182	0.0604	5.48E-156	0.0633
Volume of layer 4	-0.09	1.68E-04	-0.0752	0.0524	1.16E-117	0.0454
Volume of layer 5	0.0088	7.13E-01	0.0389	0.1037	0.00E+00	0.1099
	ILVFSPTA			**ILVFSPTA**		
	≤3 mut, ≥10 counts			**≥10 counts**		
# of mutations	-0.1442	1.15E-12	-0.117	0.1003	0.00E+00	0.099
Polarity	-0.0324	1.12E-01	-0.0247	0.0392	8.69E-54	-0.0189
Hydrophobicity	-0.1523	5.69E-14	-0.116	0.0057	2.57E-02	-0.0637
Alpha-helical propensity	-0.1534	3.76E-14	-0.1538	-0.0327	6.47E-38	-0.0236
Sidechain volume	0.0092	6.50E-01	-0.0017	0.0886	1.04E-267	0.0709
Volume of layer 1	0.0829	4.62E-05	0.0646	0.0516	6.63E-92	0.059
Volume of layer 2	0.0577	4.63E-03	0.0453	0.0236	1.29E-20	0.0223
Volume of layer 3	-0.0301	1.39E-01	-0.0368	0.0205	5.84E-16	0.0172
Volume of layer 4	-0.0361	7.62E-02	-0.0484	0.0383	1.86E-51	0.0324
Volume of layer 5	0.0114	5.75E-01	-0.0073	0.0656	2.70E-147	0.0575
	S-Helix			**S-Helix**		
	≤3 mut, ≥10 counts			**≥10 counts**		
# of mutations	0.118	4.01E-04	0.1086	-0.0312	2.77E-58	-0.0794
Polarity	0.124	1.99E-04	0.1025	-0.089	0.00E+00	-0.124
Hydrophobicity	0.2481	4.95E-14	0.2207	-0.0796	0.00E+00	-0.1144
Alpha-helical propensity	0.3127	9.01E-22	0.2963	-0.0071	2.72E-04	0.0006
Charge	-0.2145	8.68E-11	-0.2174	0.0185	1.55E-21	0.0182

Table 2.7 - Physiochemical properties of amino acids

Amino acid	Polarity (Pliska)	Alpha-Helical propensity (Kcal mol^{-1})	Sidechain Volume (norm to Ala)	Charge	Hydrophobicity (Eisenberg)
A	0.31	-0.71	1.00	0	0.62
C	1.54	-0.22	2.43	0	0.29
D	-0.77	-0.1	2.78	-1	-0.90
E	-0.64	-0.21	3.78	-1	-0.74
F	1.79	-0.37	5.89	0	1.19
G	0.00	0.00	0.00	0	0.48
H	0.13	0.03	4.66	0	-0.40
I	1.8	-0.17	4.00	0	1.38
K	-0.99	-0.58	4.77	+1	-1.50
L	1.70	-0.52	4.00	0	1.06
M	1.23	-0.42	4.43	0	0.64
N	-0.60	-0.01	2.95	0	-0.78
P	0.72	0.00	2.72	0	0.12
Q	-0.22	-0.33	3.95	0	-0.85
R	-1.01	-0.70	6.13	+1	-2.53
S	-0.04	-0.27	1.60	0	-0.18
T	0.26	-0.09	2.60	0	-0.05
V	1.22	-0.16	3.00	0	1.08
W	2.25	-0.45	8.08	0	0.81
Y	0.96	-0.06	6.47	0	0.26

References

1. C. Ma, *et al.*, Improved Peptide Retention Time Prediction in Liquid Chromatography through Deep Learning. *Anal. Chem.* **90**, 10881–10888 (2018).

2. N. W. Schmidt, G. Grigoryan, W. F. DeGrado, The accommodation index measures the perturbation associated with insertions and deletions in coiled-coils: Application to understand signaling in histidine kinases. *Protein Sci. Publ. Protein Soc.* **26**, 414–435 (2017).

3. A. N. Lupas, J. Bassler, S. Dunin-Horkawicz, The Structure and Topology of α-Helical Coiled Coils. *Subcell. Biochem.* **82**, 95–129 (2017).

4. M. Inouye, Signaling by transmembrane proteins shifts gears. *Cell* **126**, 829–831 (2006).

5. S. A. Chervitz, J. J. Falke, Molecular mechanism of transmembrane signaling by the aspartate receptor: a model. *Proc. Natl. Acad. Sci. U. S. A.* **93**, 2545–2550 (1996).

6. J. J. Falke, A. H. Erbse, The piston rises again. *Struct. Lond. Engl. 1993* **17**, 1149–1151 (2009).

7. M. V. Milburn, *et al.*, Three-dimensional structures of the ligand-binding domain of the bacterial aspartate receptor with and without a ligand. *Science* **254**, 1342–1347 (1991).

8. E. C. Lowe, A. Baslé, M. Czjzek, S. J. Firbank, D. N. Bolam, A scissor blade-like closing mechanism implicated in transmembrane signaling in a Bacteroides hybrid two-component system. *Proc. Natl. Acad. Sci. U. S. A.* **109**, 7298–7303 (2012).

9. K. S. Molnar, *et al.*, Cys-scanning disulfide crosslinking and bayesian modeling probe the transmembrane signaling mechanism of the histidine kinase, PhoQ. *Struct. Lond. Engl. 1993* **22**, 1239–1251 (2014).

10. V. Stewart, The HAMP signal-conversion domain: static two-state or dynamic three-state? *Mol. Microbiol.* **91**, 853–857 (2014).

11. J. S. Parkinson, Signaling mechanisms of HAMP domains in chemoreceptors and sensor kinases. *Annu. Rev. Microbiol.* **64**, 101–122 (2010).

12. I. Tews, *et al.*, The structure of a pH-sensing mycobacterial adenylyl cyclase holoenzyme. *Science* **308**, 1020–1023 (2005).

13. V. Anantharaman, S. Balaji, L. Aravind, The signaling helix: a common functional theme in diverse signaling proteins. *Biol. Direct* **1**, 25 (2006).

14. E. Afgan, *et al.*, The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2018 update. *Nucleic Acids Res.* **46**, W537–W544 (2018).

15. S.-Q. Zhang, *et al.*, The membrane- and soluble-protein helix-helix interactome: similar geometry via different interactions. *Struct. Lond. Engl. 1993* **23**, 527–541 (2015).

16. J. Koehler, N. Woetzel, R. Staritzbichler, C. R. Sanders, J. Meiler, A unified hydrophobicity scale for multispan membrane proteins. *Proteins* **76**, 13–29 (2009).

17. A. Kim, F. C. Szoka, Amino acid side-chain contributions to free energy of transfer of tripeptides from water to octanol. *Pharm. Res.* **9**, 504–514 (1992).

18. J. L. Fauchère, M. Charton, L. B. Kier, A. Verloop, V. Pliska, Amino acid side chain parameters for correlation studies in biology and pharmacology. *Int. J. Pept. Protein Res.* **32**, 269–278 (1988).

19. S. Betz, R. Fairman, K. O’Neil, J. Lear, W. Degrado, Design of two-stranded and three-stranded coiled-coil peptides. *Philos. Trans. R. Soc. Lond. B. Biol. Sci.* **348**, 81–88 (1995).

Chapter 3: Examining the multi-conformational landscape of the HAMP domain in signal transduction.

Abstract

In previous chapters, we establish that the HAMP, in its most thermodynamically stable state, inhibits the activity of the autokinase domain. In this chapter, we examine if the HAMP domain inhabits 2 structurally distinct signaling states or has multiple *structural* states that can be categorized into one of two *functional* states. The following work is a reproduction of a publication that addresses this question by employing a novel microfluidic DNA library building technique developed by the Abate lab to build a library of structurally distinct inputs into the autokinase of a Gram-negative HK, CpxA, as well as leucine scan of the S-Helix motif that connects it to the autokinase. A library of 9x9x8 = 648 distinct structural inputs were evaluated for their effect on the output of the CpxA autokinase as a function of 8 leucine point mutants in the S-Helix. We find that we cannot relate the patterns of activity that arise as a function of these structural inputs to two distinct HAMP conformations. Instead, it appears that the HAMP is able to inhabit a continuum of structures that depend on adjacent domains.

Contributions

This chapter is a reproduction of the manuscript, lark IC, Mensa B, Ochs CJ, Schmidt NW, Mravic M, Quintana FJ, DeGrado WF, Abate AR. Protein design-scapes generated by microfluidic DNA assembly elucidate domain coupling in the bacterial histidine kinase CpxA. Proc Natl Acad Sci U S A. 2021 Mar 23;118(12):e2017719118. The manufacture of the microfluidic device that was used in this study was conducted by the Abate lab (Drs. Iain C. Clark, Christopher J. Ochs).

The reporter library plasmid used in this study as well as the DNA parts used in the gateway assembly of the library were generated and validated by Dr. Nathan W. Schmidt. Library synthesis, selection by FACS and subsequent deep-sequencing of selected populations was conducted by Iain C. Clark. Dr. Clark and I were primarily responsible for the analysis of the sequencing library, and I was primarily responsible for interpretation of sequencing data as it relates to histidine kinase function and structure. I also conducted the experimental validation of library parts C and D in wild type CpxA by FACS. Dr. Marco Mravic conducted molecular dynamics simulations of library parts A and B, and the parsing of resulting structural features. Dr. Clark conducted hierarchical clustering of library data. Writing and review of the manuscript was conducted primarily by Dr. Clark and myself, with contributions from Dr. Mravic for sections regarding molecular dynamics and with the supervision and input of Drs. William F DeGrado and Adam R Abate.

Introduction

Protein engineers and synthetic biologists create and study biological function through design of nucleic acids, typically assembled from separate coding or non-coding building blocks. In most cases, it is difficult to predict which assembled DNA sequences will yield the desired function. Instead, DNA libraries containing large numbers of variants are screened for the trait of interest. This approach is used to optimize expression of promoter-gene-terminator combinations (1), tune biosynthetic circuits to balance metabolic flux (2, 3), and map specific amino acid changes to protein function (4-6). Molecular methods for producing DNA libraries include random mutagenesis, mutational scanning, and assembly of pre-existing sequences. Pre-defined parts allow control over the location and type of variants constructed, guaranteeing interesting mutations in the library, and excluding non-functional ones. Together with DNA synthesis, combinatorial assembly eliminates the codon bias of random mutagenesis and facilitates construction of gene circuit and multi-domain protein libraries (7-10). Consequently, a suite of molecular methods are now available to assemble pre-defined DNA parts into combinatorial libraries, including Golden gate (11), Gibson (12), SLICE (13), SLIC (14), and BioBrick (15).

Generating libraries in a one-pot assembly of DNA parts is simple and commonplace but has shortcomings. These assembly reactions are subject to bias because all parts are free to react with all others; thus, DNA parts that efficiently combine are overrepresented in the final library. In addition, because all combinations are possible, such approaches generate massive libraries that cannot be thoroughly screened. For a library of *m* components and *n* variants per component, n^m possible ordered combinations exist. For instance, even a relatively simple biosynthetic pathway comprised of four genes, each with its own promoter and ribosomal binding site (12 parts), and 4 variants per part, would result in over 16 million unique combinations. If 6 variants per part are used, a modest number when seeking to enhance a pathway, over 2 billion combinations are possible. For libraries of this size and larger, it can be difficult to assay a significant fraction of members even in pooled high throughput screens, for example by fluorescence-activated cell sorting (FACS). However, it is likely that most of the sequence space contains non-functional variants that neither need to be constructed nor screened. This has driven approaches that limit variant assembly based on intuition or modeling (16), and therefore concentrate experimental effort on subsets most likely to be functional (**Figure 3.1A**). Rationally reduced libraries can be constructed with many different assembly reactions that contain different starting parts. However, this manual approach is not scalable with current technology. Moreover, methods to automate DNA synthesis by microfluidics, although promising, have yet to produce large subset libraries at high throughput (17-20). Thus, a system that precisely generates targeted libraries from a complex part list would enable large combinatorial sequence spaces to be screened at a fraction of effort and cost.

To address this challenge, we develop a microfluidic system that can be programmed to create specific combinations of pre-defined DNA parts at high throughput and with minimal reagent consumption. Our system uses microfluidic valves to combine specific DNA parts, followed by compartmentalization into individual water-in-oil droplets to perform isolated assembly reactions. The system steps through desired DNA combinations sequentially and selectively (**Figure 3.1B**) at about one per second. Less than 150 µL is consumed to produce 5,000 DNA combinations, a 300-fold reduction in reagent consumption compared to automated well plate assembly. The resultant droplet reactors, each of which contains all requisite parts and enzymes to construct a unique variant, are thermally incubated in parallel for molecular assembly. Additionally, droplet compartmentalization suppresses bias since efficient reactions do not compete with less efficient ones, and all have ample time to achieve saturation. Instead of relying on hybridization to co-localize oligonucleotides (21), a process that can be difficult to optimize as library complexity increases, our approach uses programmed microfluidics to create diverse and even libraries from pre-defined oligonucleotide parts at optimal concentrations. This process can generate libraries of any desired combination from starting parts, and greatly reduces reagent consumption and screening effort when complex combinations of DNA parts are needed.

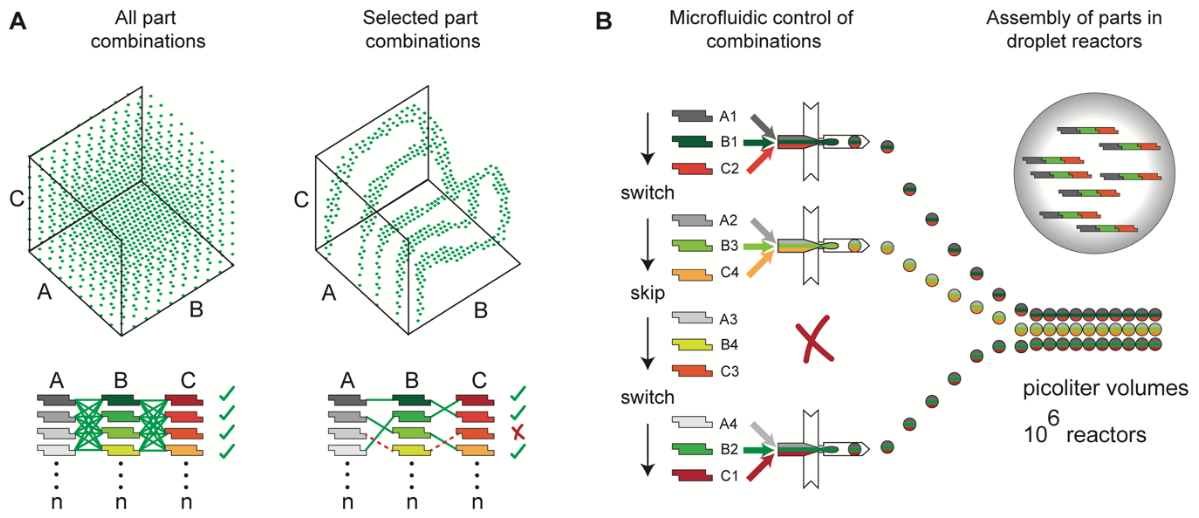


Figure 3.1 - Generation of DNA libraries of pre-defined composition. Non-random walks through sequence space require libraries with pre-defined variants. **(A)** Visual representation of the sequence space. Left: combinations generated with a one-pot all-by-all approach yield, ideally, all combinations in the sequence space. Right: combinations generated by selectively combining specific parts allow pre-determined walks through the sequence space and reduced screening effort. **(B)** Microfluidic approach for generating subset libraries. Switching rapidly between inlet channels combines parts and enzymes for assembly inside droplet reactors.

Using this technique, we construct a four-part combinatorial library encoding multiple domains of the canonical bacterial histidine kinase (HK), CpxA (22-24). Transmembrane HKs are multidomain proteins that often relate the binding of an extracellular ligand to the activity of an intracellular kinase. The signal passes along an extended series of homo-dimeric signal-transducing domains, which communicate via coaxial two and four-helix bundles along the dimer interface. The signal transducing domains include a number of modules, such as HAMP, GAF domains and helical linkers, frequently swapped between HK family members (25). Here, we use protein design to investigate how systematic modifications to transmembrane (TM) and linker domains modulate the activity of the downstream catalytic domains. Addressing this question is key to understanding how HKs transmit conformational information through multiple linker domains. We replace the TM domain of CpxA with a series of engineered TM dimers, which vary in sequence, structure, and length of the cytoplasmic connector to the N-terminus of CpxA's signal-transducing HAMP domain. We also vary the sequence of the signaling helix (S-Helix) located at the C-terminus of the HAMP domain, just prior to the catalytic domain. The sequence of the intervening HAMP domain is held constant, allowing examination of how it responds to and transmits diverse structural signals. We find that the S-helix sets the basal rate of activity, which is modulated up or down in a roughly additive manner by changes in the linker, when examined in the population average. In addition, when the variants are examined individually, we find a single TM sequence can elicit opposing effects on the catalytic activity of different S-helix variants – enhancing activity in one variant and inhibiting in another. Thus, the intervening HAMP domain not only passively transmits signals, but fundamentally changes the signaling patterns depending on even single-site variants in the neighboring domains. We posit that this flexibility yields a spectrum of functional outputs as HKs vary in response to changing evolutionary pressures, providing a potential explanation for the retention of multiple transmitting domains throughout evolution.

Results

Microfluidic design and operation.

To enable rational construction of a multi-domain protein library, we design a multi-valve microfluidic device to encapsulate specific combinations of pre-synthesized DNA library parts in 90 µm (382 pL) droplets (**Figure 3.2A, Figure 3.8**). The device consists of 38 reagent inlets, each with its own microfluidic membrane valve (26) and an oil inlet for droplet formation. The valves are arranged in two arrays on either side of the microfluidic drop maker (**Figure 3.2B**, Valve array). Open valves dispense reagents, which flow into a T-junction drop maker. Two valves downstream of the drop maker direct droplets into off-chip waste or collection tubes (**Figure 3.2C**, Collection switch). Each fluidic channel is connected via flexible tubing to a single well in a 96 deep well plate housing DNA parts, enzymes, and buffer. The 96 well-plate is pressurized with a manifold (**Figure 3.8D**) that ensures all wells are at equal pressure. Pressurization of the manifold drives flow through serpentine resistor channels (**Figure 3.2D**, Pressure Balancing), fabricated on a separate layer of the chip, which maintains equivalent flow rates from each inlet channel and, thus, ensures controlled final reagent concentrations. The design is scalable since additional inlets can be added along the length of the central channel. Following the design considerations outlined in the SI Appendix Materials and Methods section, the device can scale to 100k libraries with minimal changes to design or operation.

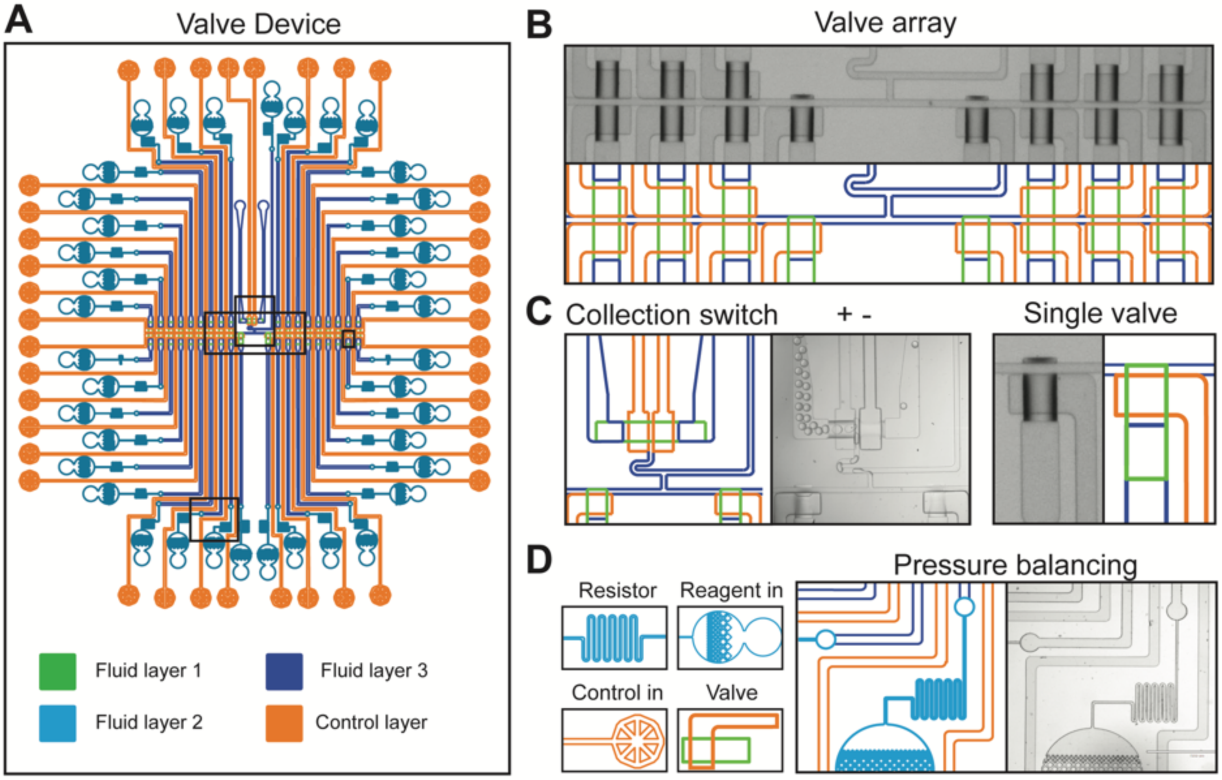


Figure 3.2 - Hybrid microfluidic device that uses valves and droplets to deterministically combine and assemble DNA parts. (A) Schematic of the device showing fluidic and control layers, inlets, valves and resistors. **(B)** Schematics and images of the fabricated device. The valve-array contains 38 reagent valves leading to a common channel where reagents are combined, and drops are formed. **(C)** The collection switch is a two-valve system for directing droplets to a collection tube or into waste. **(D)** Fluid flow rates in inlets with different locations on the main collection channel are equalized by pressure balancing with on-chip resistors.

Each microfluidic valve can be actuated in <10 ms (**Figure 3.3A-B**), as measured using a fluorescent dye and droplet cytometer (27), allowing rapid switching between inlets (**Movie 1**). The device operates on a ~720 millisecond (ms) duty cycle consisting of 500 ms of library collection, 200 ms of channel flushing to waste, with 10 ms to switch valves to the next part combination (**Figure 3.3A-C**). To start a cycle, the collection valve closes, and the waste valve opens (**Movie 2**). Four reagent valves open simultaneously, allowing enzyme master mix and DNA parts to flow through the central channel, into the drop maker, and out through the waste valve. After this flush, the collection valve opens, the waste valve closes, and droplets containing the desired library components are collected. To switch to a new part combination, the collection valve closes, the waste valve opens, and the cycle is repeated (**Figure 3.3D, Movie 3-4**). This cycle continues until all desired combinations have been generated. A library of ~10,000 defined combinations can be created in ~2 hours. Droplets containing DNA parts and enzymes are incubated off chip to complete the assembly reaction. Any DNA assembly molecular biology can be used, including those requiring temperature changes, because the droplets are stable to heat and collected in PCR tubes that can be thermocycled.

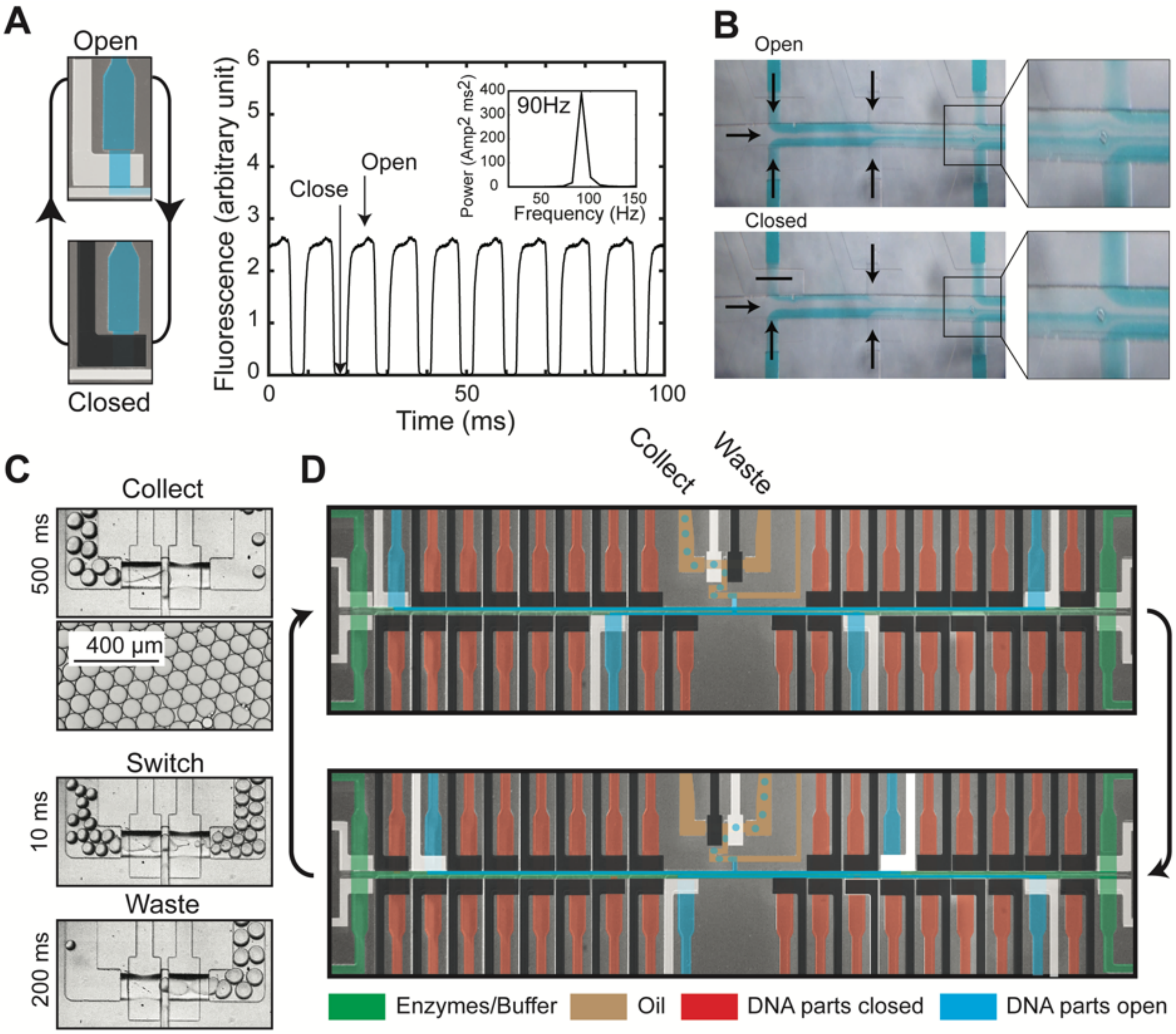


Figure 3.3- Valve array duty cycle. (A) Valve opening and closing speed measured by fluorescence intensity. **(B)** Dye labeled inlets show how reagents are combined in the central channel. Switching a single inlet from open (top) to closed (bottom) eliminates its stream from the central channel. **(C)** Flushing to waste reduces contamination from residual DNA parts from the prior cycle. **(D)** The device cycles between collecting drops (top) and wasting drops (bottom) as part of its duty cycle. Closed valves are colored black and open valves are colored white. Flowing DNA parts are colored blue, and blocked DNA parts are colored red. Enzymatic reagents are colored green. Oil is colored brown.

Rationale for the multi-domain CpxA library design

We use our microfluidic library construction method to build a multi-domain protein library from 4 DNA parts, each with 8 or 9 variants per part (**Figure 3.4A**). This library allows us to explore how diverse structural inputs from a designed transmembrane (TM) domain and cytoplasmic linkers influence the kinase activity of the bacterial histidine kinase (HK) CpxA (28). The transmembrane protein CpxA is a prototypical HK that senses cell envelope and protein folding stress in the periplasm (29). The wild type protein consists of (from the periplasmic to the C-terminal cytoplasmic kinase): 1) a periplasmic sensor domain; 2) an antiparallel four-helix transmembrane domain (TM); 3) a parallel four-helix bundle signal transducing HAMP domain; 4) a short S-Helix domain; and 5) the two conserved catalytic histidine kinase domains (**Figure 3.4B**). The signal-transducing domains are conserved and repeated in a modular fashion within many different HKs and combine to yield sensors for diverse molecular and environment cues. It is hypothesized that HKs function by a similar mechanism of transmitting conformational changes across their modular domains to influence kinase domain activity (30-36). Functional diversity is thought to be achieved by fine tuning HK domain conformational coupling via specific inter-domain protein geometry. Here, we rationally design a family of CpxA variants, which systematically vary in expected inter-domain geometry and coupling, to directly test the range of structural features that yield high-kinase activity. The designed components evaluated in this study consist (from N- to C-terminus) of: 1) an N-terminal MBP domain and HA affinity tag, 2) a set of 81 TM dimers of systematically varying structure, 3) a variable-length alanine linker, 4) the wildtype HAMP domain, 5) a series of eight mutants in the S-helix region and 6) the wildtype DHp and ATP binding domains of the CpxA catalytic domain (**Figure 3.4A-B**). The N-terminal MBP tag results in periplasmic location of MBP, and is often used in similar genetic screens to select for expression and proper membrane insertion (37, 38).

The TM helices contain a strong TM helix homo-dimerization "G-X_3-G" motif, taken from the protein Glycophorin A (GpA, LILL**G**VMA**G**VIGT)(39, 40). This TM design simplifies the conformational input propagated by the second TM helices of the native CpxA 4-helix bundle into a 2-helix bundle. The TM helix is constructed by combining two library variants (parts A+B), which allow flexibility in varying the length of the helix and the position of its GX_3G motif within the membrane (**Figure 3.9**). Furthermore, a single Trp residue, which prefers to localize near the headgroup region of the bilayer, is scanned through multiple positions of the C-terminal third of the TM helix to influence the location of the helix in the membrane, as in previous studies of HKs (**Figure 3.4A**) (41). The helical phase of the last residue of the TM helix is held roughly constant, but the length of the neighboring linker is changed to allow systematic variation of phase. Short linkers are generally present between the TM helices and cytoplasmic domains of HK proteins. We therefore place 0 to 7 helix-promoting alanine insertions in a linker (part C of the library, **Figure 3.4A-B**) between the TM helix and the cytoplasmic domains of $CpxA_{cyto}$ (HAMP, S-helix and catalytic domains). Assuming helicity is maintained, each Ala would result in a ~100 degree phase shift and ~1.5 Å axial translation of the terminal linker residue as it connects to $CpxA_{cyto}$. Finally, in library part D, we introduce a series of eight consecutive single-site leucine substitutions into the S-helix of $CpxA_{cyto}$ to alter its basal kinase activity. Evaluating the effects of TM-linker combinations over multiple S-helix variants offers a rigorous test for how conformational information coded by the upstream TM-linker variants couples through the native HAMP domain to the kinase domain. In all, this 5184-member library consists of 81 TM helices, 8 linkers and 8 S-helix variants.

Construction and screening the multi-domain protein library

We use our microfluidic system to generate a full library of all part combinations, as well as a library that has been reduced in size by restricting specific combinations of the TM (parts A and B). We use Golden Gate assembly (42) of DNA parts with BsaI recognition sites to produce libraries from four DNA parts (A,B,C,D) (**Figure 3.4A**, **Table 3.1**). There are 9 A parts, 9 B parts, 8 C parts and 8 D parts. Combinations of parts A and B range from 16 to 28 amino acids, but we hypothesize that some AB combinations, which encode a synthetic transmembrane helix, may not be well accommodated in the membrane. Therefore, in addition to generating a library from all parts (full library, 5184 possible variants), we restrict part AB combinations based on the rule 18 AA ≤ [A+B] ≤ 25 AA (subset library, 3648 possible variants). We sequence each library after on-chip assembly and cloning and compare the diversity and evenness to standard tube assembly. The library is designed such that multiple paired end 150 bp reads cover the entire ABCD sequence, allowing us to assess the abundance of library members by next generation sequencing. Libraries generated in droplets have a higher number of unique members: the on-chip constructed full (4351 of 5184, 84%) and subset (3120 of 3648, 86%) libraries have greater coverage and evenness than standard in-tube assembly (1280 of 5184, 25%) (**Figure 3.4C**). The subset library has substantially de-selected (29.9% to 2.7% of the total library) the undesired part combinations from the final clone library (**Figure 3.4D**, 18 AA ≤ [A+B] ≤ 25 AA, black boxes), demonstrating that our microfluidic system can construct libraries with pre-defined members.

To experimentally evaluate how these protein domains interact and influence CpxA signal transduction, we design a pooled assay to screen our engineered protein library (**Figure 3.4E**). On-chip libraries are cloned into a plasmid to create an intact, in-frame, designed *cpxA* protein-coding gene. The plasmid contains a constitutive promoter driving CpxA and a green fluorescent protein (GFP) reporter under the control of the *cpxP* promoter which is activated by CpxR, the cognate response regulator of CpxA. This is a reliable reporter of overall CpxAR activation (43). Libraries are expressed in *E.coli cpxA*::km (JW3882-1, The Coli Genetic Stock Center, Yale) and the GFP^{high} population is sorted by FACS, corresponding to ~4.86% of cells (**Figure 3.4E**). The abundance of library members before and after sorting is determined by Illumina sequencing and enrichment of each ABCD sequence is calculated (see Supplemental Materials and Methods).

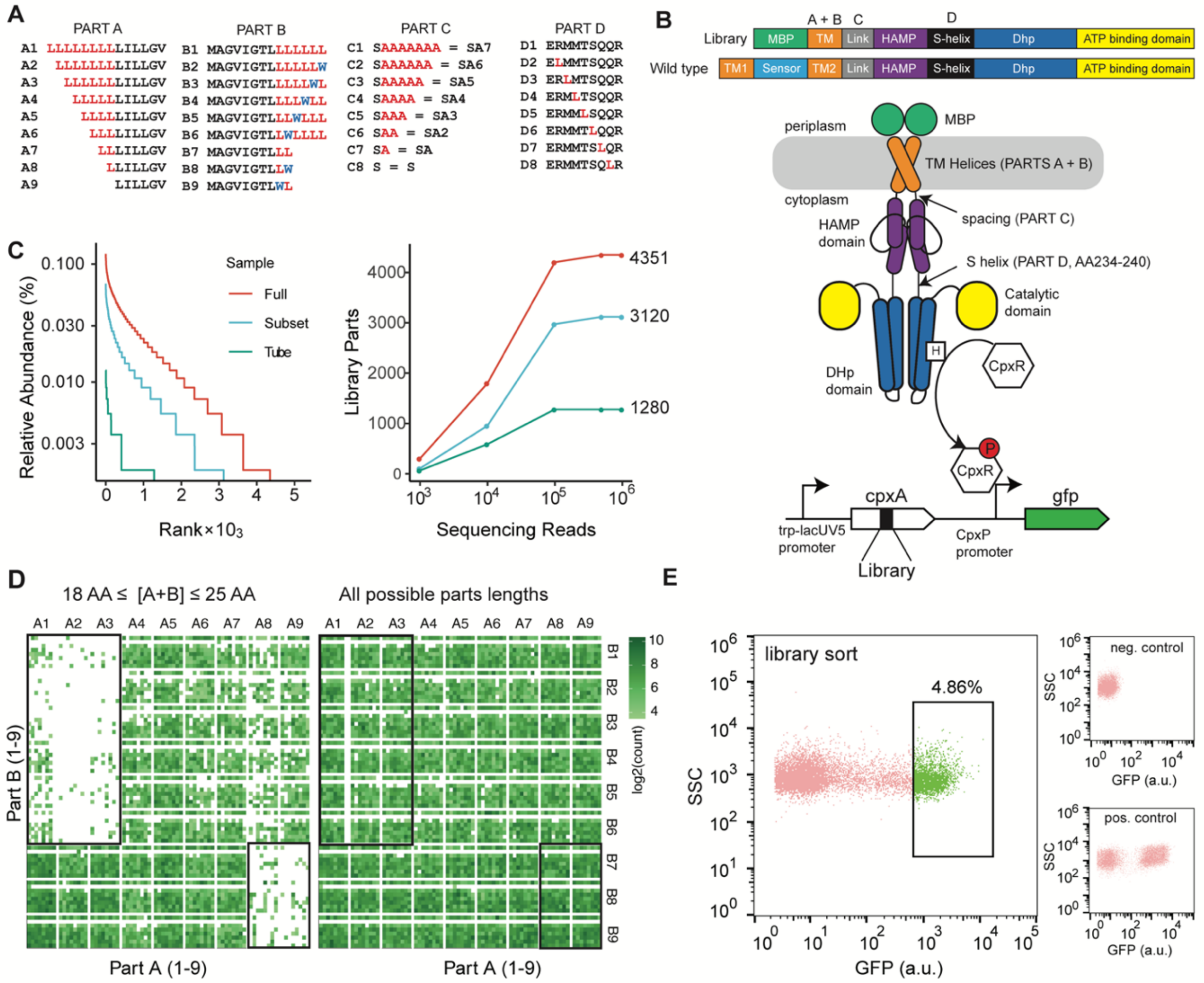


Figure 3.4 - Construction and sequencing of the DNA library demonstrates efficient on-chip library subsetting. (A) Sequences of the four parts used to construct the combinatorial DNA library. **(B)** Top: comparison of the synthetic and wild type CpxA structures. Bottom: The engineered CpxA contains maltose-binding protein (MBP) replacing the signal domain, followed by a two-helix transmembrane (TM) region encoded by parts A and B. The variable juxtamembrane linker (part C) between the TM and HAMP domain is followed by leucine substitutions in the S-helix (part D). CpxA phosphorylates the response regulator CpxR, which activates transcription of a GFP reporter via the cpxP promoter. **(C)** Left: Rank-abundance curves comparing the full library generated on-chip (full), full library generated in a tube (tube), and the subset library generated on-chip (subset). Microfluidic assembly enhances variant coverage compared to pooled, tube-based assembly. Right: Subsampling reads and counting library members quantifies diversity and confirms adequate sequencing depth. **(D)** Distribution of library sequences in the subset and full libraries, displayed as $\log_2$(read counts). Parts C and D are nested within parts A and B. Left: AB combinations constrained by the length of A and B such that 18 AA ≤ [A+B] ≤ 25 AA (subset). Right: No restriction is placed on which parts were combined to generate the library (full). **(E)** Sorting of GFP^{high} reporter cells expressing the CpxA library. Positive (CpxA L243S) and negative (CpxA Q229V) controls are shown.

The phenotypic effects of variations in the A – D library components

Our library construction and screening method allows us to evaluate variations in one library part, while either varying the sequence of the other parts or holding them constant. The resulting activity and enrichment profiles can be evaluated in the context of different models of conformation coupling. For example, the interplay between the effects of substitutions in the S-helix, which set the basal kinase level, and the TM-linker informs models for coupling through the HAMP domain to the DHp. At one extreme, the HAMP can be postulated to have just two conformations whose relative energetics change in response to structural transitions of upstream signaling domains (in this case the different variants coded by parts A-C). In this "two-state HAMP model", each TM-linker variant would have the same effect on the activity each of the S-helix variants, and vice versa. At the other extreme, the "multi-state HAMP model" the HAMP might have a more dynamic structure, which continuously varies in response to the TM-linker input conformation (44). In the HAMP multi-state model, different inputs from the TM helix-linker would be expected to induce differing conformations in the HAMP domain, which might couple differently to mutants of the neighboring S-helix. In this scenario we would see different relative activity patterns for the various S-helix mutants in response to structural changes in upstream domains, arising from the complex coupling of a multi-state HAMP to the adjacent S-helix (and *vice versa*).

We first analyze enrichment of the individual members of two library components, averaged over the remaining variants. For example, **Figure 3.5A** shows the enrichment for each of the eight S-helix variants (part D) as a function of the eight linkers (part C), in each case averaged over all possible A and B components. Other pairwise combinations are considered in

Figure 3.5 to identify which components contribute most to variations in enrichment. When viewed in this context, the most important determinant of enrichment score is the identity of part D, corresponding to the S-Helix (**Figure 3.5A**). Given its position between the wildtype HAMP and DhP domains, it is not surprising that S-mutants have a large effect on the activity of the kinase. There is also high concordance between the enrichment results from the two libraries (full and subset), which were independently generated, screened, and sequenced. Importantly, the enrichment profiles for S-helix mutants move up and down in concert as the linker length is systematically varied. This matches the result expected in the HAMP two-state model (44-46). To determine the effect of these mutations in full-length native CpxA we introduce each leucine mutant into WT protein and test each mutant's activity using a *cpxP*::GFP reporter strain by flow cytometry. These CpxA mutants display the same pattern of GFP fluorescence as the library enrichment (parts A+B averaged, part C = WT) (**Figure 3.5E**). The pattern, as a function of the leucine substitution position, is consistent over many replicates (**Figure 3.10A**).

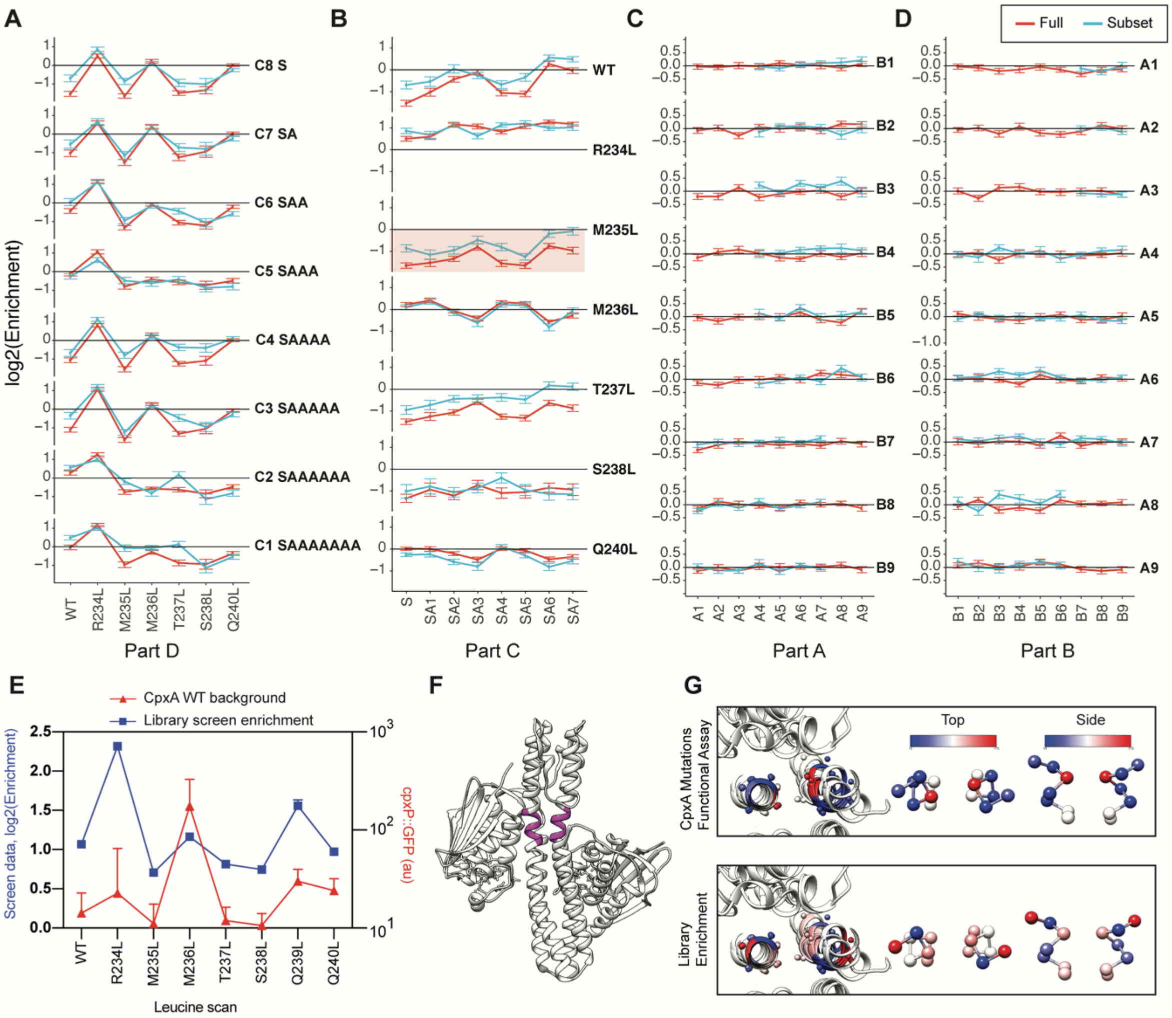


Figure 3.5- Expression and screening of the DNA library encoding CpxA identifies the S-helix as a major determinant of signaling. (A) Enrichment as a function of leucine substitution (part D) for each alanine insertion (part C) (averaged over parts A and B). **(B)** Enrichment as a function of linker alanine insertion (part C) for each Leu substitution (part D) (averaged over parts A and B). **(C)** Enrichment as a function of part A for each part B (averaged over parts C and D). **(D)** Enrichment as a function of part B for each part A (averaged over parts C and D). **(A-D)** Enrichment is calculated as the fold change in the normalized abundance of each variant sequence between post and pre-sort. Error bars are standard error, calculated as the mean divided by the square root of the sample size. Blue is the subset Library and red is the full library. Q239L (full and subset) and 18 AA ≤ [A+B] ≤ 25 AA (subset) were removed from the library because of low counts. **(E)** Comparison of screen (enrichment) data to the functional cpxP::GFP assay. For the functional assay, L mutants are made in otherwise wild-type CpxA. **(F)** Location of the S-helix in CpxA. g) Mapping of reporter fluorescence (top) and enrichment scores (bottom) on the S-helix of aligned CpxA structures (PDB: 4BIV, 4BIU chains A/B, D/E) shows enriched L variants segregate to the core of the dimer, and de-enriched variants on the outward face.

Non-additive effects of the transmembrane and juxtamembrane linker on enrichment and kinase activity

While S-helix variants are the dominant contributor to library enrichment and appear to respond additively to substitutions in a manner consistent with the two-state model, in-depth analysis also reveals more subtle deviations from the expectations of such a simple model. The linker length made significant, although less dominant, contributions to library enrichment, as shown in the profiles generated for linkers of differing length while holding the S-helix constant and averaging over parts A and B (**Figure 3.5B**). For example, while M235L is deleterious to signaling, the linker sequence modulates activity up or down (**Figure 3.5B**, shaded red). Earlier studies of HK's have shown that when residues are inserted into the helical inter-domain linkers of signaling domains the transcriptional activity is modulated by linker length in a sinusoidal manner, with a repeat roughly matching that of the alpha-helix (47-50). We observe similar sinusoidal results in both libraries, but the phase changes between different S-helix mutants (**Figure 3.6B**). Thus, linker variants that promote HK activity in one S-helix variant inhibit it in another. This finding departs from the expectations of a strict two-state coupling model.

We also examine the effects of varying parts A and B, which together comprise the TM helix. Parts A and B do not appear to significantly affect enrichment when they were examined individually while averaging over parts C and D (**Figure 3.5C-D**). However, examination of individual datasets shows large variations in the response to A and B variants, depending on the nature of the C and D components (**Figure 3.10B**). To better understand the origin of this variation we perform hierarchical clustering of the enrichment data, grouping sequences with similar variation with respect to part C (**Figure 3.6**). This allows us to understand how individual members of the A and B components cooperate with the linker Ala-insertions to affect activity in the context of different S-helix variants. The periodic effects of alanine insertion observed in the averaged background (**Figure 3.5B**) become even more pronounced in the individual clusters (**Figure 3.6**), and, again, the effects are consistent between the screen replicates (full vs. subset library). Furthermore, we remake A-insertion mutations in full-length native CpxA and test each mutant's activity using the *cpxP*::GFP reporter assay (**Figure 3.10C**). Similar periodicity is observed in this dataset to patterns seen in cluster 15 and in the M236L (averaged over A and B) (**Figure 3.10D**), suggesting that linker effects are reproducible, but highly dependent on domain context.

The periodicity and phase of insertional profiles provides important information about the geometry of the helical linkers between signaling domains. The phase relates to the location of the Ca atoms as they wind around an a-helix, while the period relates to the inter-helical interaction pattern in a helical dimer. For example, a parallel pair of canonical a-helices has a repeat of 3.6 residues, while a left-handed coiled coil (left-handed crossing near 20°) has a 3.5-residue repeat, and a right-handed glycophorin-like motif (right-handed crossing near 40°) has a period of 4.0. These periods can also be modulated by nonideality of the α-helix, particularly when the registry of the input and output helices being connected are not easily spanned in a helical conformation. To determine the phase and period we fit the enrichment profiles for the linkers to a sine function. The majority of the highly populated clusters were well described by a sine wave with repeats within 3 to 4.2 residues (**Figure 3.6**, **Dataset 1**), which is within the range seen in helical dimerization motifs, including the dimeric helical linkers in HKs (51). Since clusters contain different S-helix and TM substituents, the range of phases observed is consistent with the observation that linker variants that promote HK activity in one TM/S-helix context often inhibit it in another.

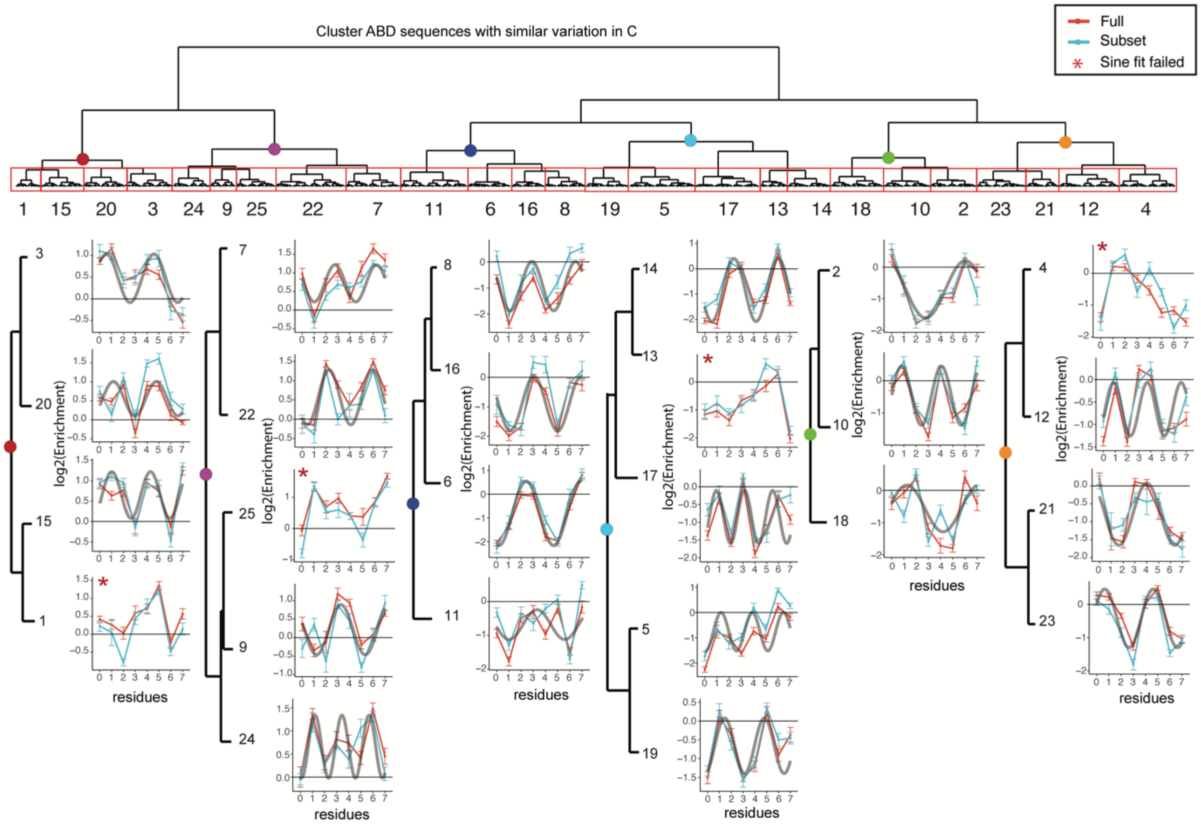


Figure 3.6 - Analysis of variable length juxtamembrane Ala linkers (part C) in the context of different TM and S-helix domains. ABD sequences with similar variation in part C are clustered and enrichment is plotted as a function of the linker (part C). Sine curves fitted to data are superimposed on each dataset. The period of each sine curve falls within the expected range of 2.7 to 4.2, with few exceptions.

Structure-function links revealed by all-atom molecular dynamics simulations of the synthetic transmembrane domain

Next, we next used molecular dynamics (MD) simulations to determine which structural properties of the TM dimers encoded by parts A and B associate with the different behaviors seen in the functional clustering. The synthetic transmembrane domain built by parts A and B encodes a G-X_3-G motif with variable leucine spacers and a shifted C-terminal tryptophan (**Figure 3.4A**). This results in 81 variants with sequence lengths between 16 and 28 residues, which are expected to alter the depth of the crossing motif in the membrane, its crossing angle, and the Ca-Ca distance at the end of the TM region. To structurally define the expected conformations of the 81 TM variants, we ran 80 nsec all-atom MD simulations of each TM dimer in 1-palmitoyl-2-oleoyl-sn-glycero-3-phosphocholine (POPC) lipid bilayers (**Figure 3.7A**) in explicit solvent with CHARMM36 parameters (52) as implemented in GROMACS 2018.

Across all simulations, the TM segments (Parts A + B) are generally stable (**Figure 3.9B-E**) and hold the intended initial dimeric conformation: the canonical G-X_3-G-type interaction with close interhelical backbone interactions (3.5 - 4.5 Å) with a right-handed helix-helix crossing angle. The mean RMSD over TM amino acids between final versus initial simulation frames is 1.3 Å. For longer TM helices, structural re-arrangements accommodate the increased length without significantly disrupting the strong G-X_3-G motif interaction. **Figure 3.9A** shows three representative cases of significant deviations from the starting structure for TM sequences that are longer than the canonical 20-residues of single-pass TM helices; the helices widen their crossing angle (**Figure 3.9A**, left) or adopt helix kink distortions (**Figure 3.9A**, middle and right) at the N-terminal half of the dimer. These conformations help maintain their apolar sidechains embedded in lipid. However, because the structurally stable G-X_3-G motif lies close to the cytoplasmic side of the membrane, the grossest structural variation occurs primarily on the outer leaflet of the bilayer, distant from the region that connects to the linker and HAMP domains (**Figure 3.9A**). Therefore, in the future it would be desirable to design libraries that include a signal-responsive periplasmic sensor domain and remove sequences that encode physically unreasonable extensions of the TM helix as in our subset library.

To allow comparison with phenotypically defined clusters, we compute structural and geometric features likely to influence the conformation and energetics of the downstream domains (**Figure 3.7B**). Specifically, we clustered the 81 models based on the depth of the G-X_3-G and the C-terminal end of the helices in the membrane, the interhelical crossing angle, the Ca-Ca distances near the end of the TM region, the helical registry and the degree to which the dimers deviate from C2 symmetry. The TM domain geometries cluster into eight structurally distinct groups (**Figure 3.11A**), each expected to pose a unique conformational input to the subsequent juxtamembrane linker (part C) and HAMP domain, and to influence inter-domain coupling and kinase activity (53).

We next compare how structurally similar TM domains relate to functionally similar clusters from the screen. We cluster the screen enrichment data based on sequences with similar variation in downstream domains (parts C and D) to find which A+B combinations lead to similar functional outcomes. This procedure leads to phenotypically similar clusters, in which each member is identified by its A and B parts (**Figure 3.11B**). We evaluate the overlap between the individual AB combinations within structural and functional clusters. The resulting plot shows that functionally similar AB sequences correlate with structural similar ones (**Figure 3.7C**). For example, the two classes of TM helical pairs with the G-X$_3$-G located closest to the cytoplasm (structure class 7 and 8) associate with only a limited number of phenotypic outcomes (predominantly function classes 3+4) based on the cooccurrence of the same AB sequences in the structural and functional clusters parts (**Figure 3.11C**). These findings establish a clear structure-function link between TM geometry and signaling output. It is, however, difficult to determine a single parameter that uniquely controls signaling, in part because of the interdependence of the geometric features governing the structure. For example, helical ends can be shifted up or down by changing the interhelical crossing angle (scissoring motion) or by altering the depth caused by C-terminal Trp substitutions and Leu insertions. Additional structural studies with the C-terminal domains attached will be required to decipher the relative contributions of these and other features to the activation of the kinase domain.

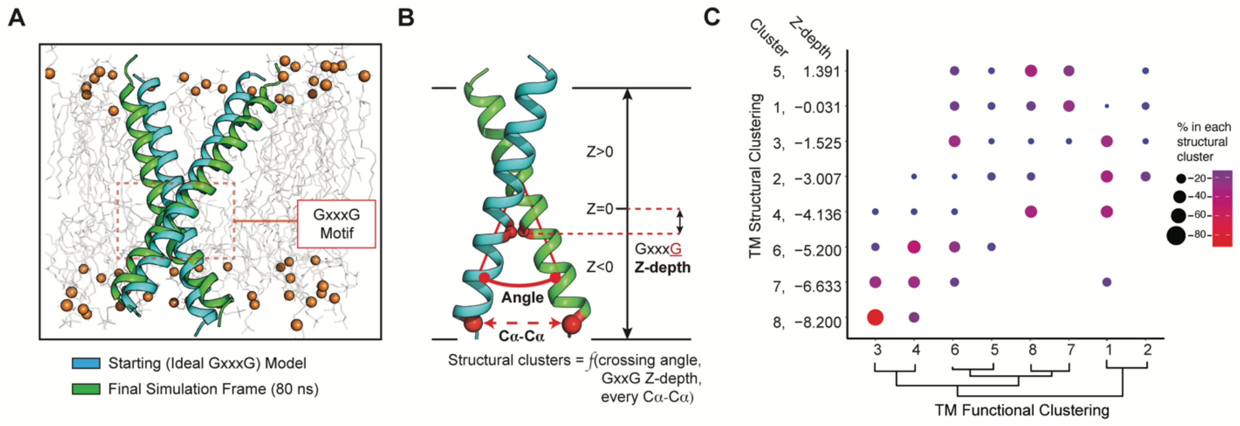


Figure 3.7- Structure-function relationships revealed by molecular dynamics (MD) simulations. (A) MD simulations predict all TM dimer geometries simulated in lipid bilayers, shown as cartoon ribbon mainchain traces (initial model, cyan; final MD frame, green; lipid headgroup phosphates, orange spheres). A sequence with the largest backbone RMSD of the final MD simulation frame (80 ns) versus the initial models is shown as a representative example. **(B)** Visual representation of the TM dimer and structural parameters extracted from MD simulation data. TM domain structural features are predicted based on MD simulations for each AB combination (also see **Figure 50F**). **(C)** TM clustering based on structural features compared to TM clustering based on functional enrichment (screen). Functionally similar AB sequences segregate with structural similar ones.

Discussion

We report a microfluidic system that deterministically combines DNA parts for enzymatic assembly. The microfluidic system can be programmed to generate any library from a set of starting parts, and therefore allows non-random walks through sequence space, both in composition and abundance. Because the system uses physical compartmentalization instead of sequence or molecular biology optimization, it is fully compatible with such advances as they arise. The system uses resistors to ensure balanced flow rates, allowing additional part inlets to be added along the main channel to scale synthesis to more complex assemblies. An additional feature of our approach is that it allows biasing of the quantities of each part combination by adjusting the collection times to favor specific library members. If higher representation of specific part combinations is desired, the device can be programmed to collect a larger number of droplets for those combinations. Thus, our microfluidic construct assembler represents an efficient tool for the generation of precise combinatorial libraries with applications in biosynthetic pathway design, protein engineering, and deep mutational scanning.

We apply our microfluidic system to study signal transduction in the prototypical bacterial histidine kinase CpxA. Our microfluidically generated library allows us to ask how geometric “signals” from engineered membrane-spanning domains transmit through HAMP and S-helix domains to a four-helix bundle that serves as the histidine phosphoryl acceptor in the phosphorelay catalytic mechanism of the HKs. Addressing this question is key to understanding not only how HKs signal, but why they incorporate multiple linker domains (53-56). What evolutionary advantage is there to retaining and shuffling signal transduction domains and interdomain linkers that have no apparent function in binding or catalysis? They appear to "passively" transmit signals in most proposed mechanisms, but it is hard to understand the conservation of multiple inserted domains rather than a single more concise connection.

To probe the mechanism of conformation coupling at a distance, the library scans Leu substitutions across the S-helix, located one full domain (>20 Å) away from the connection between the TM-linker and the N-terminus of the HAMP. The S-helix variants contribute most to the phenotypic variation in the library, as expected from the S-helix's proximity to the catalytic center. This dominant effect was also confirmed in native full-length CpxA (**Figure 3.5E**). When enrichment scores are mapped onto available CpxA cytoplasmic structures (pdb: 4BIV, 4BIU), high-signal variants predominantly fall on the dimeric interface of the two-helix bundle of the S-helix while low-signal variants fall on the outside (**Figure 3.5E-F**). Substitutions to outward facing residues near contacts with the ATP binding domains of the kinase also impacted activity. Together, the Leu substitutions in the S-helix served their intended purpose of significantly altering the energetic landscape of the intermediates required in the enzymatic cycle of the catalytic domain (i.e. autphosphorylation, phosphotransfer and nucleotide exchange) (48), providing a range of basal activities in both the WT and engineered CpxA library.

When evaluated over the average of the TM variants, the phenotypes of the S-helix and Ala-linker library members covary in an additive manner. The S-helix sets a baseline level of signaling, and the spacer modulates activity around that baseline. This behavior is expected from a simple two-state model, in which the HAMP has two different conformations that can either promote or inhibit kinase activity. In this model, the energetic difference between the two states – but not their structures – are altered by changes in the sequence and structure of the upstream TM-linker. However, by examining the phenotypic variation of each member of the library, we are able to also observe significant non-additive behavior, in which mutations in one domain can have opposing effects, depending on the sequence of the distant domain. This is seen in the varied patterns of enrichment observed when the juxtamembrane linker is changed in different S-helix backgrounds (**Figure 3.5B, Figure 3.6**). Such behavior is more consistent with the possibility that HAMP domains have multiple conformations, which can vary significantly with respect to the input conformation from the TM domain and linker, and couple differently to the multiple conformational states populated as the kinase transitions through functional states (51, 57-59). In depth MD simulations of the TM variants confirm a correlation between structural output from the membrane domain and the corresponding profile of activity of the corresponding variants (**Figure 3.7**). This finding encourages us to speculate on the consistency of our phenotypic screening with structural mechanisms of kinase activation (30, 32-35).

Our results suggest that transducing domains provide opportunities to not only vary the basal levels of HKs, but to also more radically change their energy landscapes and signaling patterns in response to evolutionary pressures. This is consistent with other reports documenting multiple continuously varying conformers of variants of HAMP domains (58). It also is consistent with the large structural changes seen in CpxA, as it transitions from a symmetrical resting state, to an asymmetric Michaelis complex, to the intermediate covalent phospho-histidine intermediate, which then binds and transfers the phosphoryl group to its cognate response regulator. Finally, CpxA also has a distinct phosphatase activity, and it is the balance of the phosphatase to the kinase activity that sets the overall transcriptional response. In this view, a multi-conformational HAMP domain would engender functionally diverse outcomes to sequence alterations by allowing differential coupling to each of the individual functional structural states of the catalytic core machinery.

Conclusion

We describe a new microfluidic system for the construction of rationally reduced DNA libraries. Our approach combines valves with droplet microfluidics to rapidly select and enzymatically assemble pre-defined DNA parts into constructs. This approach allows construction of libraries with targeted part combinations automatically, at high speed, and with low reagent usage. Thus, our microfluidic construct assembler affords a facile way to take efficient walks through sequence space. We use the system to assemble a multi-domain protein library of the canonical bacterial sensor kinase, CpxA. We show that mutations to the S-helix of CpxA globally change kinase activity by directly altering the stability of catalytic-states in the kinase domain, which are then energetically modulated by an interplay between sequence features in the TM and linker domain. Thus, we observed additive two-state coupling as the dominant theme, but with significant contributions to non-additive components, consistent with multi-state signaling through the HAMP module.

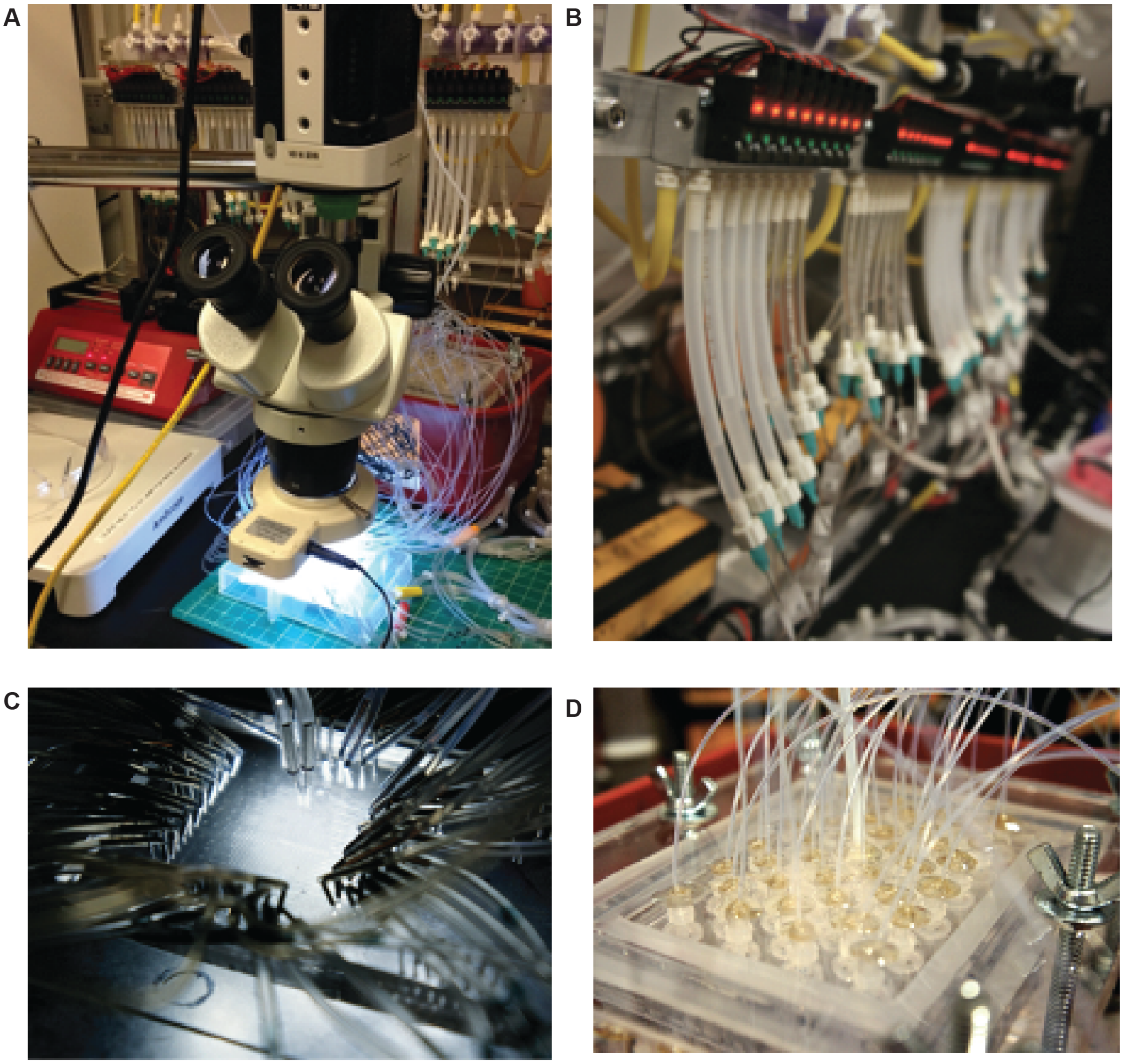
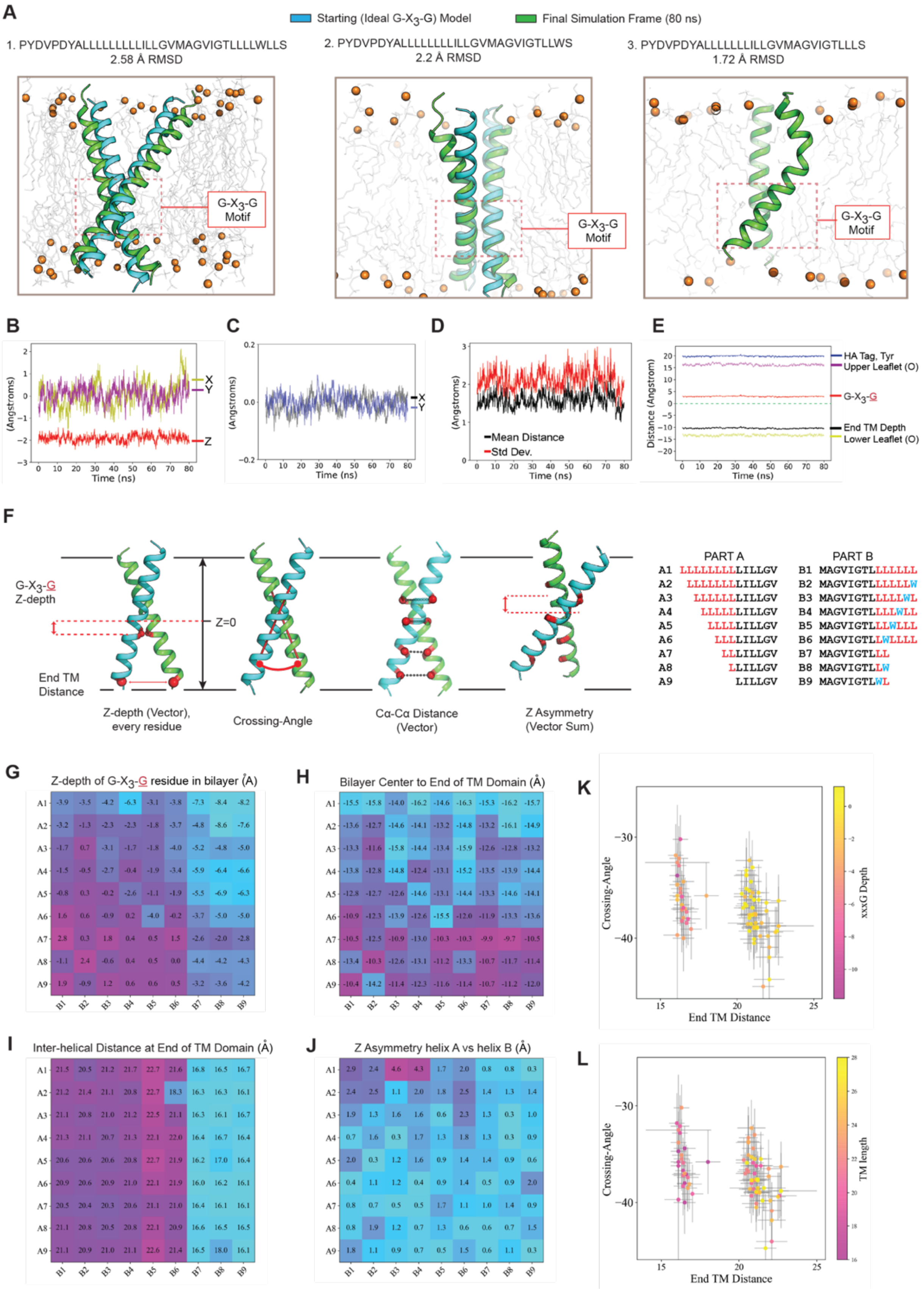


Figure 3.8- Instrumentation and microfluidic hardware. (A) Microscope and camera used to observe valve operation. **(B)** solenoid valves that control the on-chip valves**. (C)** Image of the microfluidic chip with control air and reagent lines. **(D)** Pressurized manifold to evenly dispense DNA parts and reagents to the microfluidic chip.



PART A		PART B	
A1	LLLLLLLLLILLGV	B1	MAGVIGTLLLLLLL
A2	LLLLLLLLILLGV	B2	MAGVIGTLLLLLLW
A3	LLLLLLLILLGV	B3	MAGVIGTLLLLLWL
A4	LLLLLLILLGV	B4	MAGVIGTLLLLWLL
A5	LLLLLILLGV	B5	MAGVIGTLLLWLLL
A6	LLLLILLGV	B6	MAGVIGTLLWLLLL
A7	LLLILLGV	B7	MAGVIGTLLL
A8	LLILLGV	B8	MAGVIGTLLW
A9	LILLGV	B9	MAGVIGTLWL

G Z-depth of G-X3-G residue in bilayer (Å)

	B1	B2	B3	B4	B5	B6	B7	B8	B9
A1	-3.9	-3.5	-4.2	-6.3	-3.1	-3.8	-7.3	-8.4	-8.2
A2	-3.2	-1.3	-2.3	-2.3	-1.8	-3.7	-4.8	-8.6	-7.6
A3	-1.7	0.7	-3.1	-1.7	-1.8	-4.0	-5.2	-4.8	-5.0
A4	-1.5	-0.5	-2.7	-0.4	-1.9	-3.4	-5.9	-6.4	-6.6
A5	-0.8	0.3	-0.2	-2.6	-1.1	-1.9	-5.5	-6.9	-6.3
A6	1.6	0.6	-0.9	0.2	-4.0	-0.2	-3.7	-5.0	-5.0
A7	2.8	0.3	1.8	0.4	0.5	1.5	-2.6	-2.0	-2.8
A8	-1.1	2.4	-0.6	0.4	0.5	0.0	-4.4	-4.2	-4.3
A9	1.9	-0.9	1.2	0.6	0.6	0.5	-3.2	-3.6	-4.2

H Bilayer Center to End of TM Domain (Å)

	B1	B2	B3	B4	B5	B6	B7	B8	B9
A1	-15.5	-15.8	-14.0	-16.2	-14.6	-16.3	-15.3	-16.2	-15.7
A2	-13.6	-12.7	-14.6	-14.1	-13.2	-14.8	-13.2	-16.1	-14.9
A3	-13.3	-11.6	-15.8	-14.4	-13.4	-15.9	-12.6	-12.8	-13.2
A4	-13.8	-12.8	-14.8	-12.4	-13.1	-15.2	-13.5	-13.9	-14.4
A5	-12.8	-12.7	-12.6	-14.6	-13.1	-14.4	-13.3	-14.6	-14.1
A6	-10.9	-12.3	-13.9	-12.6	-15.5	-12.0	-11.9	-13.3	-13.6
A7	-10.5	-12.5	-10.9	-13.0	-10.3	-10.3	-9.9	-9.7	-10.5
A8	-13.4	-10.3	-12.6	-13.1	-11.2	-13.3	-10.7	-11.7	-11.4
A9	-10.4	-14.2	-11.4	-12.3	-11.6	-11.4	-10.7	-11.2	-12.0

I Inter-helical Distance at End of TM Domain (Å)

	B1	B2	B3	B4	B5	B6	B7	B8	B9
A1	21.5	20.5	21.2	21.7	22.7	21.6	16.8	16.5	16.7
A2	21.2	21.4	21.1	20.8	22.7	18.3	16.3	16.3	16.1
A3	21.1	20.8	21.0	21.2	22.5	21.1	16.3	16.1	16.7
A4	21.3	21.1	20.7	21.3	22.1	22.0	16.4	16.7	16.4
A5	20.6	20.6	20.6	20.8	22.7	21.9	16.2	17.0	16.4
A6	20.9	20.6	20.9	21.0	22.1	21.9	16.0	16.2	16.1
A7	20.5	20.4	20.3	20.6	21.1	21.0	16.4	16.1	16.1
A8	21.1	20.8	20.5	20.8	22.1	20.9	16.6	16.5	16.5
A9	21.1	20.9	21.0	21.1	22.6	21.4	16.5	18.0	16.1

J Z Asymmetry helix A vs helix B (Å)

	B1	B2	B3	B4	B5	B6	B7	B8	B9
A1	2.9	2.4	4.6	4.3	1.7	2.0	0.8	0.8	0.3
A2	2.4	2.5	1.1	2.0	1.8	2.5	1.4	1.3	1.4
A3	1.9	1.3	1.6	1.6	0.6	2.3	1.3	0.3	1.0
A4	0.7	1.6	0.3	1.7	1.3	1.8	1.3	0.3	0.9
A5	2.0	0.3	1.2	1.6	0.9	1.4	1.4	0.9	0.6
A6	0.4	1.1	1.2	0.4	0.9	1.4	0.5	0.9	2.0
A7	0.8	0.7	0.5	0.5	1.7	1.1	1.0	1.4	0.9
A8	0.8	1.9	1.2	0.7	1.3	0.6	0.6	0.7	1.5
A9	1.8	1.1	0.9	0.7	0.5	1.5	0.6	1.1	0.3

Figure 3.9- Molecular dynamics (MD) simulations. (A) MD simulations predict all TM dimer geometries maintain the inter-helical G-X_3-G interaction but adopt structural re-arrangements. Transmembrane helix dimers simulated in lipid bilayers, shown as cartoon ribbon mainchain traces (initial model, cyan; final MD frame, green; lipid headgroup phosphates, orange spheres). The sequences with the two largest backbone RMSDs of the final MD simulation frame (80 ns) versus the initial models are shown as representative examples: (1) and (2). Both adopt unique structural rearrangements but maintain the close inter-helical interaction of the G-X_3-G motif with near-canonical geometry. (1), an increased inter-helical crossing angle; (2), modest helical kink N-terminal to the G-X_3-G motif. (3), an additional representative example of more pronounced helix kinking towards the N-terminal third of the TM helix, whereas the C-terminal portion containing the G-X_3-G motif maintains close interaction with its partner TM domain (right-handed crossing angle ~ -35°), denoted by red dashed region. **(B-E)** Representative example of calculated bilayer plane stability and relative features. **(B)** Bilayer midline plane calculated from POPC tail terminal carbon. Stability of bilayer plane center X,Y,Z positions across a trajectory. **(C)** Stability of X and Y component of normal vector for bilayer plane; Z remains nearly 1 within 1%. **(D)** For each simulation frame and its calculated bilayer plane, the average distance of POPC tail terminal carbon atoms linearly projected onto the calculated bilayer midline (Z=0) for that frame (black) and the standard deviation of those distances (red); these are essentially the residual error values of the bilayer plane calculation. **(E)** Given the calculated bilayer midline for each frame, the distance of protein and lipid atom types are shown plotted across the simulation. Membrane insertion depth for Cα atoms of specific residues: first Tyr of the HA tag (blue), second Gly of the G-X_3-G motif (red), and the 2nd to last TM domain amino acid (black, 2 before Ser); the distance of the center of mass of the TM dimer's two equivalents atoms linearly projected onto the bilayer plane. POPC leaflet positions for each frame calculated the mean distance of glycerol oxygen atoms projected onto the bilayer place, resulting in a membrane bilayer length of the expected ~30 Å. **(F)** Visual representation of the TM dimers and structural parameters extracted from MD simulation data. **G-J)** TM domain structural features predicted at final MD simulation step. **(G)** Membrane depth of the G-X_3-G motif (2nd glycine) relative to the lipid bilayer midline. **(H)** Membrane depth of the 2nd to last TM domain residue relative to the bilayer midline. **(I)** Distance between Cα atoms for equivalent residues in the dimer (helix A to helix B) at the end of the TM domain (2nd to last TM domain residue). **(J)** Mean asymmetry in membrane depth between equivalent Cα atom in the dimer, helix A versus helix B. **(K-L)** Geometric features of TM domains vary across MD trajectories. Predicted TM domain geometries for 81 unique TM domain sequences (A + B part combinations) shown as their trajectory mean (points) and standard deviation (grey bars). The geometries are well-separated by the distance across the interacting TM domain dimer between the C-terminal-most residues. **(K)** Plot of each trajectories' crossing-angle and C-terminal inter-helical distance (End TM Distance), colored as a heat map including the TM helix length of the A + B part. **(L)** Plot of each trajectories' crossing-angle and C-terminal inter-helical distance (End TM Distance), colored as a heat map including the Z-depth of the G-X_3-G motif relative to the bilayer normal midpoint (best fit plane).

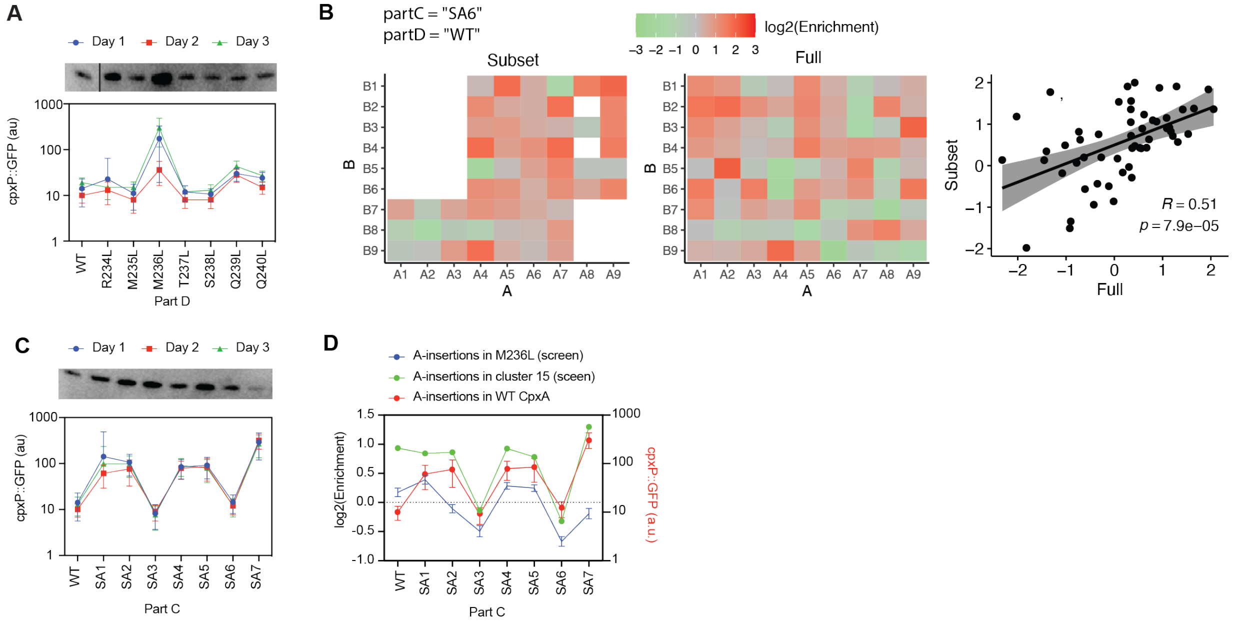
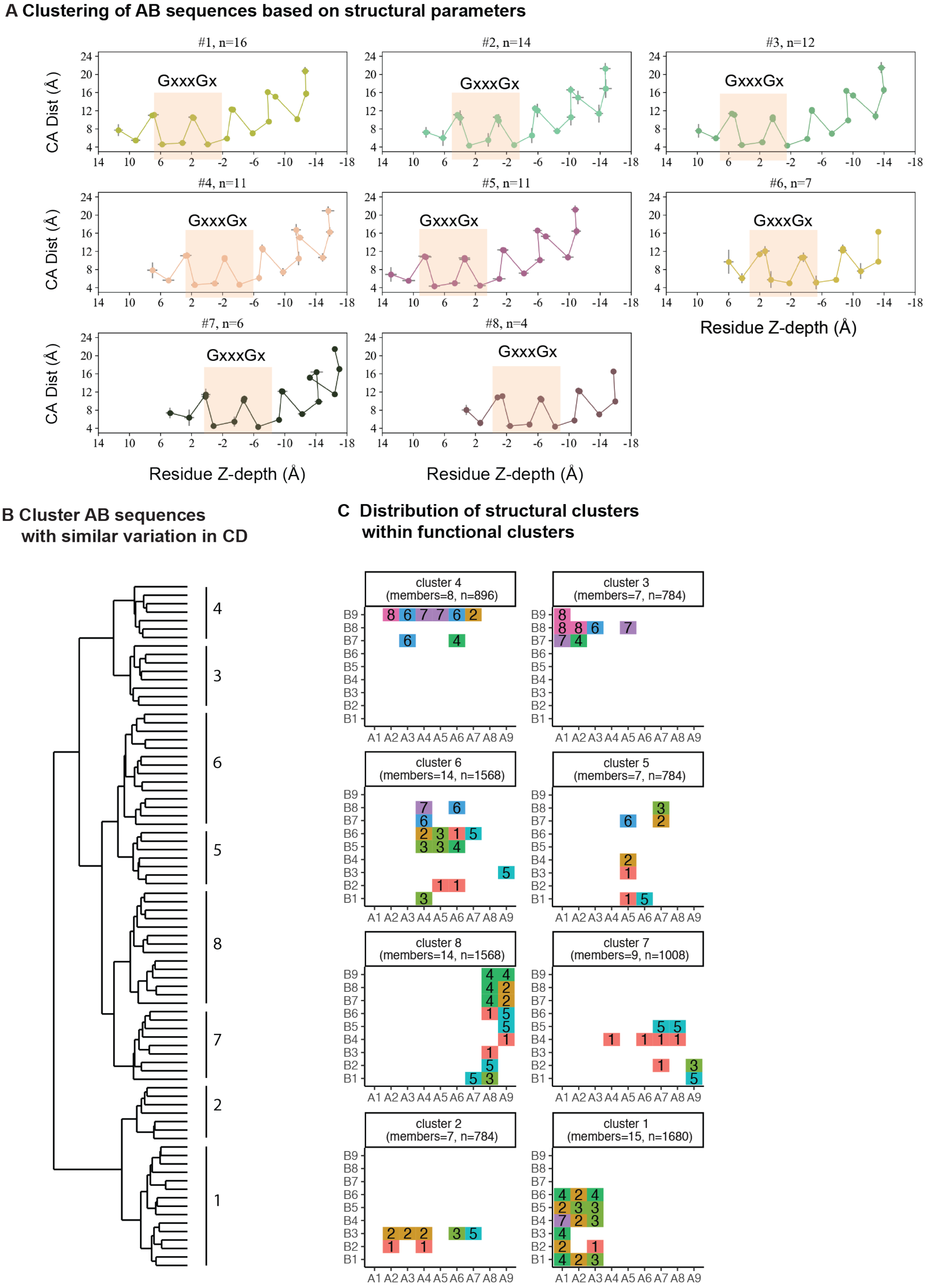


Figure 3.10 - Functional reporter assay on single mutants made in otherwise wild-type CpxA. (A) Fluorescence of the cpxP::GFP as a function of S-helix Leu scan made in wild-type CpxA. **(B)** Heatmaps of enrichment as a function of parts A and B for a single CD combination (part C = SA6, part D = WT). Correlation between subset and full libraries is shown on the right. **(C)** Fluorescence of the cpxP::GFP as a function of variable length alanine linkers . **(A,C)** Reproducibility over 1-3 days of CpxA activation demonstrates consistent patterns of signaling. Western blots show protein expression using the antibodies against the HA tag. Error bars are standard error. **(D)** Comparison of A-insertion linkers between mutants made in a WT background and screen data. Cluster 15 and M236L (averaged A+B) display a similar periodic pattern to a WT TM and S-helix.



A Clustering of AB sequences based on structural parameters

B Cluster AB sequences with similar variation in CD

C Distribution of structural clusters within functional clusters

cluster 4
(members=8, n=896)

	A1	A2	A3	A4	A5	A6	A7	A8	A9
B9		8	6	7	7	6	2		
B8									
B7			6			4			
B6									
B5									
B4									
B3									
B2									
B1									

cluster 3
(members=7, n=784)

	A1	A2	A3	A4	A5	A6	A7	A8	A9
B9	8								
B8	8	8	6		7				
B7	7	4							
B6									
B5									
B4									
B3									
B2									
B1									

cluster 6
(members=14, n=1568)

	A1	A2	A3	A4	A5	A6	A7	A8	A9
B9									
B8				7		6			
B7				6					
B6				2	3	1	5		
B5				3	3	4			
B4									
B3									5
B2					1	1			
B1				3					

cluster 5
(members=7, n=784)

	A1	A2	A3	A4	A5	A6	A7	A8	A9
B9									
B8							3		
B7					6		2		
B6									
B5									
B4					2				
B3					1				
B2									
B1					1	5			

cluster 8
(members=14, n=1568)

	A1	A2	A3	A4	A5	A6	A7	A8	A9
B9								4	4
B8								4	2
B7								4	2
B6								1	5
B5									5
B4									1
B3								1	
B2								5	
B1							5	3	

cluster 7
(members=9, n=1008)

	A1	A2	A3	A4	A5	A6	A7	A8	A9
B9									
B8									
B7									
B6									
B5							5	5	
B4				1		1	1	1	
B3									
B2							1		3
B1									5

cluster 2
(members=7, n=784)

	A1	A2	A3	A4	A5	A6	A7	A8	A9
B9									
B8									
B7									
B6									
B5									
B4									
B3		2	2	2		3	5		
B2		1		1					
B1									

cluster 1
(members=15, n=1680)

	A1	A2	A3	A4	A5	A6	A7	A8	A9
B9									
B8									
B7									
B6	4	2	4						
B5	2	3	3						
B4	7	2	3						
B3	4								
B2	2		1						
B1	4	2	3						

Figure 3.11- TM data clustering based on structural and functional data. (A) Two-feature structural clustering of TM domain geometries at the C-terminal half predicted from MD simulation of all 81 unique TM sequences. The results of structural clustering all MD simulations of the 81 unique TM domain sequence (A + B part combinations) by per residue Cα-Cα inter-helical distance and per residue Z-depth versus the bilayer midline. Cluster averages (points) with per position standard deviation (gray error bars) are shown for each cluster. The 5 TM domain residues corresponding to the shared G-X_3-G motif are highlighted (beige shaded box). Clusters are numbered and number of members listed above each plot. This clustering suggests across the 81 sequences a number of unique TM domain geometries are presented to the juxtamembrane linker and HAMP domain within the engineered CpxA construct. **(B).** Functional clustering of the TM based on enrichment data generated in the high throughput screen. The dendrogram shows clustered data, which was used to define 15 groups with similar variation across parts C and D. **(C)** Structural clusters segregate within distinct functional clusters. For example, functional cluster 7 is composed of predominantly structural clusters 1 and 5.

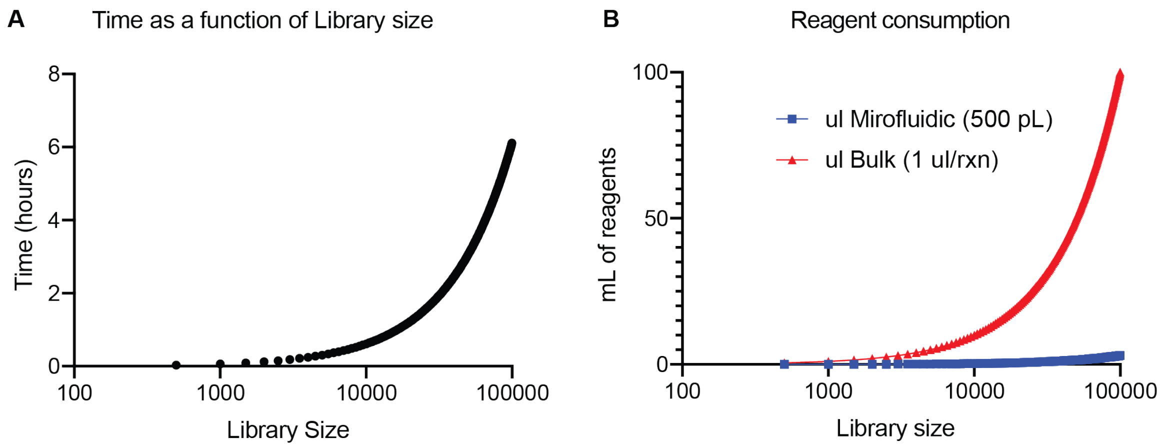


Figure 3.12 - Estimated time and reagent expenditure for microfluidic versus well-based subset library generation. (A) Microfluidic time required for a given library size given a 220 ms duty cycle. **(B)** Reagent consumption of microfluidic (500 pL) vs well-based (1 µL) library generation.

Materials and Methods

Device fabrication: The microfluidic devices are fabricated using standard photolithography (1). Feature heights for the 4 layer fluidic wafer are as follows: 5 µm primer layer (SU8-2005, 3000 rpm, 30 sec), 35 µm valve seats (AZ50 XT positive photoresist, 1750 rpm, 30 sec), 20 µm resistors (SU8-3010, 3000 rpm, 30 sec), 60 µm main fluidic channels (SU8-3050, 4000 rpm, 30 sec)(2). Pneumatic wafers are fabricated at 25 µm height (SU8-3025, 4000 rpm, 30 sec). PDMS replicas are cast by pouring PDMS at 10:1 weight ratio of base:curing agent (Sylgard 184, Dow Chemical, MI, USA) or spin coating a thin membrane on the pneumatic wafer (2000 rpm, 30 sec). After baking at 60 °C for 1 hour, the fluidic replica is cut out, and access ports are punched with a 0.5 µm biopsy punch. The channel side of the fluidic slabs is then stamped on a clean wafer spin-coated with a thin layer of PDMS curing agent (5000 rpm, 1 min) and aligned to the valve features on the pneumatic wafer. The devices are baked overnight, and access ports are punched for the valve channels. The devices are plasma-bonded to a glass slide and baked for 2 days at 60 °C.

Device design: Our design objective is to ensure that the pressure drops ($\mathrm{dP_{inlet}}$) down the inlet channels are ten times greater than the pressure drop down the main channel ($\mathrm{dP_m}$). Since the flow rate through each inlet is proportional to the pressure drop across it, this ensures that all flow rates are within 10%. The pressure drop ratio between an inlet and main channels is:

$$\mathrm{dP_{inlet}}/_{\mathrm{dP_m}} = (\mathrm{R_{inlet}} * \mathrm{Q_{inlet)}}/(\mathrm{R_m} * \mathrm{Q_m}\ )\ (\mathrm{Eqn.\,1})$$

where $\mathrm{R_{inlet}}$ and $\mathrm{R_m}$ are the hydraulic resistances and $\mathrm{Q_{inlet}}$ and $\mathrm{Q_m}$are the flow rates of the inlet and main channels, respectively. $\mathrm{R_{inlet}}$ is set by the serpentine resistor, while $\mathrm{R_m}$ is set by the length of the main channel. To ensure that pure formulations are generated at each collection step, the flow rate of the main channel must be sufficient to fully replace the main channel volume per cycle:

$$\mathrm{Q_m} = \frac{\mathrm{V_m}}{\mathrm{t}} \text{ (Eqn. 2)}$$

where $\mathrm{V_m}$ is volume of the main channel, and t the cycle time. Moreover, since the main channel flow rate is driven by the flows of the open inlet channels, the flow rate of each channel is given by:

$$\mathrm{Q_{inlet}} = \frac{\mathrm{Q_m}}{\mathrm{N_{inlets}}} \text{ (Eqn. 3)}$$

where $\mathrm{N_{inlets}}$ is the number of inlets. Thus, substituting Eqn. 3 into Eqn. 1 yields:

$$\mathrm{dP_{inlet}}/\mathrm{dP_m} = \frac{\mathrm{R_{inlet}}}{\mathrm{R_m N_{inlets}}} \text{ (Eqn. 4)}$$

The hydraulic resistance in the main channel ($\mathrm{R_m}$) is constrained by the practical issue of fitting the inlet channels and their control valves into a single main channel. On one hand, narrow main channels are preferable because they have less volume and therefore can be flushed more quickly with less reagent waste (Eqn. 2); however, they also have higher hydraulic resistance, necessitating that $\mathrm{R_{inlet}}$ be higher to ensure that $\mathrm{dP_{inlet}/dP_m}$remains above 10. In principle, $\mathrm{R_m}$ can be reduced by minimizing the channel length, but this packs the inlet channels and their valves closer together, making fabrication difficult. Thus, in general, $\mathrm{R_m}$ is minimized by packing the inlets as close together as practical. As a starting point, we choose $\mathrm{V_m}$ to be equal to the minimum amount of volume we want to recover per part combination, since a given combination cannot be collected until an entire $\mathrm{V_m}$ of reagent has flushed through the main channel, ensuring that all component solutions have arrived to the droplet maker at the time of collection.

Eqn. 4 provides a simple framework for designing the microfluidic system. $\mathrm{N_{inlets}}$ is fixed by the number of parts we want to assemble. If more parts are to be assembled, more valves must be opened, and thus the inlet resistances $\mathrm{R_{inlet}}$ must be increased to compensate. Thus, in general, the smaller the device can be made, the faster it can be run, but at the cost of requiring higher input pressures. To withstand higher pressures, the devices must be fabricated out of stiffer materials. In addition, since all inlets are maintained at equal input pressure, the housing manifold must be constructed to withstand the increased pressures. Based on these insights, we believe that buttressing of the pressure manifold and tighter packing of valve inlets to reduce the main channel length will allow ~5x increases in speed, beyond which additional challenges emerge that will require a more significant upgrade of the device. Therefore, following the design considerations above, it should be currently possible to generate defined 100k libraries in approximately 6 hours with greatly reduced reagent consumption compared to a well-based approach (**Figure 53B**).

Device operation: We use HFE-7500 fluorinated oil with 2% (w/v) triblock surfactant (RAN Biotechnologies, Inc.) for all experiments at a flow rate of 1000 µL per hour. An aqueous flow rate of 500 µL per hour is produced by combining 4 DNA parts with Enzyme Mix and Buffer Mix. DNA parts and enzymes are supplied to their individual inlet channels through a custom-built 38x pressure manifold, pressurized at 4 psi (**Figure 3.8**). Multilayer membrane valves are pressurized to 40 psi and controlled by solenoid valves (Pneumadyne, S10MM-30-24-3) activated by custom LabView software.

The LabView software cycles through a pre-defined list of ABCD combinations. One duty cycle has four steps. First, the waste valve is opened for 200 ms to flush the main channel. Second, valves corresponding to each of the parts (for example valves controlling A_1, B_2, C_5, D_3) are opened to assemble the desired construct ($A_1B_2C_5D_3$). Third, the waste valve closes. Fourth, the collection valve opens and droplets are collected for 500 ms. Three libraries are generated. The first is generated with a tube reaction of all the possible parts ("Tube"). The second is generated microfluidically using all the parts ("Full"). The third is generated microfluidically with a subset of the parts ("Subset"). To generate the subset library, a list of all ABCD combinations that encoded a final protein of restricted transmembrane length (18 amino acids ≤ [A+B] ≤ 25 amino acids) is created and read by the LabView software. Each ABCD is created in a single duty cycle using the series outlined above.

Library assembly, cloning, and sequencing: The combinatorial libraries are built using Golden Gate assembly (3). All DNA parts contain the BsaI recognition site (GGTCTC) as well as 4bp sticky overhangs that determine the position of each DNA part in the final construct. Sequences of the DNA parts, and primers, can be found in **Table 3.1**. After running the microfluidic instrument, each droplet contains 0.2 µM DNA, 1x NEB Golden Gate Buffer, and 1x NEB Golden Gate Assembly Mix. The reactions are incubated in droplets off-chip in a thermocycler to perform the Golden Gate assembly using the following protocol: (1 min /37°C, 1 min/16°C) x30, 5 min/55°C. After assembly, 20% (v/v) perfluoro-1-octanol (Sigma Aldrich) is added to break the emulsion and recover assembled DNA. Sequencing libraries are prepared from on-chip assembled material using a dual primer PCR with primer set 1 (ChipLibSeq_F and ChipLibSeq_R) and primer set 2 (Nextera XT Index 1 N7XX and Nextera XT Index 2 S5XX) with KAPA HiFi Mastermix. The concentration of the primers is 50 nM and 500nM for set 1 and set 2, respectively. The thermocycling program used is: 3 min/95 °C, (20 sec/98 °C, 20 sec/50 °C, 45 sec/72 °C)x5, (20 sec/98 °C, 20 sec/70 °C, 45 sec/72 °C)x25, 10 min/72 °C.

Assembled DNA is PCR-amplified (Kapa Hifi Mastermix, Kapa Biosystems) using primers ChipLib_F (500nM) and ChipLib_R (500nM) for 3min/95 °C, (20 sec/98 °C, 20 sec/50 °C, 45 sec/72 °C)x14 and cloned using Golden Gate Assembly into plasmid pNS001. This plasmid has a chloramphenicol resistance cassette, p15A ORI, CpxR and CpxA fusion proteins under a pTrc promoter and the CpxAR two component system reporter (pCpxP-GFP). The plasmid library is transformed into chemically competent cells, plated onto 4 large petri dishes (Thermo Scientific Nunc, # 240835) per library and incubated until single colonies are visible. Plates are scraped and the plasmid library is purified (QIAGEN Plasmid Midi Kit). PCR is used to prepare plasmid libraries for sequencing using a staggered primer on the P5 side (CpxA_P5_s1-s8) to avoid clustering problems and indexed primers on the P7 side for sample indexing (CpxA_P7_B01-B04). The plasmid libraries are sequenced using a MiSeq PE150. Sequences of oligonucleotides used in this work are found in **Tables 3.1-3.2**.

Expression and screening of the CpxA library: Three plasmid libraries (Tube, Full, Subset) are transformed into chemically competent JW3882-1 (cpxA771(del)::kan) obtained from the Coli Genetic Stock Center Molecular (Cellular & Developmental Biology, Yale University). Libraries are plated to onto 4 large petri dishes with LB (with chloramphenicol and kanamycin) per library and scraped after 18 hours of incubation at 37°C. Growth of ΔmalE strains expressing the construct on M9+maltose plates validates proper membrane insertion (4, 5). Cells are washed with LB media and freezer stock aliquots are made for subsequent use. To perform screening experiments, single freezer stocks are recovered by shaking in LB at 37C for 4 hours, followed by a single wash in M9 minimal media. Cells are diluted in M9 media to OD_{600} and 1 million cells are sorted by Fluorescent activated cell sorting (MoFlo Astrios) using a 100µm nozzle. Plasmids from sorted and unsorted cells are purified (QIAGEN Plasmid Mini Kit). Libraries are prepared as above and sequenced using a MiSeq PE150.

Sequence analysis: Paired end raw sequencing reads are assembled using bbmerge to generate an intact sequence containing the ABCD fragment and trimmed to remove primer contaminants. The usearch program (-search_global -top_hit_only -id .99 -query_cov 1 -target_cov 1 -maxdiffs 1 -strand plus) is used to assign each assembled read to a variant ($A_iB_jC_kD_l$). A custom python script counts reads for each part. Normalized variant abundance is calculated as the sum of the reads that map uniquely to a variant, plus one, divided by the average of the number of variants detected in the sample. Enrichment is calculated as the fold change in the normalized abundance of each variant between post and pre-sort (normalized abundance post-sort / normalized abundance pre-sort). Data from the screen is found in **Dataset 2**.

Generation of linker alanine insertions and S-helix leucine substitutions in full length CpxA: Full length, C-terminally His-tagged CpxA is cloned into the NcoI/SalI sites of the modified pTrc99a plasmid (pSau) by Gibson Assembly. Alanine insertions in Part C and leucine substitutions in Part D are introduced using blunt-ligation cloning strategy. These constructs are transformed into the AFS51 reporter strain (ΔcpxAΔpta::Kan cpxP::GFP). Overnight cultures of AFS51 containing either empty plasmid or various CpxA constructs are grown from single colonies in LB with 50 ug/mL Kanamycin and subsequently diluted 200-fold into fresh LB media with 50 ug/mL Ampicillin. This culture is grown to mid-log phase (OD600 = 0.4 - 0.6), and 50 µL is removed for analysis by FACS. Five microliters of culture is also analyzed by western blot to confirm expression of CpxA. Experiments are repeated in triplicate. The responsiveness of the CpxP::GFP reporter is confirmed in WT CpxA using Brilacidin, an antimicrobial peptide mimetic and inducer of the CpxAR two-component system (6). pSau plasmid with no CpxA insert is used as a negative control. CpxA function is evaluated by FACS on a BD FACS caliber instrument. 20,000 cells, gated by forward and side-scatter, are evaluated for GFP fluorescence (Ex 488 nm, EM 515 nm) per sample. Sample average fluorescence and standard error are determined by standard analysis in Flo-Jo. Protein expression of CpxA variants is determined by Western blotting. Five microliters of each sample is boiled in 1X LDS buffer at 95°C and separated by MES-SDS electrophoresis, followed by dry transfer onto a nitrocellulose membrane (iBlot). Membrane is blocked with 1% BSA in TBST buffer (20 mM Tris, 50 mM NaCl, 1 mM EDTA, 0.1% Tween-20), washed 5X with TBST buffer and probed with anti-pentaHis HRP antibody. HRP signal is measured using luminescent substrate and imaged (Bio-Rad ChemiDoc MP imaging system).

Clustering of screen data: Hierarchical clustering is performed on screen enrichment data calculated above, by combining both screens. Two clustering analyses are conducted. **Figure 3.6** clusters ABD with similar variation in enrichment with respect to C. Data is transformed into a matrix where rows are ABD and columns are C. **Figure 3.11B** clusters AB with similar variation in enrichment with respect to CD. Data is transformed into a matrix where rows are AB and columns are CD. For each clustering analysis, the distance matrix is calculated using the get_dist(method = "spearman") function in R. Next, the hclust(method="ward.D") is used to identify clusters. Data manipulation and plotting is performed using the tidyverse package and ggplot in R.

Molecular Model Building: For each of the 81 engineered TM domain sequences we build and simulate an atomic level molecular model. The sequence of interest includes the N-terminal HA tag, the TM domain defined by parts A and B, and a final C-terminal serine residue. First, template molecular model is built starting with the X-ray crystal structure of the G-X_3-G -containing TM helix dimer GlycophorinA (GpA) in lipid (PDB: 5EH4) (7). Given the variable length of each A + B combination, the TM domain termini are extended by adding alanine residues in ideal α-helical conformation to match the length of the engineered construct. Then, the engineered TM sequence is threaded onto this model with helix register defined by the A part's G-X_3-G motif. An all-atom model with sidechains is built and placed in the implicit lipid bilayer solvation model given GpA's. Next, the model is relaxed in RosettaMP with energy function weights "ref2015_memb" (8). The structure is first minimized with harmonic restraints on the main chain atoms, then put through the membrane-conscious FastRelax protocol including 8 rounds of iterative sidechain repacking, Cartesian minimization (with main chain constraints), and rigid-body transformation that minimizes the TM helix depth and tilt in the implicit lipid bilayer given its apolar solvation energy. The lowest energy model produced from Rosetta in 10 relaxation trials is used to build an all-atom system of the protein in a hydrated lipid bilayer for MD simulation for each of the 81 engineered TM domain sequences.

All-atom Molecular Dynamics Simulations: Given the final position of the TM dimer model based on GpA within the Rosetta implicit membrane bilayer plane (30 Å long), the protein is embedded in a square 1-palmitoyl-2-oleoyl-sn-glycero-3-phosphocholine (POPC) lipid bilayer box of 6.0 x 6.0 nm. First, the POPC lipid box is generated using VMD and overlaid with the implicit bilayer (9). Lipid tails overlapping with the protein (1 Å) are removed. The system is treated as a periodic box, 7.0 nm in the Z direction, and hydrated with TIP3P water (ca. 20 Å of water regions above and below the bilayer). NaCl is added to the system to neutralize protein charge and to yield a final 0.15 M ion concentration. The system is prepared and treated with CHARMM36 parameters (10) using the GROMACS 2018 engine for minimization and simulations (11), using the recommended cut-offs (rcoulomb, rvdw = 1.2 nm) switching (1.0 nm), and Particle-Mesh Ewald distances for CHARMM36 with a 2 fs time step for Langevin dynamics.

The system is minimized using 5000 steps using steepest decent with harmonic positional restraints of 1 kcal mol^{-1} Å^{-2} on all non-hydrogen protein atoms. Next, a 50 ps NVT dynamics simulation is conducted at 310.15 K with identical restraints using the V-rescale algorithm to maintain this temperature with a 0.1 ps coupling constant. A 12 ns NPT simulation is next run with 1 kcal mol^{-1} Å^{-2} harmonic restraints on protein Cα atoms, with a berendsen thermostat at 310.15 K with a 1.0 ps coupling constant and a semiisotropic berendsen barostat at 1 bar with a pressure coupling time constant of 1 ps. In all cases, the periodic box dimensions become stable within 8-12 ns. Unrestrained Langevin dynamics production simulations are then run for 80 ns with a Nose-Hoover thermostat at 310.15 K with a 1.0 ps coupling constant and a semiisotropic Parrinello-Rahman barostat at 1 bar with a pressure coupling time constant of 1 ps.

Analysis of Structural Features from MD Trajectories: To calculate the insertion depth of each residue in the 2-helix TM domain, the bilayer center plane (z=0) is calculated using the terminal carbon atoms of the POPC acyl tails, calculated separately for each frame in the simulation. From these atoms' coordinates, the bilayer plane's center is calculated by the coordinates' mean. The bilayer's normal vector is calculated by singular value decomposition, producing a plane that minimizes the mean distance of these atoms to the plane. The bilayer behavior and atom-specific depths calculated relative to this bilayer are plotted across the simulation trajectory for a representative example in **Figure 3.9B-E**. Given that two equivalent atoms exist for each dimer, a residue's depth is calculated as the center of mass of these two atoms linearly projected onto the bilayer plane. Each residue's Z-depth relative to the bilayer midline is calculated, producing a vector later used for 2-feature structural clustering (see below). The Z-depth of the second glycine of each TM domains' G-X_3-G motif and the Z-depth of the 2nd to last TM domain residue are shown for a single representative simulation (**Figure 3.9E**) as well as for all 81 simulations of each unique TM domain sequences in **Figure 3.9G-H**.

The Asymmetry feature aims to capture the TM dimer structural re-arrangements such as tilt (**Figure 3.9F**). For each pair of equivalent residues across the TM domain dimer, the Z-depth asymmetry is calculated as the difference in the projected distance to the bilayer midline for each of these two equivalent residues' Cα atoms. The Asymmetry feature is the mean of Z-depth asymmetry across all residues in the TM segment beginning at the Glycophorin G-X_3-G motif to the end of the TM domain (LILLGVMAGVIGT...) (**Figure 3.9J**). The inter-helical dimer TM domain crossing angle is calculated first by calculating a helical axis vector for each TM helix within the 13-residue Glycophorin G-X_3-G motif region (LILLGVMAGVIGT), then taking the dihedral angle. **Figure 3.9K-L** shows inter-helical crossing angle for all 81 sequences averaged across the trajectories, with dynamic variation in this feature within each simulation trajectory depicted as error bars, representing the crossing angle standard deviation across simulation frames. The inter-helical distance of each equivalent pair of residues' Cα atoms across the dimer interface is calculated, resulting in a vector used for 2-feature structural clustering (see below). We focus on Cα-Cα interhelical distance at the end of the TM domain (2nd to last TM domain residue, 2 prior to final Ser), shown across the 81 sequences by A and B parts in **Figure 3.9I**.

Structural clustering of predicted TM domain geometries from MD simulations: We performed a two-feature weighted structural clustering of the TM helix geometries across the 81 TM domain dimer simulations (each a unique A + B part combination). The structural clustering used the structural features of each residue's Cα-Cα inter-helical distance and Z-depth in membrane relative to the bilayer normal. For each simulation's final frame, we calculated these features for every residue in the TM domain dimer region beginning at the Glycophorin G-X_3-G motif (LILLGVMAGVIGT) and extending to the end of the TM domain (variable length depending on the B part). This results in each sequence being described by two vectors: the per residue Cα-Cα inter-helical distances and the per residue membrane Z-depth (described above). An all-by-all distance matrix was calculated for each pairwise distance between simulations by taking the root mean squared (RMS) difference of each simulations' two vectors separately (residue Cα-Cα inter-helical distances, residue Z-depth), followed by taking the weighted sum of these two vector differences with the Z-depth feature given a weight of 3 and the Cα-Cα distance feature given a weight of 1. With this square pairwise distance matrix, a hierarchical clustering was performed using the Ward criteria (minimizing cluster variance). Clusters were defined by a linkage cut-off distance equal to the mean of the pairwise distance matrix (non-redundant, upper triangular portion) plus 1 standard deviation. This analysis resulted in 8 clusters, summarized in **Figure 3.11A** by the mean values of each feature vector (Cα-Cα distances, residue Z-depth) within each cluster. The sequences within each cluster are listed in **Dataset 3**.

Table 3.1 – Primers used in this study

Name	Sequence	Notes
ChipLib_F	ATATACGTCTCTTGCG	For amplification and cloning of parts assembled on chip
ChipLib_R	AGATACGTCTCTGCAG	
ChipLibSeq_F	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGNNNNNATATACGTCTCTTGCG	For sequencing of parts assembled on chip
ChipLibSeq_R	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGNNNNNAGATACGTCTCTGCAG	
Nextera XT Index 1 Primers (N7XX)	from the Nextera XT Index kit (FC-131-1001 or FC-131-1002)	
Nextera XT Index 2 Primers (S5XX)	from the Nextera XT Index kit (FC-131-1001 or FC-131- 1002)	
CpxA_P5_s1	AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCTCGTATGAT**GTGCCGGATTATGCG**	For sequencing plasmid libraries after sorting GFP high expressing cells Combined P5 primer (CpxA_P5_s1-8) introduces stagger for amplicon sequencing. These area added at equamolar concentrations to improve cluster generation. See: "Protocol: PCR of sgRNAs for Illumina sequencing" Broad MIT
CpxA_P5_s2	AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCTGCGTATGAT**GTGCCGGATTATGCG**	
CpxA_P5_s3	AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCTAGCGTATGAT**GTGCCGGATTATGCG**	
CpxA_P5_s4	AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCTCAACGTATGAT**GTGCCGGATTATGCG**	
CpxA_P5_s6	AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCTTGCACCGTATGAT**GTGCCGGATTATGCG**	
CpxA_P5_s7	AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCTACGCAACGTATGAT**GTGCCGGATTATGCG**	
CpxA_P5_s8	AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCTGAAGACCCGTATGAT**GTGCCGGATTATGCG**	
CpxA_P7_B01	CAAGCAGAAGACGGCATACGAGATATTGGATTGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT**GCTCGTGAGAGATATCAGAAAGCAG**	One CpxA_P7_B01 primer used with pooled P5 primer
CpxA_P7_B02	CAAGCAGAAGACGGCATACGAGATATACTCGGGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT**GCTCGTGAGAGATATCAGAAAGCAG**	
CpxA_P7_B03	CAAGCAGAAGACGGCATACGAGATTATGAGAAGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT**GCTCGTGAGAGATATCAGAAAGCAG**	
CpxA_P7_B04	CAAGCAGAAGACGGCATACGAGATGCACAGTTGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT**GCTCGTGAGAGATATCAGAAAGCAG**	

Table 3.2 - Oligonuclotide used for CpxA library Assembly

Part Name	Sequence
A1	**ATATACGTCTCTTGCG***CTTCTCCTGCTGTTGTTACTGCTCTTAATCCTGTTAGGCGTG*AT**GGCATGAGACCTGTGT**
A2	ATATACGTCTCTTGCGCTCCTGCTGTTGTTACTGCTCTTAATCCTGTTAGGCGTGATGGCATGAGACCTGTGT
A3	ATATACGTCTCTTGCGCTGCTGTTGTTACTGCTCTTAATCCTGTTAGGCGTGATGGCATGAGACCTGTGT
A4	ATATACGTCTCTTGCGCTGTTGTTACTGCTCTTAATCCTGTTAGGCGTGATGGCATGAGACCTGTGT
A5	ATATACGTCTCTTGCGTTGTTACTGCTCTTAATCCTGTTAGGCGTGATGGCATGAGACCTGTGT
A6	ATATACGTCTCTTGCGTTACTGCTCTTAATCCTGTTAGGCGTGATGGCATGAGACCTGTGT
A7	ATATACGTCTCTTGCGCTGCTCTTAATCCTGTTAGGCGTGATGGCATGAGACCTGTGT
A8	ATATACGTCTCTTGCGCTCTTAATCCTGTTAGGCGTGATGGCATGAGACCTGTGT
A9	ATATACGTCTCTTGCGTTAATCCTGTTAGGCGTGATGGCATGAGACCTGTGT
B1	**GTGTGGGTCTCT**GGCAGGTGTTATTGGCACCCTGTTACTGTTGCTGCTTCTG**AGTTGAGACCACACA**
B2	GTGTGGGTCTCTGGCAGGTGTTATTGGCACCCTGTTACTGTTGCTGCTTTGGAGTTGAGACCACACA
B3	GTGTGGGTCTCTGGCAGGTGTTATTGGCACCCTGTTACTGTTGCTGTGGCTGAGTTGAGACCACACA
B4	GTGTGGGTCTCTGGCAGGTGTTATTGGCACCCTGTTACTGTTGTGGCTTCTGAGTTGAGACCACACA
B5	GTGTGGGTCTCTGGCAGGTGTTATTGGCACCCTGTTACTGTGGCTGCTTCTGAGTTGAGACCACACA
B6	GTGTGGGTCTCTGGCAGGTGTTATTGGCACCCTGTTATGGTTGCTGCTTCTGAGTTGAGACCACACA
B7	GTGTGGGTCTCTGGCAGGTGTTATTGGCACCCTGTTACTGAGTTGAGACCACACA
B8	GTGTGGGTCTCTGGCAGGTGTTATTGGCACCCTGTTATGGAGTTGAGACCACACA
B9	GTGTGGGTCTCTGGCAGGTGTTATTGGCACCCTGTGGCTGAGTTGAGACCACACA
C1	**ATATAGGTCTCA**GAGTGCGGCAGCTGCGGCAGCTGCCCTGGCAAAACCGGCGCGTAAGCTGAAAAACGCTGCCGATGAAGTTGCCCAGGGAAACTTACGCCAGCACCCGGAACTGGAAGCGGGGCCACAGGAATTCCTTGCCGCAGGTGCCAGTTTTAACCAGATGGTTACCGC**GCTGCGAGACCATATA**
C2	ATATAGGTCTCAGAGTGCGGCAGCTGCGGCAGCTCTGGCAAAACCGGCGCGTAAGCTGAAAAACGCTGCCGATGAAGTTGCCCAGGGAAACTTACGCCAGCACCCGGAACTGGAAGCGGGGCCACAGGAATTCCTTGCCGCAGGTGCCAGTTTTAACCAGATGGTTACCGCGCTGCGAGACCATATA
C3	ATATAGGTCTCAGAGTGCGGCAGCTGCGGCACTGGCAAAACCGGCGCGTAAGCTGAAAAACGCTGCCGATGAAGTTGCCCAGGGAAACTTACGCCAGCACCCGGAACTGGAAGCGGGGCCACAGGAATTCCTTGCCGCAGGTGCCAGTTTTAACCAGATGGTTACCGCGCTGCGAGACCATATA
C4	ATATAGGTCTCAGAGTGCGGCAGCTGCGCTGGCAAAACCGGCGCGTAAGCTGAAAAACGCTGCCGATGAAGTTGCCCAGGGAAACTTACGCCAGCACCCGGAACTGGAAGCGGGGCCACAGGAATTCCTTGCCGCAGGTGCCAGTTTTAACCAGATGGTTACCGCGCTGCGAGACCATATA
C5	ATATAGGTCTCAGAGTGCGGCAGCTCTGGCAAAACCGGCGCGTAAGCTGAAAAACGCTGCCGATGAAGTTGCCCAGGGAAACTTACGCCAGCACCCGGAACTGGAAGCGGGGCCACAGGAATTCCTTGCCGCAGGTGCCAGTTTTAACCAGATGGTTACCGCGCTGCGAGACCATATA
C6	ATATAGGTCTCAGAGTGCGGCACTGGCAAAACCGGCGCGTAAGCTGAAAAACGCTGCCGATGAAGTTGCCCAGGGAAACTTACGCCAGCACCCGGAACTGGAAGCGGGGCCACAGGAATTCCTTGCCGCAGGTGCCAGTTTTAACCAGATGGTTACCGCGCTGCGAGACCATATA
C7	ATATAGGTCTCAGAGTGCGCTGGCAAAACCGGCGCGTAAGCTGAAAAACGCTGCCGATGAAGTTGCCCAGGGAAACTTACGCCAGCACCCGGAACTGGAAGCGGGGCCACAGGAATTCCTTGCCGCAGGTGCCAGTTTTAACCAGATGGTTACCGCGCTGCGAGACCATATA
C8	ATATAGGTCTCAGAGTCTGGCAAAACCGGCGCGTAAGCTGAAAAACGCTGCCGATGAAGTTGCCCAGGGAAACTTACGCCAGCACCCGGAACTGGAAGCGGGGCCACAGGAATTCCTTGCCGCAGGTGCCAGTTTTAACCAGATGGTTACCGCGCTGCGAGACCATATA
D1	**GCGCGGGTCTCC**GCTGGAACGCATGATGACCTCTCAACAGCGT**CTGCAGAGACGTATCT**
D2	GCGCGGGTCTCCGCTGGAACTGATGATGACCTCTCAACAGCGTCTGCAGAGACGTATCT
D3	GCGCGGGTCTCCGCTGGAACGCCTGATGACCTCTCAACAGCGTCTGCAGAGACGTATCT
D4	GCGCGGGTCTCCGCTGGAACGCATGCTGACCTCTCAACAGCGTCTGCAGAGACGTATCT
D5	GCGCGGGTCTCCGCTGGAACGCATGATGCTGTCTCAACAGCGTCTGCAGAGACGTATCT
D6	GCGCGGGTCTCCGCTGGAACGCATGATGACCCTGCAACAGCGTCTGCAGAGACGTATCT
D7	GCGCGGGTCTCCGCTGGAACGCATGATGACCTCTCTGCAGCGTCTGCAGAGACGTATCT
D8	GCGCGGGTCTCCGCTGGAACGCATGATGACCTCTCAACTGCGTCTGCAGAGACGTATCT

References

1. H. Alper, C. Fischer, E. Nevoigt, G. Stephanopoulos, Tuning genetic control through promoter engineering. *Proc Natl Acad Sci U S A* **102**, 12678-12683 (2005).
2. M. Jeschek, D. Gerngross, S. Panke, Combinatorial pathway optimization for streamlined metabolic engineering. *Curr Opin Biotechnol* **47**, 142-151 (2017).
3. P. Xu *et al.*, Modular optimization of multi-gene pathways for fatty acids production in E. coli. *Nat Commun* **4**, 1409 (2013).
4. L. M. Starita, S. Fields, Deep Mutational Scanning: A Highly Parallel Method to Measure the Effects of Mutation on Protein Function. *Cold Spring Harbor Protocols* **2015**, pdb.top077503-077505 (2015).
5. P. A. Romero, T. M. Tran, A. R. Abate, Dissecting enzyme function with microfluidic-based deep mutational scanning. *Proceedings of the National Academy of Sciences of the United States of America* **112**, 7159-7164 (2015).
6. D. M. Fowler, S. Fields, Deep mutational scanning: a new style of protein science. *Nature Methods* **11**, 801-807 (2014).
7. R. E. Cobb, J. C. Ning, H. Zhao, DNA assembly techniques for next-generation combinatorial biosynthesis of natural products. *J Ind Microbiol Biotechnol* **41**, 469-477 (2014).
8. C. Neylon, Chemical and biochemical strategies for the randomization of protein encoding DNA sequences: library construction methods for directed evolution. *Nucleic Acids Res* **32**, 1448-1459 (2004).

9. C. H. Reynolds *et al.*, Diversity and Coverage of Structural Sublibraries Selected Using the SAGE and SCA Algorithms. *Journal of Chemical Information and Computer Sciences* **41**, 1470-1477 (2001).

10. S. E. Osborne, A. D. Ellington, Nucleic Acid Selection and the Challenge of Combinatorial Chemistry. *Chemical Reviews* **97**, 349-370 (1997).

11. C. Engler, R. Kandzia, S. Marillonnet, A One Pot, One Step, Precision Cloning Method with High Throughput Capability. *PloS one* **3**, e3647 (2008).

12. D. G. Gibson *et al.*, Enzymatic assembly of DNA molecules up to several hundred kilobases. *Nat Methods* **6**, 343-345 (2009).

13. Y. Zhang, U. Werling, W. Edelmann, SLiCE: a novel bacterial cell extract-based DNA cloning method. *Nucleic Acids Research* **40**, e55-e55 (2012).

14. M. Z. Li, S. J. Elledge, Harnessing homologous recombination in vitro to generate recombinant DNA via SLIC. *Nat Methods* **4**, 251-256 (2007).

15. R. P. Shetty, D. Endy, T. F. Knight, Jr., Engineering BioBrick vectors from BioBrick parts. *J Biol Eng* **2**, 5 (2008).

16. M. Jeschek, D. Gerngross, S. Panke, Rationally reduced libraries for combinatorial pathway optimization minimizing experimental effort. *Nat Commun* **7**, 11163 (2016).

17. D. S. Kong, P. A. Carr, L. Chen, S. Zhang, J. M. Jacobson, Parallel gene synthesis in a microfluidic device. *Nucleic Acids Research* **35**, e61-e61 (2007).

18. S. C. C. Shih *et al.*, A Versatile Microfluidic Device for Automating Synthetic Biology. *ACS Synthetic Biology* 10.1021/acssynbio.5b00062 (2015).

19. U. Tangen *et al.*, DNA-library assembly programmed by on-demand nano-liter droplets from a custom microfluidic chip. *Biomicrofluidics* **9**, 044103 (2015).

20. C. J. Ochs, A. R. Abate, Rapid modulation of droplet composition with pincer microvalves. *Lab Chip* **15**, 52-56 (2015).

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

22. R. F. Weber, P. M. Silverman, The cpx proteins of Escherichia coli K12. Structure of the cpxA polypeptide as an inner membrane component. *J Mol Biol* **203**, 467-478 (1988).

23. E. Batchelor, D. Walthers, L. J. Kenney, M. Goulian, The Escherichia coli CpxA-CpxR envelope stress response system regulates expression of the porins ompF and ompC. *J Bacteriol* **187**, 5723-5731 (2005).

24. A. E. Mechaly, N. Sassoon, J. M. Betton, P. M. Alzari, Segmental helical motions and dynamical asymmetry modulate histidine kinase autophosphorylation. *PLoS Biol* **12**, e1001776 (2014).

25. T. Krell *et al.*, Bacterial sensor kinases: diversity in the recognition of environmental signals. *Annu Rev Microbiol* **64**, 539-559 (2010).

26. M. A. Unger, H. P. Chou, T. Thorsen, A. Scherer, S. R. Quake, Monolithic microfabricated valves and pumps by multilayer soft lithography. *Science* **288**, 113-116 (2000).

27. J. J. Agresti *et al.*, Ultrahigh-throughput screening in drop-based microfluidics for directed evolution. *Proceedings of the National Academy of Sciences of the United States of America* **107**, 4004-4009 (2010).

28. I. Gushchin *et al.*, Mechanism of transmembrane signaling by sensor histidine kinases. *Science* **356** (2017).

29. T. L. Raivio, T. J. Silhavy, Transduction of envelope stress in Escherichia coli by the Cpx two-component system. *Journal of Bacteriology* **179**, 7724-7733 (1997).

30. A. E. Mechaly, N. Sassoon, J.-M. Betton, P. M. Alzari, Segmental helical motions and dynamical asymmetry modulate histidine kinase autophosphorylation. *PLoS Biology* **12**, e1001776 (2014).

31. Manasi P. Bhate, Kathleen S. Molnar, M. Goulian, William F. DeGrado, Signal Transduction in Histidine Kinases: Insights from New Structures. *Structure* **23**, 981-994 (2015).

32. M. Hulko *et al.*, The HAMP domain structure implies helix rotation in transmembrane signaling. *Cell* **126**, 929-940 (2006).

33. H. U. Ferris *et al.*, The mechanisms of HAMP-mediated signaling in transmembrane receptors. *Structure* **19**, 378-385 (2011).

34. M. V. Airola, K. J. Watts, A. M. Bilwes, B. R. Crane, Structure of concatenated HAMP domains provides a mechanism for signal transduction. *Structure* **18**, 436-448 (2010).

35. K. E. Swain, J. J. Falke, Structure of the conserved HAMP domain in an intact, membrane-bound chemoreceptor: a disulfide mapping study. *Biochemistry* **46**, 13684-13695 (2007).

36. J. S. Parkinson, Signaling mechanisms of HAMP domains in chemoreceptors and sensor kinases. *Annual Review of Microbiology* **64**, 101-122 (2010).

37. P.-C. Su, B. W. Berger, A novel assay for assessing juxtamembrane and transmembrane domain interactions important for receptor heterodimerization. *Journal of Molecular Biology* **425**, 4652-4658 (2013).

38. P.-C. Su, B. W. Berger, Identifying key juxtamembrane interactions in cell membranes using AraC-based transcriptional reporter assay (AraTM). *Journal of Biological Chemistry* **287**, 31515-31526 (2012).

39. K. R. MacKenzie, J. H. Prestegard, D. M. Engelman, A transmembrane helix dimer: structure and implications. *Science* **276**, 131-133 (1997).

40. D. T. Moore, B. W. Berger, W. F. DeGrado, Protein-protein interactions in the membrane: sequence, structural, and biological motifs. *Structure* **16**, 991-1001 (2008).

41. R. R. Draheim, A. F. Bormans, R. Z. Lai, M. D. Manson, Tryptophan residues flanking the second transmembrane helix (TM2) set the signaling state of the Tar chemoreceptor. *Biochemistry* **44**, 1268-1277 (2005).

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

43. P. A. DiGiuseppe, T. J. Silhavy, Signal detection and target gene induction by the CpxRA two-component system. *J Bacteriol* **185**, 2432-2440 (2003).

44. V. Stewart, The HAMP signal-conversion domain: static two-state or dynamic three-state? *Mol Microbiol* **91**, 853-857 (2014).

45. L. G. Mondejar, A. Lupas, A. Schultz, J. E. Schultz, HAMP domain-mediated signal transduction probed with a mycobacterial adenylyl cyclase as a reporter. *J Biol Chem* **287**, 1022-1031 (2012).

46. M. Doebber *et al.*, Salt-driven equilibrium between two conformations in the HAMP domain from Natronomonas pharaonis: the language of signal transfer? *J Biol Chem* **283**, 28691-28701 (2008).

47. A. Moglich, R. A. Ayers, K. Moffat, Design and signaling mechanism of light-regulated histidine kinases. *J Mol Biol* **385**, 1433-1444 (2009).

48. V. Stewart, L. L. Chen, The S helix mediates signal transmission as a HAMP domain coiled-coil extension in the NarX nitrate sensor from Escherichia coli K-12. *J Bacteriol* **192**, 734-745 (2010).

49. K. Winkler, A. Schultz, J. E. Schultz, The S-helix determines the signal in a Tsr receptor/adenylyl cyclase reporter. *J Biol Chem* **287**, 15479-15488 (2012).

50. B. Wang, A. Zhao, R. P. Novick, T. W. Muir, Activation and inhibition of the receptor histidine kinase AgrC occurs through opposite helical transduction motions. *Mol Cell* **53**, 929-940 (2014).

51. N. W. Schmidt, G. Grigoryan, W. F. DeGrado, The accommodation index measures the perturbation associated with insertions and deletions in coiled-coils: Application to understand signaling in histidine kinases. *Protein Sci* **26**, 414-435 (2017).

52. J. Huang, A. D. MacKerell, Jr., CHARMM36 all-atom additive protein force field: validation based on comparison to NMR data. *J Comput Chem* **34**, 2135-2145 (2013).

53. M. V. Airola *et al.*, HAMP domain conformers that propagate opposite signals in bacterial chemoreceptors. *PLoS Biol* **11**, e1001479 (2013).

54. H. U. Ferris *et al.*, Mechanism of regulation of receptor histidine kinases. *Structure* **20**, 56-66 (2012).

55. H. U. Ferris, K. Zeth, M. Hulko, S. Dunin-Horkawicz, A. N. Lupas, Axial helix rotation as a mechanism for signal regulation inferred from the crystallographic analysis of the E. coli serine chemoreceptor. *J Struct Biol* **186**, 349-356 (2014).

56. K. E. Swain, J. J. Falke, Structure of the conserved HAMP domain in an intact, membrane-bound chemoreceptor: a disulfide mapping study. *Biochemistry* **46**, 13684-13695 (2007).

57. L. Zhu, P. G. Bolhuis, J. Vreede, The HAMP signal relay domain adopts multiple conformational states through collective piston and tilt motions. *PLoS Comput Biol* **9**, e1002913 (2013).

58. R. Z. Lai, J. S. Parkinson, Functional suppression of HAMP domain signaling defects in the E. coli serine chemoreceptor. *J Mol Biol* **426**, 3642-3655 (2014).

59. J. S. Parkinson, Signaling mechanisms of HAMP domains in chemoreceptors and sensor kinases. *Annu Rev Microbiol* **64**, 101-122 (2010).

Publishing Agreement

It is the policy of the University to encourage open access and broad distribution of all theses, dissertations, and manuscripts. The Graduate Division will facilitate the distribution of UCSF theses, dissertations, and manuscripts to the UCSF Library for open access and distribution. UCSF will make such theses, dissertations, and manuscripts accessible to the public and will take reasonable steps to preserve these works in perpetuity.

I hereby grant the non-exclusive, perpetual right to The Regents of the University of California to reproduce, publicly display, distribute, preserve, and publish copies of my thesis, dissertation, or manuscript in any form or media, now existing or later derived, including access online for teaching, research, and public service purposes.

DocuSigned by:
Bruk Mensa
B34ED14ECA97452... Author Signature

3/18/2021
Date