THE UNIVERSITY OF CHICAGO

STRUCTURAL PRINCIPLES OF PROTEIN EVOLVABILITY

A DISSERTATION SUBMITTED TO

THE FACULTY OF THE DIVISION OF THE BIOLOGICAL SCIENCES

AND THE PRITZKER SCHOOL OF MEDICINE

IN CANDIDACY FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY

GRADUATE PROGRAM IN BIOCHEMISTRY AND MOLECULAR BIOPHYSICS

BY

CRAIG DEVALK

CHICAGO, ILLINOIS

DECEMBER 2024

Dedication

This work is dedicated to my family. To Mom, Dad, Amber, and Tyler, thank you for your steadfast support, love, and encouragement. To Michelle, my partner, I am profoundly grateful for your unwavering support and enduring belief in me.

Table of Contents

List of Figures

List of Tables

Acknowledgement

Countless people have contributed, both directly and indirectly, to this thesis. I will attempt to list them all here, but to those I may not mention, I offer a heartfelt thank you for making my time in graduate school smoother and easier.

First, I'd like to thank my advisor, Rama. His infectious love for the challenging aspects of science was invaluable when my project seemed to be falling apart. His attitude and scientific insights have inspired me to work harder in all areas of my life. Much of what I've learned over the past seven years began with him. I'd also like to thank my committee members, Dr. Allan Drummond, Dr. Joy Bergelson, and Dr. Bryan Dickinson. You and your lab members have been invaluable resources during my time as a graduate student.

I'd like to thank the members of the Ranganathan lab. Mike, thank you for keeping the lab running smoothly and for helping me with the actual science of biochemistry when I needed it. Rathi, Steph, and Miranda, thank you for keeping our lab organized and ensuring we had everything necessary for our research. BoRam, I appreciate both your scientific help and your friendship. Your work setting up PACE and teaching me to use it was crucial to this thesis. Steven, thank you for organizing social events and for encouraging me, daily, to be more social in lab. Finally, thank you to all other past and present lab members (Bryan, Fabian, Peter, Emily, Pierce, Xinran, Jeremy, Yaakov, Tyler, Lauren, Nikša, Sean, Riccardo, Eric, Diane, Nitya, Ankita, Sarah, Yujiao, Andrew, and Sonia) who contributed to this work in various ways.

I'd also like to thank Kabir Husain, formerly a post-doc in Arvind Murugan's lab and now a Lecturer at University College London. You've been like a second advisor, helping

me think through theory and design experiments I wouldn't have considered on my own. A significant portion of this thesis is the result of our conversations. Thank you.

Outside the lab, I'm grateful to the BMB department for supporting me throughout graduate school, especially the administrative staff (Lisa, Shani, Giovanni) and fellow graduate students.

Thank you to my friends, both within UChicago (Mel, Katie, Jordan, Brittany, Madeline, Chad, Cassidy) and outside (Alex, Mahmoud, Andrew, Laura), for helping me unwind and for never talking about work with me. I truly appreciated it.

Lastly, and most importantly, I want to thank my family. To my pet rabbit, Christopher Robun, who will undoubtedly pour over the details of this thesis, thanks for being a calming (but demanding) presence. Mom, Dad, Amber, Tyler, your support has meant everything to me. Mom and Dad, thank you for instilling a work ethic that made this possible. Michelle, thank you for your constant love and support. I don't know how I would have made it through the pandemic and these last few years without you. Thank you.

Abstract

This thesis examines the relationship between conditional neutrality, protein sequence encoding, and the role of evolutionary history in shaping protein adaptability. By analyzing algorithmically designed synthetic variants of the ligand-binding protein $PSD95^{pdz3}$ – where only the core epistatic unit was preserved, and the surrounding constraints were scrambled – I show that these surrounding constraints are essential for adaptation to new functional challenges. This finding, that certain sequence constraints are crucial for evolution, combined with the understanding that these constraints are themselves shaped by evolutionary history, suggests that a protein's past influences its capacity for adaptation. Preliminary results using a continuous evolution system (BTH-PACE) support this, demonstrating that different rates of environmental fluctuation lead to distinct patterns of constraints and varying abilities to generate conditional neutrality for alternate ligands. This underscores the role that evolutionary history has in shaping the pattern of sequence constraints on proteins. To further investigate evolutionary dynamics, I introduce HiDenSeq, a novel method for quantifying key evolutionary parameters using the Luria-Delbrück distribution. These parameters include the distribution of fitness effects, selection pressure, and ability to generate conditional neutrality. Together, these findings offer new insights and tools for understanding the constraints that govern protein evolution.

Chapter 1. Introduction

Proteins have the capacity to fold, function, and adapt to changing selection pressures. Significant work has gone into understanding the pattern of amino acid interactions, defined as the constraints, that are necessary for folding and function[1-5]. However, how the need to adapt has shaped these constraints on protein sequences is not as well understood[6-7]. From a phenomenological perspective, an example of a constraint that seems unique to adaptive fitness is conditional neutrality (CN), the capacity of proteins to generate mutations that have no (or minimal) effect on current fitness but that provide a selective advantage when environments change[8]. The key feature of CN mutants is that these initially neutral mutations can persist long enough in the standing variation to facilitate paths to new fitness peaks as selection pressures vary. CN enhances adaptability by essentially decoupling the generation of phenotypic diversity from the need for that diversity. Thus, when selection pressures randomly vary, pre-existing CN mutations in the population can initiate a path of adaptation[8-9].

The concept of CN as a facilitator of evolution might be traced back to the work by Luria and Delbrück in the 1940s on viral resistance in bacterial cells[10]. In their initial work they describe what is now known as standing genetic variation (SGV) where bacterial cells survived a viral infection due to “mutations which arise independently of the action of the virus.” Not all SGV is relevant to adaptation, but subsequent work has shown that for a specific new function, the portion of SGV which allows for that function (productive variation) can be classified as either CN or direct switching (DS); the distinction between the two being the timescale of their persistence after occurrence.[11-12] Direct switching

variation is quickly selected against while CN, which maintains most (or all) of the current function, will persist in all but the strongest purifying selection.

The impact of conditional neutrality extends beyond individual mutations to influence broader evolutionary processes. Context dependent robustness to mutation can drive the retention of genetic diversity within populations by providing a reservoir of heritable genetic variation through a process called exaptation[13-16]. Exaptation is a fundamental concept in evolutionary biology that describes the process by which a trait, originally evolved for one function, is co-opted for a different purpose[17-18]. Unlike adaptations, which arise through natural selection for a specific function, exaptations are traits that acquire new functions, often without significant modifications to their original structure[16]. A classic example of an exaptation is the evolution of feathers. Initially, feathers likely evolved in dinosaurs for thermal regulation, but they later became essential for flight in birds[19-21]. Exaptation also occurs at the molecular scale such as the co-option of pancreatic trypsinogen as an antifreeze protein in Antarctic fish[22]. These shifts in function highlight the opportunistic nature of evolution, where existing structures can be repurposed to meet new ecological challenges.

By definition, CN mutations, which allow for future adaptation after a shift in selection, are an example of exaptations[23]. Such standing variation can be a source of evolutionary innovation, allowing populations to adapt quickly to shifting environments[24]. This phenomenon suggests, and modeling has shown, that the evolutionary significance of CN mutations can only be understood in a full ecological and genetic context, emphasizing the importance of epistasis (how different mutations interact) and pleiotropy[25]. Pleiotropy refers to the situation where a single gene affects multiple traits,

meaning that a mutation in one gene can have widespread, and occasionally conflicting, effects an organism's fitness[26]. Experimental investigation has shown the evolutionary capacity of CN holds true in experimental systems as broad as RNA molecules, cellular metabolic networks, and field studies of flowering plants (e.g., *Boechera stricta*)[27-29]. Put simply, CN allows for increased persistence of productive genetic variation, facilitating evolution through exaptation at all scales of biology.

In recent years, the structural basis for conditional neutrality has been explored. Evolution-based models reveal an architecture for proteins in which higher-order and collective epistasis are loaded in sparse, physically connected networks spanning from the active site to allosteric sites across the protein[1-3,30-31,]. This collective epistatic unit is called a "protein sector". Functionally relevant sectors have been seen in a broad range of protein families suggesting they may be a general feature of proteins. In WW domains the pattern of correlations that define the sector alone were shown to be sufficient to specify a soluble, foldable, and functional protein (Figure 1.1)[32-33]. In S1A serine proteases three quasi-independent sectors were found, each controlling a distinct phenotype for the overall protein (Figure 1.2)[3]. Furthermore, when mutations were generated within each respective sector the phenotypes were shown to vary independently, demonstrating that sectors represent a modular unit by which evolution can act on to affect change. It is with this logic that sectors have been proposed to represent a structural organization of proteins that reflect their evolutionary histories[34].

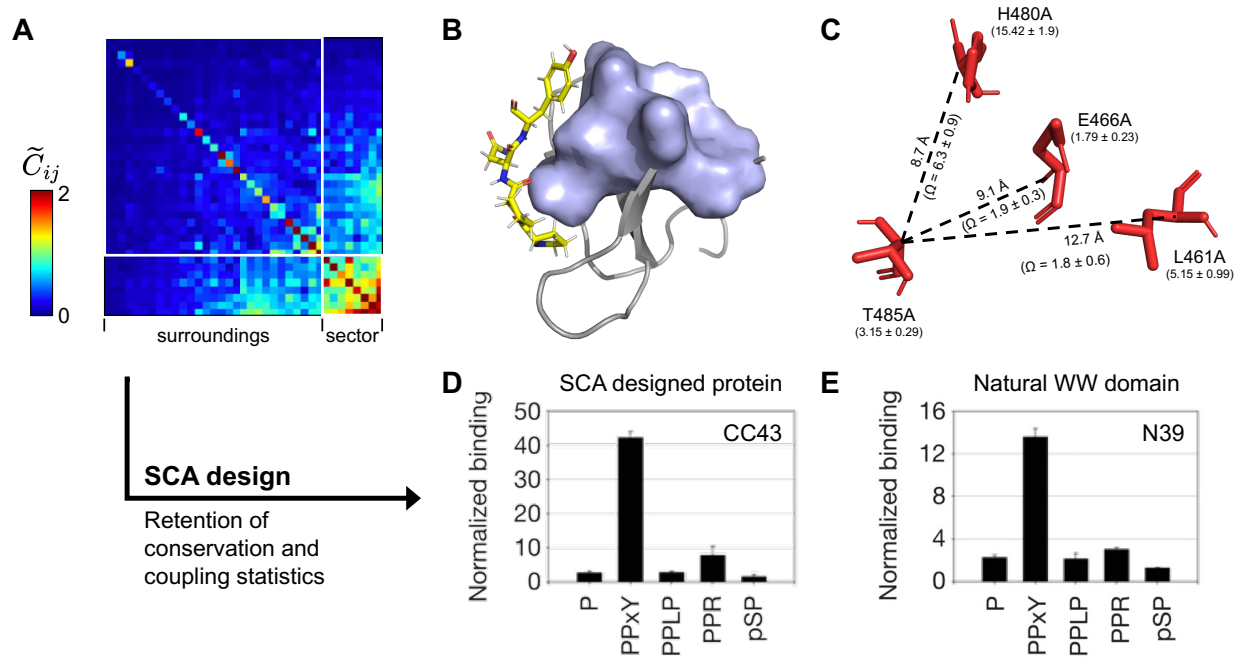

Figure 1.1. The sector defines a functional and folded protein

A. Adapted from Reynolds K, et al, 2013[35]. A clustering of the matrix of statistical coupling values ($\tilde{C}_{ij}$) in WW domains. Statistical coupling represents coevolution and a higher value (more red) indicates two residues have a high degree of correlation in the WW domain protein alignment. Clustering separates the positions in the protein into two groups: sectors and surroundings.

B. A three-dimensional structure of the representative Nedd4.3 WW domain bound to its ligand (PDB ID 1I5H, ligand in yellow). The sector is shown in blue and spans from the active site across the protein.

C. Adapted from Russ WP, et al, 2005[33]. Mutant cycle analysis of selected coevolving positions within the sector of the WW domain. The coupling parameter, Ω, is the ratio of the relative effect of a single mutation in the background of WT (X_1) to the relative effect of that same mutation in the background of another single mutation (X_2; $\Omega = X_1/X_2$); that is, the degree to which the effect of one mutation depends on the second. Residues are shown as they exist in B. Single mutations at all sites affect peptide binding (effect relative to wild type in parentheses), and mutant cycle analysis demonstrates energetic coupling ($\Omega > 1$) between position 485 and positions 461, 466 and 480. Measurements are mean ± s.d. This result indicates statistical coupling is predictive for epistasis.

D. Adapted from Russ WP, et al, 2005[33]. SCA designed proteins, which retain conservation and coupling statistics, were created. One such protein CC43, which has a 37% average and 68% top-hit identity to natural WW domains in the MSA, was assayed for ligand specificity using a peptide library screening. Binding is reported relative to background in the absence of target peptides.

E. Same as in D for N39, a natural WW domain with PPxY ligand specificity. CC43 is indistinguishable.

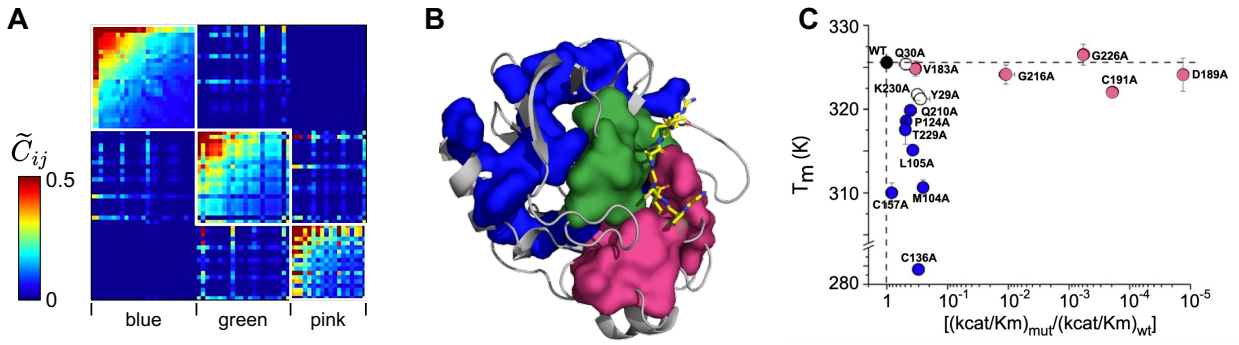

Figure 1.2. Sectors are evolvable units of a proteins architecture

A. Adapted from Halabi N, et al, 2009[3]. Data are depicted as in (1.1.A) for the S1A Protease family with all non-sector positions removed for clarity. The sector is composed of three independent coevolving units, termed the blue, green and pink sectors.

B. A three-dimensional structure of rat trypsin bound to its ligand (PDB ID 3TGI, ligand in yellow). Sectors are colored as outlined in (C). All three sectors have a distinct, orthogonal effects on protein function. The blue sector primarily spans the two β barrels of the protein and affects thermal stability. The green sector contains the catalytic triad and other residues known to be essential for the chemical mechanism of the enzyme. The pink sector comprises the S1 pocket and its surroundings and primarily controls catalytic activity.

C. Adapted from Halabi N, et al, 2009[3]. Plotted is the effect on thermal stability and catalytic power for alanine mutations made to rat trypsin. Dots are colored as in (B). The black dot is wildtype rat trypsin and white dots are mutations outside the sector. Mutations within a single sector of trypsin proteins effect function in only one dimension.

Examining sectors through the lens of CN provides further evidence that sectors are both shaped by evolution and have an active role in it. In a case study involving PSD95^{pdz3}, a member of the PDZ family of protein interaction modules, a comprehensive mutational analysis exposed the spatial organization of CN[36]. A deep mutational scan (DMS) determined the functional effects of all single mutations for binding of the PDZ domain to its canonical class I ligand and an alternative class II ligand. Mutations which provided class II function were designated either CN or direct switching mutations (DS); a designation that is dependent on the quantitative selection pressure. When mapped onto the structure of PSD95^{pdz3}, CN mutations, that is mutations which allowed for binding of the class II ligand without meaningfully altering class I ligand binding, almost exclusively occurred away from the active site (Figure 1.3). This contrasted with direct

switching mutations, where class II ligand binding was introduced at the cost of class I ligand binding, which occurred exclusively at the active site (Figure 1.3).

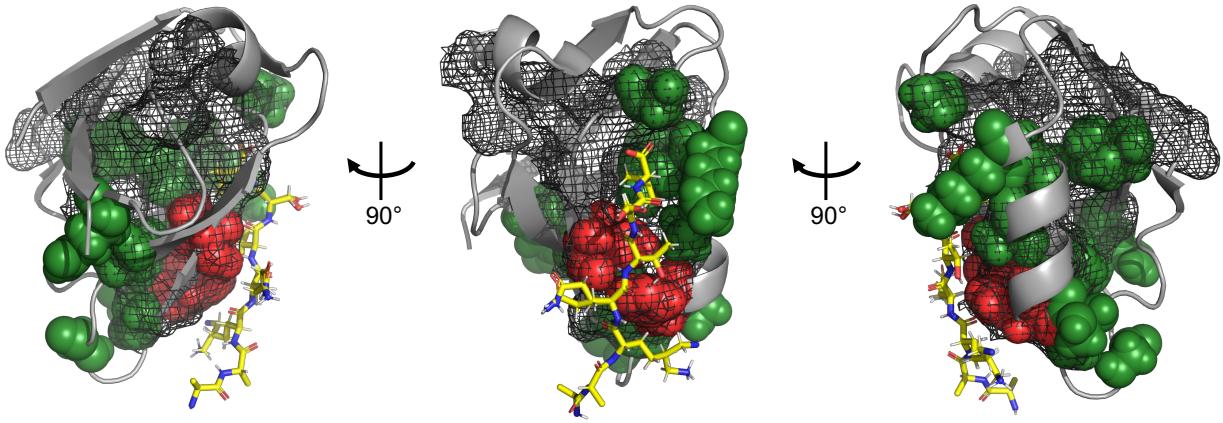

Figure 1.3. Spatial distribution of adaptive sites for class II ligand binding in PSD95^{pdz3}
Adapted from Raman AS, et al, 2016[36]. A three-dimensional structure of PSD95^{pdz3} bound to a class II ligand (ligand in yellow, PDB ID 5HED). The sector is shown in grey mesh. Class II ligand adaptive positions that are conditionally neutral for class I ligand binding are shown in green. Direct switching positions are shown in red. Nearly all adaptive positions are contained within the sector while direct switching mutations localize to the active site.

Strikingly, the conditionally neutral mutants occurred exclusively within and along the sector – the collectively epistatic unit – linking this structural feature of a protein to allostery and evolution. Thus, in addition to folding and function, sectors also play a role in enabling adaptation. Turned around, this implies that the origin and maintenance of sector architecture in proteins may stem from the need to adapt to changing selection pressures[36]. In this model, allosteric networks develop not by direct selection for protein regulation but are a byproduct of the need to support CN mutants which are statistically advantageous in fluctuating environments.

This model motivates the three main goals of this work. First, to define the architectural constraints necessary for adaptability, both in the sector and the non-sector surroundings. It is currently unclear whether the connection between the sector and its

surroundings is essential in natural proteins. Second, perform an initial test of the hypothesis that the pattern of mutations throughout evolution and the ability for CN function are dependent on the environmental fluctuation rate. Third, develop a high-throughput Luria-Delbrück style assay for quantification of evolutionary parameters as a foundation for future studies.

1.1. Role for non-sector surroundings in generating an evolvable protein

In Chapter 3 of this thesis, I explore the role for sector and non-sector constraints in adapting to the novel functional challenge of binding a new ligand. I use a previously developed, algorithmically designed protein ($C2_{34}$), of the small ligand-binding domain, $PSD95^{pdz3}$. This protein was generated by scrambling the non-sector positions (surroundings) and leaving the sector positions unchanged. Prior functional and biophysical measurements show that a thermally stabilized version of this synthetic protein is indistinguishable from $PSD95^{pdz3}$, but its ability to generate CN and therefore its ability to adapt to new selection conditions had not yet been tested. Here, I have developed a bacterial two-hybrid-based phage-assisted continuous evolution assay (BTH-PACE). This allowed direct competition of the natural and synthetic proteins in the context of random mutagenesis. I also used BTH-PACE to assay the capacity for adaptation by quantifying the maintenance of productive SGV, which is statistically likely to be CN, of each PDZ domain. Results from this work are consistent with the statement that the constraints specifying function and folding, which exist in the sector, are insufficient for specifying an evolvable protein. This provides, for the first time, a possible role for the surrounding constraints within the architecture of a protein.

1.2. The impact of evolutionary history on conditional neutrality

Every natural protein is the result of a set of selection pressures that can change over the course of their evolutionary history. A long standing hypothesis is that this pattern of selection pressures has impacted the constraints placed on individual protein sequences[24,37-39]. If this hypothesis is true, any two proteins which have different evolutionary histories should have differing abilities to generate CN. In this chapter, BTH-PACE was modified to allow for selection of binding to multiple ligands, with selection between ligands fluctuating at an experimentally defined rate. Six natural PDZ domains were each evolved under three distinct evolutionary regimes leading to proteins with different evolutionary histories. The resulting proteins were compared for their ability to bind to new ligands. An examination of the pattern of mutations for each evolutionary condition showed a dependence on the rate of fluctuation within their environment. This work represents the first *in vivo* attempt at quantifying the effect of evolutionary history on protein architecture.

1.3. Development of a high throughput Luria-Delbrück assay

The parameters defining the distribution of mutants in the classic Luria-Delbrück experiment offer a way to directly quantify a system's ability to generate CN as well as other evolutionary factors[40]. However, accurate parameterization of this distribution traditionally required hundreds to thousands of replicates, a logistically infeasible task until recently[41-42]. Through a novel experimental procedure called high-density Luria-Delbrück by sequencing (HiDenSeq), initially developed by Kabir Husain, we achieved

the goal of parameterizing the distribution. HiDenSeq utilizes advanced sequencing techniques to efficiently analyze a vast number of samples, thus overcoming the previous limitations. Additionally, with HiDenSeq, I provide the first experimental validation of an extension to the fundamental Luria-Delbrück distribution theory, demonstrating that the selection pressure of the system affects the distribution's shape. Although further extensions to the theory remain to be tested, this finding opens new avenues for quantifying and understanding evolutionary processes.

1.4. References

1 Lockless SW & Ranganathan R. Evolutionarily conserved pathways of energetic connectivity in protein families. *Science* **286,** 295-299, (1999).

2 Süel GM, *et al*. Evolutionarily conserved networks of residues mediate allosteric communication in proteins. *Nature Structural and Molecular Biology* **10**, 59-69, (2003).

3 Halabi N, *et al*. Protein sectors: evolutionary units of three-dimensional structure. *Cell* **138,** 774-786, (2009).

4 Weigt M, *et al.* Identification of direct residue contacts in protein-protein interaction by message passing. *PNAS* **106**, 67-72, (2009).

5 Morcos N, *et al.* Direct-coupling analysis of residue coevolution captures native contacts across many protein families. *PNAS Plus* **108**, E1293-E1301, (2011).

6 Payne JL & Wagner A. The causes of evolvability and their evolution. *Nature Reviews Genetics* **20**, 24-38 (2019)

7 Wagner A. Evolvability-enhancing mutations in the fitness landscapes of an RNA and a protein. *Nature Communications* **14**, 3624, (2023).

8 Paaby AB & Rockman MV. Cryptic genetic variation: evolution's hidden substrate. *Nature Reviews Genetics* **15** 247-258, (2014).

9 Draghi JA & Plotkin JB. Hidden diversity sparks adaptation. *Nature* **474**, 45-46, (2011).

10 Luria SE & Delbrück M. Mutations of bacteria from virus sensitivity to virus resistance. *Genetics* **28**, 491-511, (1943).

11 Shimon B, *et al*. Robustness-epistasis link shapes the fitness landscape of a randomly drifting protein. *Nature* **444**, 929-932 (2006).

12 Stiffler MA, *et al*. Evolvability as a function of purifying selection in TEM-1 ß-Lactamase. *Cell* **160**, 882-892, (2015).

13 Bloom JD, *et al*. Protein stability promotes evolvability. *PNAS* **103**, 5869-5874, (2006).

14 Tokuriki N & Tawfik DS. Stability effects of mutations and protein evolvability. *Current Opinion in Structural Biology* **19**, 596-604, (2009).

15 Wagner A. Robustness, evolvability, and neutrality. *FEBS Letters* **579**, 1873-3468, (2015).

16 Frenkel-Pinter M, *et al.* Adaptation and exaptation: from small molecules to feathers. *Journal of Molecular Evolution* **90**, 166-175, (2022).

17 Gould SJ & Vrba ES. Exaptation-A missing term in the science of form. *Paleobiology* **8**, 4-15, (1982).

18 Gould SJ. The exaptive excellence of spandrels as a term and prototype. *PNAS* **94**, 10750-10755, (1997).

19 Prum RO. Development and evolutionary origin of feathers. *Journal of Experimental Zoology* **285**, 291-306, (2002).

20 Dhouailly D, *et al.* Getting to the root of scales, feather and hair: As deep as odontodes? *Experimental Dermatology* **28**, 503-508, (2017).

21 Pan Y, *et al*. The molecular evolution of feathers with direct evidence from fossils. *PNAS* **116**, 3018-3023, (2019).

22 Chen L, *et al*. Evolution of antifreeze glycoprotein gene from a trypsinogen gene in Antarctic notothenioid fish. *PNAS* **94**, 3811-3816, (1997).

23 Brosius J. Exaptation at the molecular genetic level. *Science China Life Sciences* **61**, 437-452, (2019).

24 Dellus-Gur E, *et al*. What makes a protein fold amendable to functional innovation? Fold polarity and stability trade-offs. *Journal of Molecular Biology* **425**, 2609-2621, (2013).

25 Draghi JA, *et al.* Epistasis Increases the Rate of Conditionally Neutral Substitution in an Adapting Population. *Genetics* **187**, 1139-1152, (2011).

26 Stearns FW. One hundred years of pleiotropy: A retrospective. *Genetics* **186**, 767-773, (2010).

27 Hayden EJ, *et al*. Cryptic genetic variation promotes rapid evolutionary adaptation in an RNA enzyme. *Nature* **474**, 92-95, (2011).

28 Barve A & Wagner A. A latent capacity for evolutionary innovation through exaptation in metabolic systems. *Nature* **500**, 203-206, (2013).

29 Anderson JT, *et al*. Genetic trade-offs and conditional neutrality contribute to local adaptation. *Molecular Ecology* **22**, 699-708, (2013).

30 McLaughlin RN, *et al*. The spatial architecture of protein function and adaptation. *Nature* **491**, 138-142, (2012).

31 Rivoire O, *et al.* Evolution-Based functional decomposition of proteins. *PLoS Computational Biology* **12**, e1004817, (2016).

32 Socolich M, *et al*. Evolutionary information for specifying a protein fold. *Nature* **437**, 512-518, (2005).

33 Russ WP, *et al.* Natural-like function in artificial WW domains. *Nature* **437**, 579-583, (2005).

34 Wang SW, *et al*. Revealing evolutionary constraints on proteins through sequence analysis. *PLoS Computational Biology* **15**, e1007010, (2019).

35 Reynolds K, *et al*. Evolution-based design of proteins. *Methods in Enzymology* **523**, 213-235, (2013).

36 Raman AS, *et al*. Origins of allostery and evolvability in proteins: A case study. *Cell* **166**, 468-480, (2016).

37 Murugan A, *et al*. Roadmap on biology in time varying environments. *Physical Biology* **18**, 041502, (2021).

38 Pigliucci M. Is evolvability evolvable? *Nature Reviews Genetics* **9**, 75-82, (2008).

39 Kosonocky CW & Ellington AD. Evolving to evolve, Dan Tawfik's insights into protein engineering. *Biochemistry* **62**, 145-147, (2023).

40 Kessler DA & Levine H. Scaling Solution in the Large Population Limit of the General Asymmetric Stochastic Luria–Delbrück Evolution Process. *Journal of Statistical Physics* **158**, 783-805, (2015).

41 Lang GI & Murray AW. Estimating the per-base-pair mutation rate in Yeast Saccharomyces cerevisiae. *Genetics* **178**, 67-82, (2008).

42 Gou L, *et al*. The genetic basis of mutation rate variation in yeast. *Genetics* **211**, 731-740, (2019).

Chapter 2. Methods Development

To explore the structural principles of evolution, an experimental system is needed that allows control over key evolutionary parameters – mutation rate, population size, selection pressure, and selection conditions. This chapter describes the development of such a system, called the bacterial two-hybrid-based phage-assisted continuous evolution assay (BTH-PACE). BTH-PACE was adapted from the previously established phage-assisted continuous evolution (PACE) method, building on work by BoRam Lee in the Ranganathan lab.

2.1. Bacterial-two-hybrid phage assisted continuous evolution (BTH-PACE)

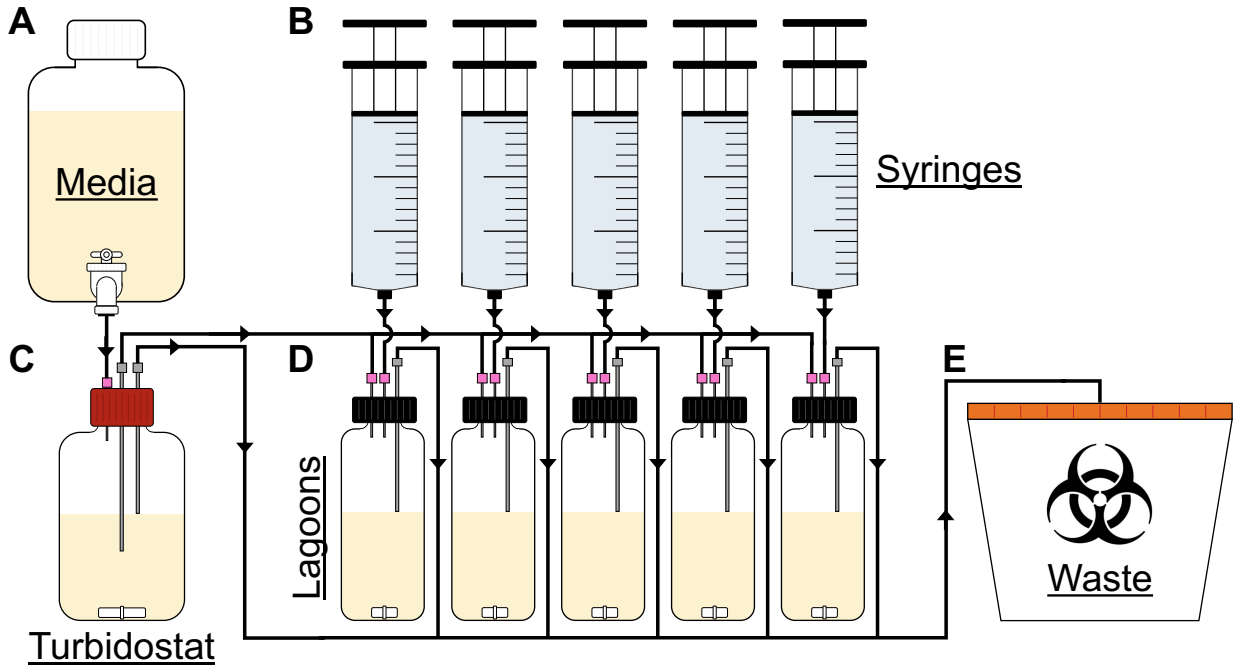

Figure 2.1. Media flows in BTH-PACE
Media is stored in a large carboy (A), which flows into a turbidostat (C) containing bacterial host cells which are kept at a constant cell density. Bacterial cells flow from the turbidostat to vials, termed lagoons, containing M13 bacteriophage and bacterial cells (D). Evolution of phage in BTH-PACE occurs in the lagoons. Selection pressure in BTH-PACE is in part determined by the dilution rate of the lagoons. Chemicals to set the mutation rate (arabinose) and the selection pressure (doxycycline) are supplied to the lagoons via syringes (B). To keep volumes constant, media from the turbidostats and lagoons is constantly flowed into a waste container (E).

In PACE (Figure 2.1), selection and subsequent evolution occur in the lagoon (Figure 2.1.D, Figure 2.2.B)[1]. Here, M13 bacteriophage, which have had their essential gene III (gIII) knocked out, infect bacterial host cells. Without the induction of gIII to subsequentially produce protein III (pIII), M13 phage are still be produced by the bacterial host cells, but they cannot reinfect new cells due to a defect in F-pilus-mediated cell entry caused by the missing pIII protein. To overcome this deficiency, pIII must be produced by an alternative mechanism, which defines the selection process in PACE. In its original form, PACE selected for improved binding of T7 RNA polymerase to a T3 promoter that regulated gIII expression. In BTH-PACE, gIII expression depends on the function of a PDZ domain. PDZ domains are small ligand-binding proteins, about 100 amino acids in size, that bind C-terminal amino acids with approximately 1 μM affinity and display sequence specificity[2,3].

A bacterial two-hybrid system was implemented in which the phage encoded PDZ domain is linked to the RNA polymerase ω subunit, while the bacterial host cell contains the 434 cI DNA-binding domain linked to a PDZ ligand (Figure 2.2.A). When the PDZ domain binds the ligand, gIII is expressed, allowing those phage encoding the PDZ domain to propagate (Figure 2.2.B). To confirm that BTH-PACE was selecting for the ligand-binding function of the PDZ domain, a standard curve was created using an existing library of 83 single mutants of the PDZ domain $PSD95^{pdz3}$, all of which had their K_D values measured against the canonical $PSD95^{pdz3}$ ligand, CRIPT (Figure 2.3.A)[4]. When this library was grown for 6 hours in BTH-PACE, the enrichment of any particular PDZ domain correlated monotonically with its known binding strength to the CRIPT ligand (Figure 2.3.B).

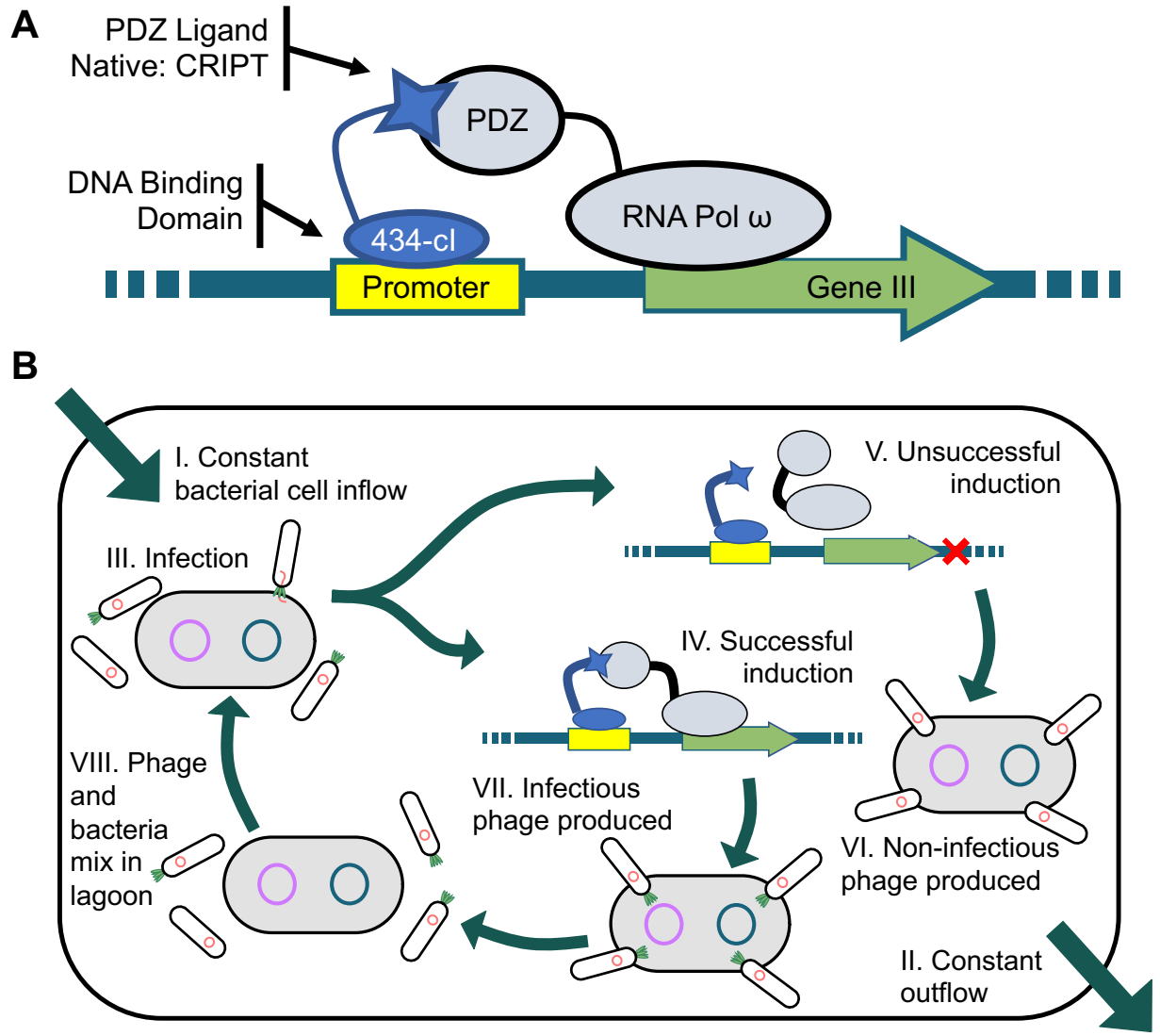

Figure 2.2. Selection in BTH-PACE

A. Selection in BTH-PACE depends on binding of a PDZ domain to its ligand. The phage-encoded PDZ domain is bound to the ω subunit of RNA polymerase. The ligand, which is bound to the 434-cI DNA binding domain, is expressed off the host-cell encoded accessory plasmid (AP). Binding of the PDZ domain to its ligand causes expression of protein III (pIII, encoded by gene III; gIII on AP), which is essential for F pilus mediated M13 bacteriophage propagation.

B. In BTH-PACE evolution occurs in a lagoon, pictured here. There is a constant inflow of bacterial host-cells (I) and outflow (II) with a dilution rate fast enough to ensure the bacteria are not evolving in the lagoon. Upon entry into the lagoon, bacterial cells are infected by phage (III) and the PDZ domain encoded in the phage genome is expressed. If the PDZ domain can bind the ligand, gIII is induced (IV). If the PDZ domain cannot bind the ligand, gIII is not induced (V) and non-infectious phage are produced which wash out of the lagoon (VI). Phage which encode a PDZ domain allowing for expression of pIII, produce infectious phage (VII). These phage are mixed into the lagoon (VIII) and are free to infect new bacterial host cells (III).

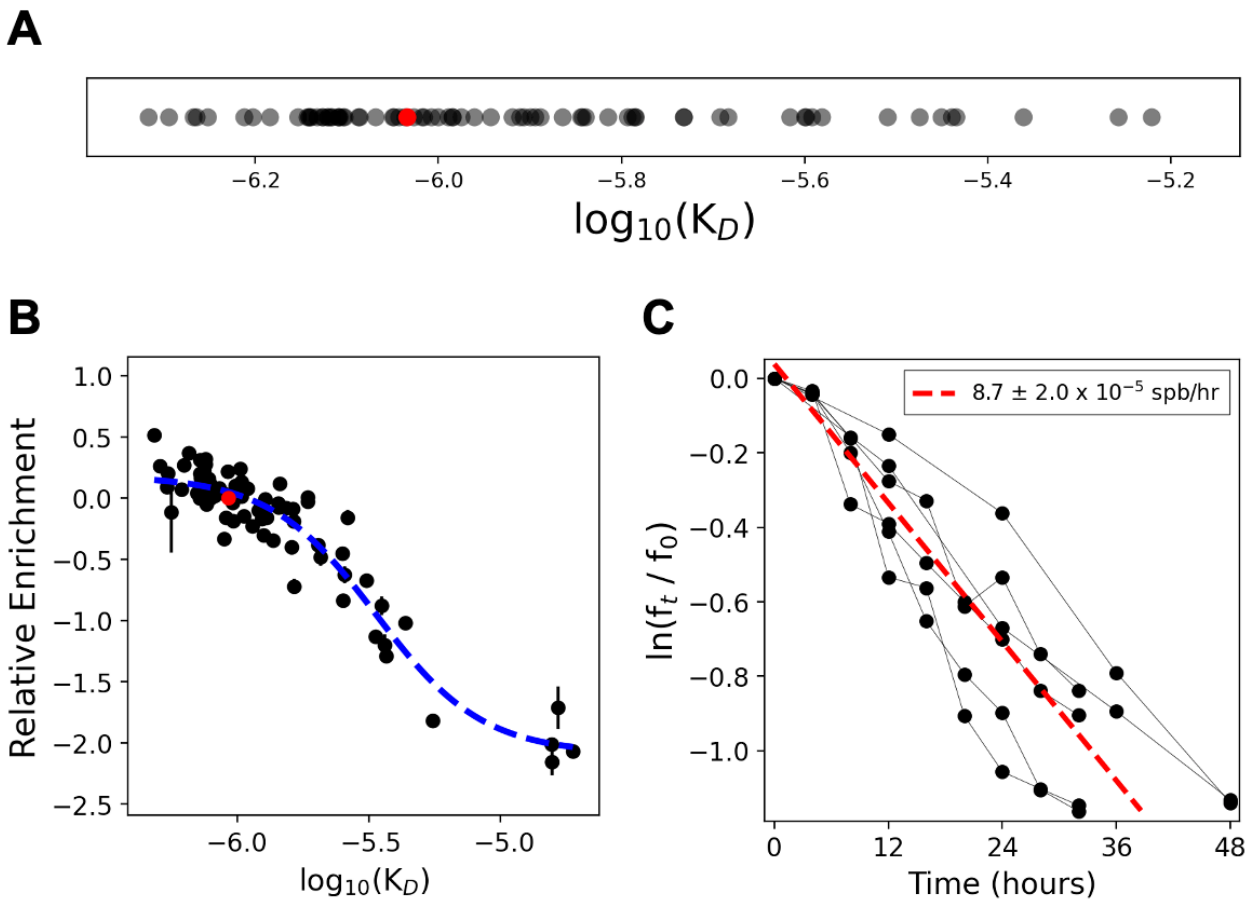

Figure 2.3. Selection and mutation in BTH-PACE

A. A library of single point mutants to PSD95^{pdz3} with known binding strengths[4]. The binding dissociation constant of each mutant (black dots) and PSD95^{pdz3} (red) to the CRIPT ligand is shown.

B. The library from (A) was grown in BTH-PACE for 6 hours. The binding strength of each mutant (black dots, PSD95^{pdz3} red dot) is plotted against a measure of fitness termed relative enrichment. Relative enrichment is defined as $RE = log_{10}(f_{i,t}/f_{i,0}) - log_{10}(f_{WT,t}/f_{WT,0})$ where $f_{i,t}$ is the frequency of a mutant i at time t and wildtype is PSD95^{pdz3}. Error bars in relative enrichment values are determined from three replicates of this experiment. A sigmoid curve (blue, dashed line) has been fit to the data.

C. The mutation rate (red, dashed line) is determined from averaging out linear fits to drops in the natural log of relative frequency ($\ln(f_t/f_0)$, where f_t is the frequency of PSD95^{pdz3} at time t) over time. Each independent experiment (black dots connected by black lines) is shown.

One of the key advantages of BTH-PACE is the ability to selectively increase the mutation rate of the phage without affecting the mutation rate in the infected bacterial cells. This ensures that only the phage evolves, simplifying both the experimental setup and the interpretation of results. The mutation rate in the phage is controlled by the concentration of arabinose in the lagoon, which induces the P_{BAD} promoter to express

mutator genes[5]. To quantify the mutation rate in BTH-PACE, an experiment was conducted where the initial population in each lagoon consisted entirely of phage containing the PSD95^{pdz3} PDZ domain. Critically, non-selective bacterial host cells (*E. coli* S2208) were used[6]. S2208 cells produce pIII upon M13 bacteriophage infection, regardless of the PDZ domain's function. In this configuration, any decrease in the frequency of PSD95^{pdz3} phage initially occurs at a rate proportional to the mutation rate.

$$\ln(f_t/f_0) = -\mu l t \quad \text{(Eq. 2.1)}$$

This relationship, where μ is the mutation rate in terms of substitutions per base pair per hour (spb/hr), l is the length of the PDZ domain, and t is time in hours, holds true as long as $\mu l \mathrm{N} >> 1$, where N is the population of phage in the lagoon. Replicates of this experiment produced a mutation rate of 8.7 ± 2.0 x 10^{-5} spb/hr at an arabinose concentration of 25 mM and a lagoon turnover rate of two times per hour (Figure 2.3.C). This is the condition that is used for all BTH-PACE experiments in this thesis.

Carlson *et al.* described a method for controlling selection pressure in a PACE system by producing pIII independently of ligand binding[7]. Specifically, pIII production is induced by anhydrotetracycline (ATc) via the small molecule-inducible TetA promoter ($P_{psp\text{-}tet}$) which decreases selection pressure as the concentration of ATc increases. However, replicating this system outside of the Liu Lab, including in the Ranganathan lab, has been challenging, likely due to ATc's light sensitivity. This issue was resolved by using doxycycline, an analog of tetracycline that is much less sensitive to light[8,9]. In a non-continuous implementation of BTH-PACE (Methods 7.3), the addition of 500 ng/mL doxycycline completely removed the selection function from the selection cells, causing them to behave like non-selection cells (Figure 2.4.A). In BTH-PACE, a doxycycline

concentration as low as 100 ng/mL in the lagoon was sufficient to eliminate selection pressure and disrupt the relationship between binding energy and relative enrichment (Figure 2.4.B).

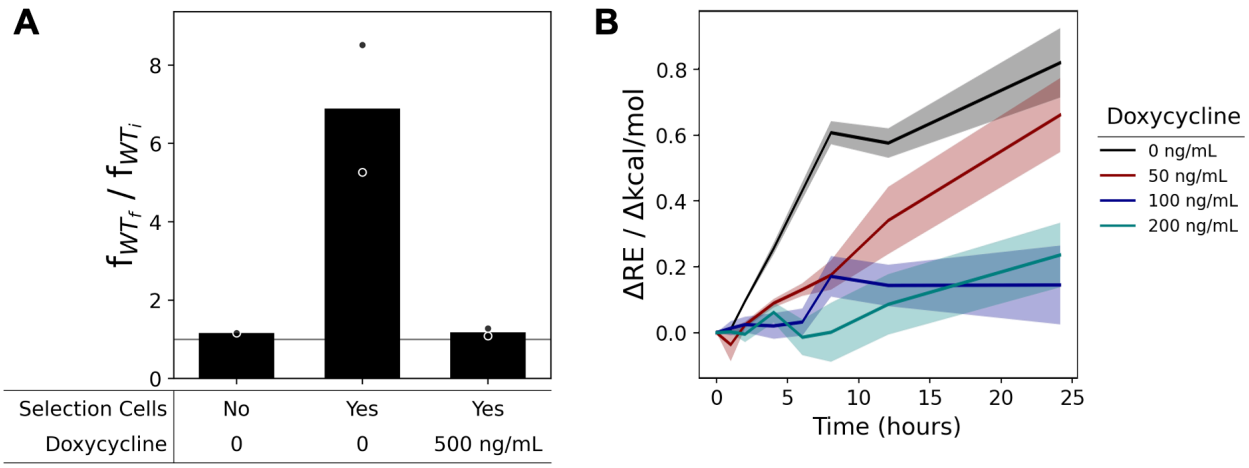

Figure 2.4. Doxycycline lowers selection pressure in BTH-PACE

A. 1% phage containing PSD95^{pdz3} were mixed with 99% of a null phage that do not contain a PDZ domain. This mixture was grown for 4 hours in batch culture under three separate conditions. Presence or absence of selection cells indicates growth in S2060 + AP-CRIPT + DP6 or S2208 bacterial host cells respectively. The change in the frequency of PSD95^{pdz3} ($f_{WT,f}/f_{WT,i}$) is assayed (dots; individual replicates, bars; condition mean).

B. The library from Figure 2.3.A was grown in BTH-PACE for 24 hours at four different doxycycline concentrations and three replicates per concentration. Plotted is a measure of the selection strength, the change in relative enrichment (ΔRE) per change in binding energy (Δkcal/mol), at each time (colored lines shown with error in shading). This quantity is found by fitting a line to data as in Figure 2.3.B where the x-axis has been transformed from binding strength to binding energy.

Using doxycycline to control the selection pressure avoids the common workaround of altering the turnover rate in the lagoon[10]. Although lowering the turnover rate will decrease the selection pressure it comes at the cost of increasing the average generation time of the phage. When hundreds of generations of evolution are required, this increase in experimental time can become prohibitive.

Complications in BTH-PACE come from non-biological sources as well. For the experiments in this study, PACE was modified to enable turbidostatic growth of bacterial cells. Turbidostatic growth keeps the bacteria at fixed density, creating a more consistent

environment for the phage and lowering the chance of phage washout. To minimize biofilm formation and buildup in BTH-PACE experiments, proper management of media is critical. This is addressed by changing the host cell turbidostat, lagoons, and media flow lines from the turbidostat to the lagoons every other day (Figure 2.1). Additionally, a fresh bacterial host cell culture, grown from an overnight culture, was used to restart the turbidostat every four days to limit biofilm growth. With the modifications and quantification of parameters discussed in this section, BTH-PACE experiments can now be run indefinitely, while controlling the mutation rate and selection pressure of the population. Detailed information on media and chemical/drug concentrations can be found in the methods section (Methods 7.1).

2.2. References

1 Esvelt KM, *et al*. A system for the continuous directed evolution of biomolecules. *Nature* **28**, 499-503, (2011).

2 Lee HJ & Zheng JJ. PDZ domains and their binding partners: structure, specificity, and modification. *Cell Communication and Signaling* **8**, 8, (2010).

3 Stiffler MA, *et al*. PDZ domain binding selectivity is optimized across the mouse proteome. *Science* **20**, 364-369, (2007).

4 McLaughlin Jr RN, *et al*. The spatial architecture of protein function and adaptation. *Nature* **491**, 138-142, (2012).

5 Badran AH & Liu DR. Development of potent in vivo mutagenesis plasmids with broad mutational spectra. *Nature Communications* **6**, 8425, (2015).

6 Badran AH, *et al*. Continuous evolution of Bacillus thuringiensis toxins overcomes insect resistance. *Nature* **5**, 58-63, (2016).

7 Carson JC, *et al*, Negative selection and stringency modulation in phage-assisted continuous evolution. *Nature Chemical Biology* **10**, 216-222, (2014).

8 Baumschlager A, *et al*. Exploiting natural chemical photosensitivity of anhydrotetracycline and tetracycline for dynamic and setpoint chemo-optogenetic control. *Nature communications* **11**, 3834, (2020).

9 Álvarex-Esmorís C, *et al.* Degradation of Doxycycline, Enrofloxacin, and Sulfamethoxypyridazine under Simulated Sunlight at Different pH Values and Chemical Environments. *Agronomy* **12**, 260, (2022).

10 Hew BE, *et al*. Directed evolution of hyperactive integrases for site specific insertion of transgenes. *Nucleic Acids Research* **52**, e64, (2024).

Chapter 3. Role for non-sector surroundings in generating an evolvable protein

3.1. Abstract

Proteins display the ability to fold into a native state, carry out biochemical reactions, and evolve as conditions of selection fluctuate in the environment. Much work has gone into understanding the sequence constraints underlying folding and function, but the structural features of proteins that enable evolvability are less well studied. Here, we algorithmically designed a synthetic variant of the ligand-binding protein, $PSD95^{pdz3}$, which retained only constraints that define the collective epistatic unit of the protein (the sector) and scrambled the non-sector surroundings ($C2_{34}$). Mutations to $C2_{34}$ generate a protein ($C2_{34}$-TM) which is nearly as thermally stabilized as $PSD95^{pdz3}$ and displays near-wildtype function but fail to adapt when asked to bind a second, class-switching ligand. Results from a deep mutational scan for binding function and bacterial-two-hybrid phage assisted continuous evolution (BTH-PACE) experiments show that the mechanistic basis for this deficiency is not simply due to a small reduction in thermal stability. More likely, a model is proposed where the synthetic variants suffer from an inability to engage an alternative allosteric network that is preferred during adaption. As a result, a special class of mutations (called conditionally neutral) that pre-exist in the standing genetic variation and facilitate adaptation are systematically depleted. These data link the intramolecular architecture of a protein to its capacity to evolve, demonstrating at the molecular level a general design principle likely to underlie all evolvable protein systems.

3.2. Introduction

Proteins exhibit the ability to fold, perform essential functions, and evolve in response to selection pressures. While significant progress has been made in understanding the constraints governing folding and function, how evolution has constrained protein sequences is less thoroughly understood[1-7]. One known constraint is conditional neutrality (CN), defined as a mutation which, in a specific genotype, environment, and quantitative selection pressure, has little effect on fitness but, when either the genotype or environment changes, provides a fitness advantage[8].

CN mutants were first shown to facilitate evolution through exaptation in work on viral resistance in bacterial cells by Luria and Delbrück in 1943[9]. The term exaptation refers to a process in evolution where a trait, structure, or feature, originally evolved for one function is co-opted for a different purpose[10-12]. Unlike adaptation, which involves the direct shaping of a trait by natural selection for a specific role, exaptation describes how existing features can acquire new functions in response to changing environmental or biological pressures[13-14]. Although Luria and Delbrück did not yet have the term exaptation in 1943, their discussion of the “immunity of hereditarily predisposed individuals” for viral infection describes exactly this process. Pre-existing CN mutants serve as textbook examples of exaptation, and further studies have demonstrated that this evolutionary role of conditional neutrality is consistent across various experimental systems, including RNA molecules, cellular metabolic networks, and field research on the flowering plant *Boechera stricta*[15-20]. Their ability to persist in standing genetic variation (SGV) enables CN mutants to facilitate exaptation across different biological scales[21].

In recent years, the structural basis for conditional neutrality has been explored. Sequence-based models reveal an architecture for proteins in which higher-order collective epistasis is loaded into sparse, physically connected networks spanning from the active site to allosteric sites across the protein[1-3,22-23]. This network, called the sector, has been linked to various protein features, including thermostability, ligand binding, catalysis, and allostery[3,24-25]. Although CN mutants have been found to exist within the sector, an open question remains as to the role the non-sector surroundings play in protein evolution and the generation of CN[26].

Previously in the Ranganathan lab, proteins were algorithmically designed in which the non-sector surroundings of a small ligand-binding protein, $PSD95^{pdz3}$, were scrambled while the sector remained intact. One such designed protein, $C2_{34}$, was shown to be compact in its native state and had near-native ligand binding function. However, $C2_{34}$ is thermally unstable, exists in a high-entropy native state, and is not robust to mutations. By introducing three stabilizing point mutations to the non-sector region of $C2_{34}$, a new protein was made ($C2_{34}$-TM) which is still functional and now approaches the thermal stability of the natural PDZ domain while being mutationally robust. Through many functional and biophysical measurements, $C2_{34}$-TM is indistinguishable from $PSD95^{pdz3}$, but its ability to generate CN and its related ability to adapt to new selection conditions has not yet been explored.

In this work, we have implemented a bacterial-two-hybrid based phage-assisted continuous evolution assay (BTH-PACE). BTH-PACE, which has an elevated mutation rate for an evolving gene of interest, allowed us to directly compete the natural and synthetic proteins in the context of mutational stress. We also extended BTH-PACE to

allow quantification of the ability of each protein to maintain SGV. This work is consistent with the statement that the constraints specifying a capacity for evolution exist, at least in part, separate from those which are required for function and folding. This provides, for the first time, a role for the non-sector surroundings within the architecture of a protein.

3.3. Results

Biophysical characterization of synthetic proteins

The key to this work lies in the design of the synthetic proteins, $C2_{34}$ and $C2_{34}$-TM. Using a Markov chain Monte Carlo (MCMC) method, the sequence of the small ligand binding PDZ domain, $PSD95^{pdz3}$, was mutated so that positions which are constrained by higher order epistasis are preferentially retained (Figure 3.1.A-B). This generated sequences where the sector remains largely intact while the non-sector surroundings are scrambled (Figure 3.1.C). One such protein, $C2_{34}$, has a similar binding affinity and specificity to the canonical Class I ligand, CRIPT (-TKNYKQ**T**SV-COOH, derived from the cysteine-rich interactor of $PSD95^{pdz3}$, 0.452 $\pm$ 0.028 μM K_D $C2_{34}$ and 1.19 $\pm$ 0.12 μM K_D for $PSD95^{pdz3}$, Figure 3.1.D). However, $C2_{34}$ is thermally unstable (32°C melting temperature, T_m, measured by differential scanning calorimetry (DSC) versus 71°C for $PSD95^{pdz3}$) and has a highly entropic native state ensemble (Figure 3.1.E-F).

Predictably, and consistent with prior work relating stability and robustness, $C2_{34}$ was shown not to be robust to mutations as measured in a deep mutation scan (DMS, Figure 3.2.B)[20,22,27]. However, three thermal stabilizing mutations in the non-sector surroundings, which produces the protein $C2_{34}$-TM (58°C T_m), are sufficient to recover much of the lost stability at physiological temperatures while retaining binding to the

CRIPT ligand (1.01 ± 0.05 μM). Additionally, $C2_{34}$-TM recovers robustness to mutations in the context of binding the CRIPT ligand (Figure 3.2.C,F). Many biochemical and biophysical analyses demonstrate that $C2_{34}$-TM and $PSD95^{pdz3}$ are nearly identical.

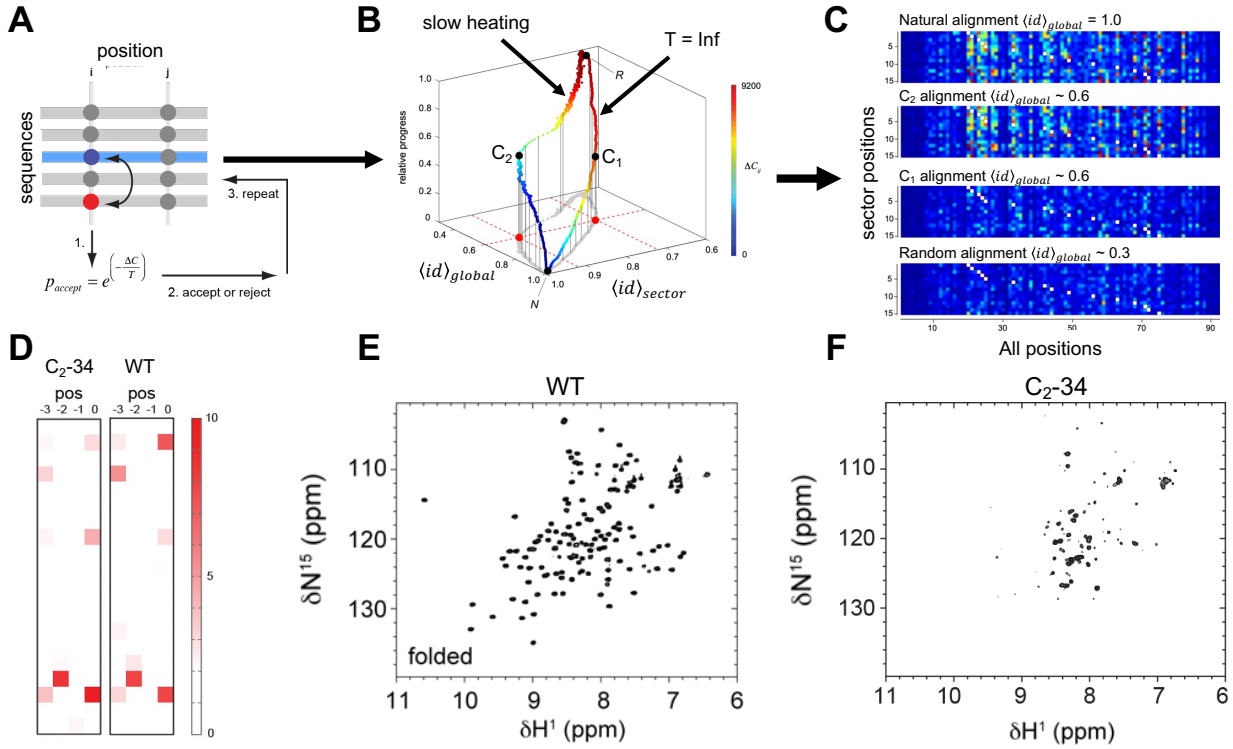

Figure 3.1. Algorithmic separation of sector constraints in $C2_{34}$

A. A MCMC approach was used to swap amino acids within the original $PSD95^{pdz3}$ (WT) protein sequence. In each iteration of MCMC, two random amino acids, both at a given site i, are selected for a swap. This ensures site specific conservation is always maintained. The swap is accepted with a probability dependent on its effect on the statistical couplings at all other sites, j, and some temperature, T. High temperatures accept all mutations and low temperatures accept only mutations which have no effect on the statistical coupling between positions in the sequence.

B. Two heating trajectories were used, C_1 and C_2. In C_1, temperature is always set to infinity meaning all swaps are accepted and only site specific conservation is retained. In C_2, the temperature starts low and then gradually increases over thousands of iterations of MCMC. This ensures that non-sector positions are preferentially mutated.

C. Matrix of statistical coupling values for four sets of proteins. The mean maximum sequence similarity, ⟨id⟩, for each sequence to a sequence in the natural alignment is shown. Although C_1 and C_2 generated proteins have similar ⟨id⟩, only the C_2 alignment preserves the coupling seen in the natural alignment.

D. Binding profile of one C_2 protein, $C2_{34}$, and $PSD95^{pdz3}$ is shown for the four C-terminal amino acids in a PDZ ligand. The natural log of the relative preference over background is shown colorimetrically.

E. ^{1}H-^{15}N HSQC NMR on 200 μM $PSD95^{pdz3}$. A dispersed spectrum of peaks indicates the existence of a well packed protein.

F. ^{1}H-^{15}N HSQC NMR on 200 μM $C2_{34}$. Collapsed peaks indicate a protein in a high entropy state.

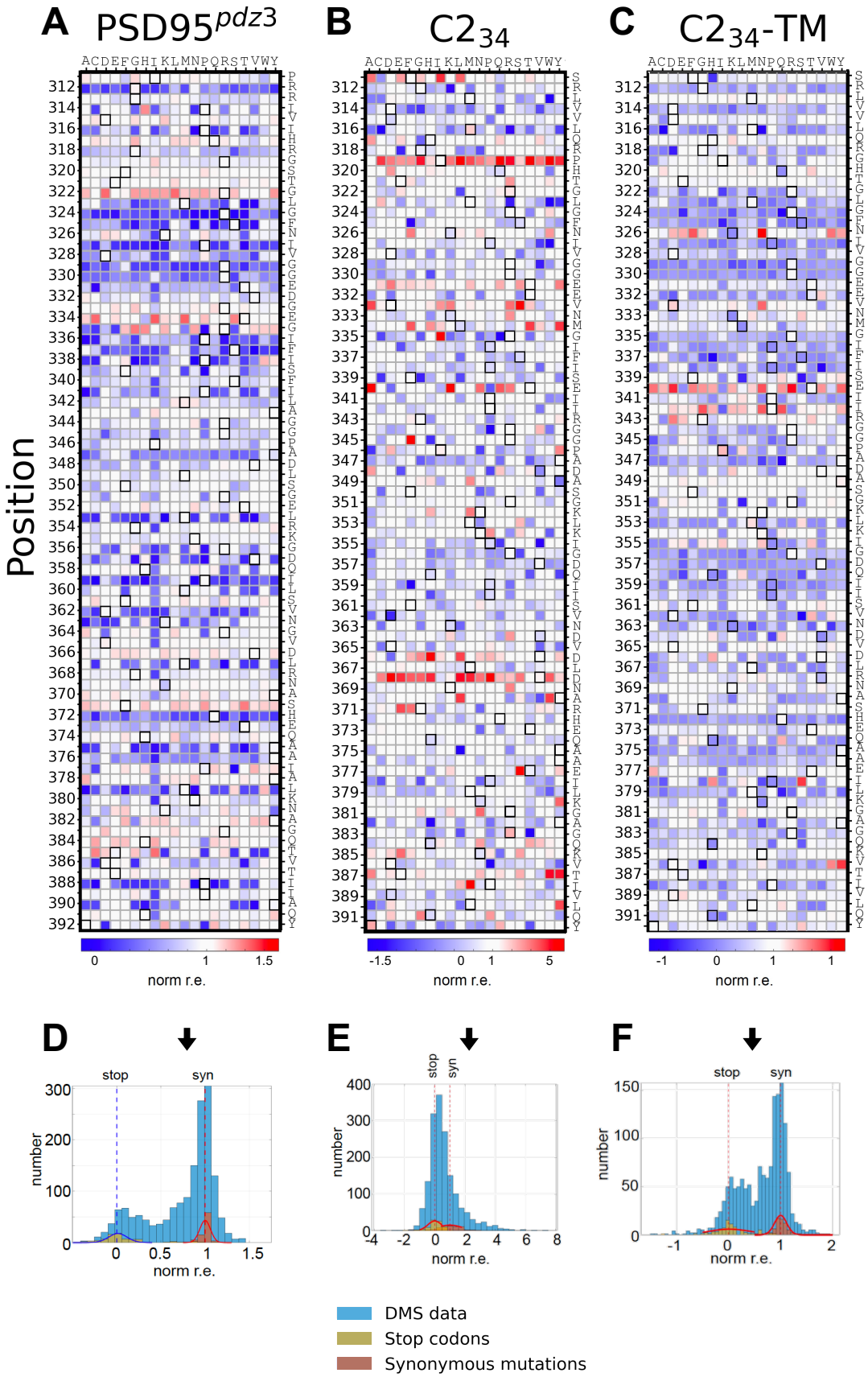
A
PSD95pdz3
B
C2 34
C
C2 34-TM
Position
norm r.e.
D
E
F
stop
syn
number
DMS data
Stop codons
Synonymous mutations

Figure 3.2. DMS for CRIPT ligand binding

A. *DMS of PSD95^{pdz3} for CRIPT ligand binding. Single mutants for each position in the protein, mutated to every other amino acid, are assayed for ligand binding function. The amino acid mutated to is shown at the top of the figure. The position in the protein is shown on the left and the original protein sequence is shown on the right as well as an outlined white box in that row. Function in comparison to the wildtype sequence from low (blue), to neutral (white), to high (red) is shown in the form of normalized relative enrichment (norm r.e.). r.e. is defined as r.e.* $= log_{10}(f_{i,t}/f_{i,0}) - log_{10}(f_{WT,t}/f_{WT,0})$ *where* $f_{i,t}$ *is the frequency of a mutant i at time t and wildtype is PSD95^{pdz3}. Norm r.e. defines 0 as the mean of the stop codon mutations and 1 as the mean of the synonymous mutations. Positions within the DMS which were missing from the assay are shown in grey.*
B. *As in (A) for C2$_{34}$. Note the existence of position 332.5 which is an insertion between positions 332 and 333 for C2$_{34}$ relative to PSD95^{pdz3}.*
C. *As in (A) for C2$_{34}$-TM.*
D. *Histograms of the DMS data from for PSD95^{pdz3} from (A). Stop codons are shown in yellow and synonymous mutations, are shown in red. Norm r.e. defines 0 as the mean of the stop codon mutations and 1 as the mean of the synonymous mutations.*
E. *As in (D) for C2$_{34}$.*
F. *As in (D) for C2$_{34}$-TM.*

Response to mutational stress in a competitive environment

Natural proteins must not only function in their existing environment but must also evolve by through heritable genetic variation in the form of mutations. The ability to maintain a standing pool of mutations for evolution to act upon requires robustness[20]. Robust mutations have no phenotypic effect in a given environment and this property has been correlated with an ability to evolve[27]. For C2$_{34}$-TM, a DMS for CRIPT ligand binding indicated a robustness that is similar to PSD95^{pdz3} (Figure 3.2.A,C). Crucially though, this is only for single mutants to the starting protein. Using BTH-PACE, where propagation of a phage encoding a single PDZ domain depends on binding to the CRIPT ligand, all three proteins were directly competed against each other in the context of mutational stress (Figure 3.3.A). Replicates in BTH-PACE take place in vials called lagoons and each lagoon represents an independent evolutionary trajectory which is inoculated with a population of 10^8 phage and subjected to a mutation rate of 8.7 x 10^5 substitutions per basepair per generation (Figure 2.3.C).

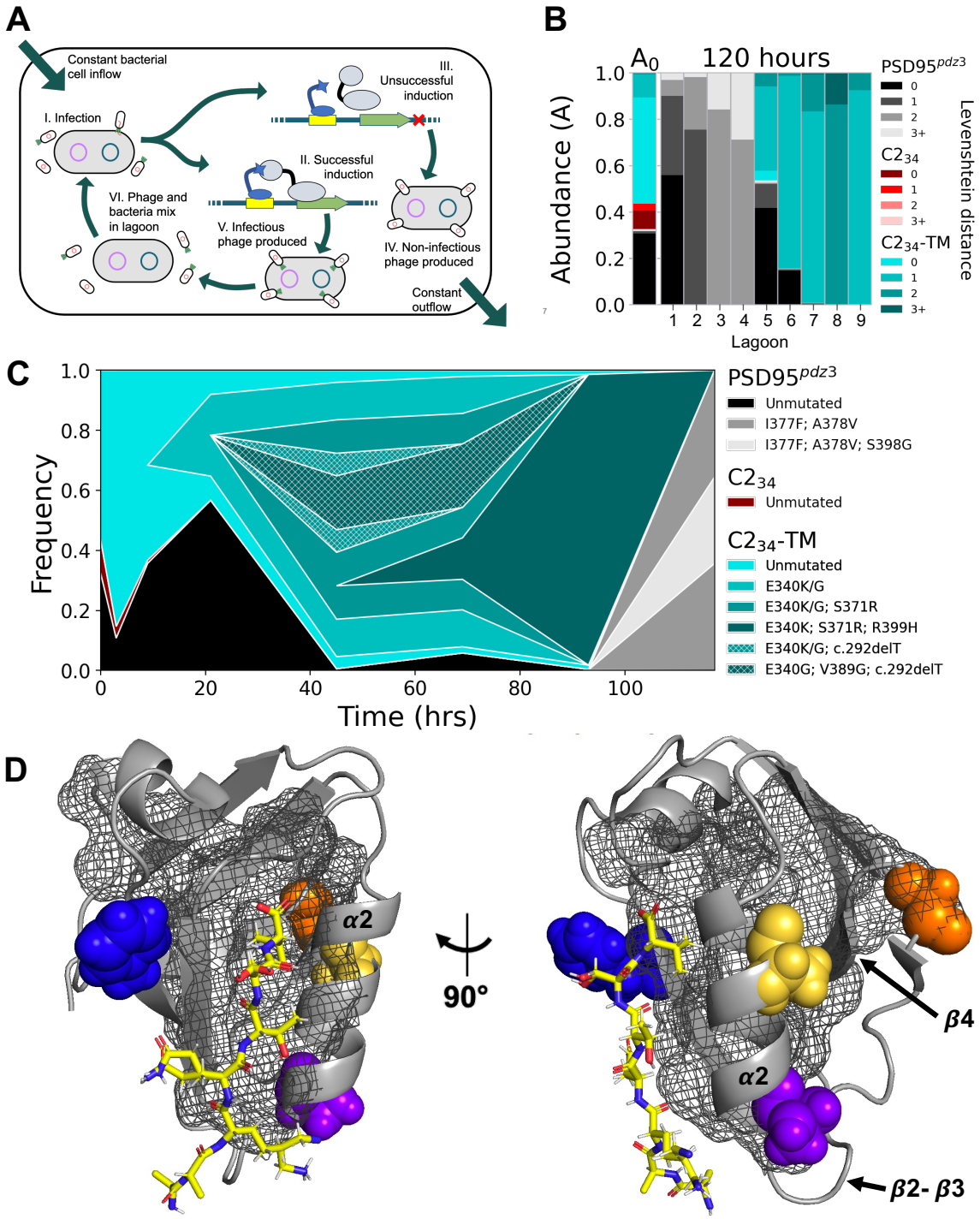

Figure 3.3. $C2_{34}$-TM holds up to mutational stress in BTH-PACE

A. During PACE, evolution occurs in a lagoon with constant bacterial host cell inflow and outflow. (I) M13 phage, lacking essential gIII, encode the evolving PDZ domain fused to RNA Polymerase's ω subunit and infect host cells containing the accessory plasmid (AP, blue) and DP6 plasmid (purple). The AP includes the ligand of interest fused to a DNA binding domain, while the DP6 plasmid controls mutation rate and selection pressure. (II) If the PDZ domain binds the ligand, gIII is induced via a bacterial two-hybrid system. (III) Without binding, gIII is not produced. (IV) Phage without pIII (from gIII) are non-infectious and wash out. (V) Phage with pIII are infectious and (VI) remain in the lagoon long enough to infect new bacterial cells (I).

B. A summary of BTH-PACE evolution experiments shown as a stacked bar plot for each independent evolutionary trajectory (each lagoon). The bars represent lineage frequency after 120 hours of competitive evolution. The first bar on the left shows the initial condition (A_0: 0 generations) common to all trajectories. Lagoon 4 is data from representative example in (C). Black represents $PSD95_{pdz3}$, red is $C2_{34}$, and cyan is $C2_{34}$-TM, with gradations and hatching indicating specific mutants.

C. A representative Muller plot showing evolution and competition between PDZ lineages tasked with binding the CRIPT ligand. Any wedge which exists within another wedge is determined to have mutated from that wedge (ex: I377F; A378V; S98G mutated from I377F; A378V). Colors are as in (B). A slash (ex: E340K/G) indicates two mutations, at the same position, that occurred at a similar rate.

D. Ribbon representation of the crystal structure of $PSD95^{pdz3}$ (PDB ID 5HED) with its native CRIPT ligand (yellow). The sector is shown in grey mesh. Positions 340, 364, 371, and 378 are shown in blue, purple, yellow, and orange respectively.

In a single lagoon, beneficial mutations allow for sweeps of a given protein variant across the population and the dominant lineage can alternate over the course of an evolutionary trajectory (Figure 3.3.C). The first mutation seen in a trajectory, such as E340K and E340G in $C2_{34}$-TM, are seen in the DMS data for CRIPT ligand binding to be beneficial. However, subsequent mutations become harder to predict. In many cases, such as A378V of the $\alpha 2$ helix in $PSD95^{pdz3}$ and S371R in $C2_{34}$-TM, the effect of the single mutation is approximately neutral or even deleterious, indicating the presence of epistasis (Figure 3.3.D). Additionally, some mutations are not possible to assay in a DMS such as the frameshift mutation in $C2_{34}$-TM (c.292delT) which truncates the protein due to a new stop codon, drastically altering the C-terminal tail of the protein. Nonetheless, in the context of the E340K and E340G mutation, this indel mutation is beneficial.

Due to the stochastic nature of any individual evolutionary trajectory, the results of a single lagoon provide little information on the general fitness of a given lineage. However, comparing lagoon results after an arbitrary but consistent time (120 hours) shows that the lineages of $C2_{34}$-TM and $PSD95^{pdz3}$ are equally competitive when asked to bind the CRIPT ligand (Figure 3.3.B). Four out of nine lagoons are seen with majority $PSD95^{pdz3}$ and four out of nine are seen with majority $C2_{34}$-TM. The remaining lagoon

(lagoon 5 in Figure 3.3.B) has a roughly equal proportion of lineage for each protein (53.7% $PSD95^{pdz3}$ versus 46.3% $C2_{34}$-TM).

Additionally, for mutations which are present above a frequency of 1 percent of the total lagoon population, there is no meaningful difference in the percent of mutations which occurred within the sector between the two proteins (9.7% $PSD95^{pdz3}$ versus 6.5% $C2_{34}$-TM). This result holds true for all frequency cutoffs tested, as low as 0.1%, and indicates that constraints placed on the non-sector surroundings, which are scrambled in $C2_{34}$-TM, do not play a major role in native function. In contrast, $C2_{34}$ is quickly depleted in every lagoon (Figure 3.3.B). The depletion of $C2_{34}$ can be explained by its lack of thermal stability and resulting inability to maintain any heritable variation. Proteins require a headroom of thermal stability for mutational robustness (mutations are on average 1-3 kcal/mol destabilizing) and $C2_{34}$ does not have this[28]. As the DMS data for CRIPT ligand binding show, $C2_{34}$ is generally intolerant to mutations and unsurprisingly, unable to adapt (Figure 3.2.B,E).

Differences in mutational tolerance to a class II ligand

To this point, the evolutionary capacity of these PDZ domains has not been tested. As the primary function of $PSD95^{pdz3}$ is presumed to be ligand binding, a logical evolutionary challenge would be to task the three proteins with binding a new ligand[29]. PDZ ligands are C-terminal peptides of target proteins and specificity is primarily determined by the amino acid sequence of the ligand. PDZ domains are often classified by the ligand that they bind and class I ligand binders, like all three proteins in this paper, have preference for C-terminal sequences with a consensus of -X-**S/T**-X-ϕ-COOH (X is any amino acid, ϕ is hydrophobic). However, another class of domains exist, termed class

II ligand binders, which bind ligands with the profile -X-$\underline{\phi}$-X-ϕ-COOH[30]. As the determining factor between class I and class II ligands is the -2 position, a mutation from threonine to phenylalanine at this position, creates a class II ligand. This ligand is referred to as the $T_{-2}F$ ligand (-TKNYKQ**F**SV-COOH). $PSD95^{pdz3}$ binds this ligand with the lower K_D of 17.0 ± 1.1 μM. $C2_{34}$ and $C2_{34}$-TM also have a lower affinity for the $T_{-2}F$ ligand with K_D's of 152 ± 57 μM and 121 ± 15 μM respectively. These binding measurements are a difference between $PSD95^{pdz3}$ and the two artificial proteins. Whereas $PSD95^{pdz3}$ binds the $T_{-2}F$ ligand only 37.6 times worse than the CRIPT ligand, $C2_{34}$ and $C2_{34}$-TM do so 128 and 120 times worse respectively.

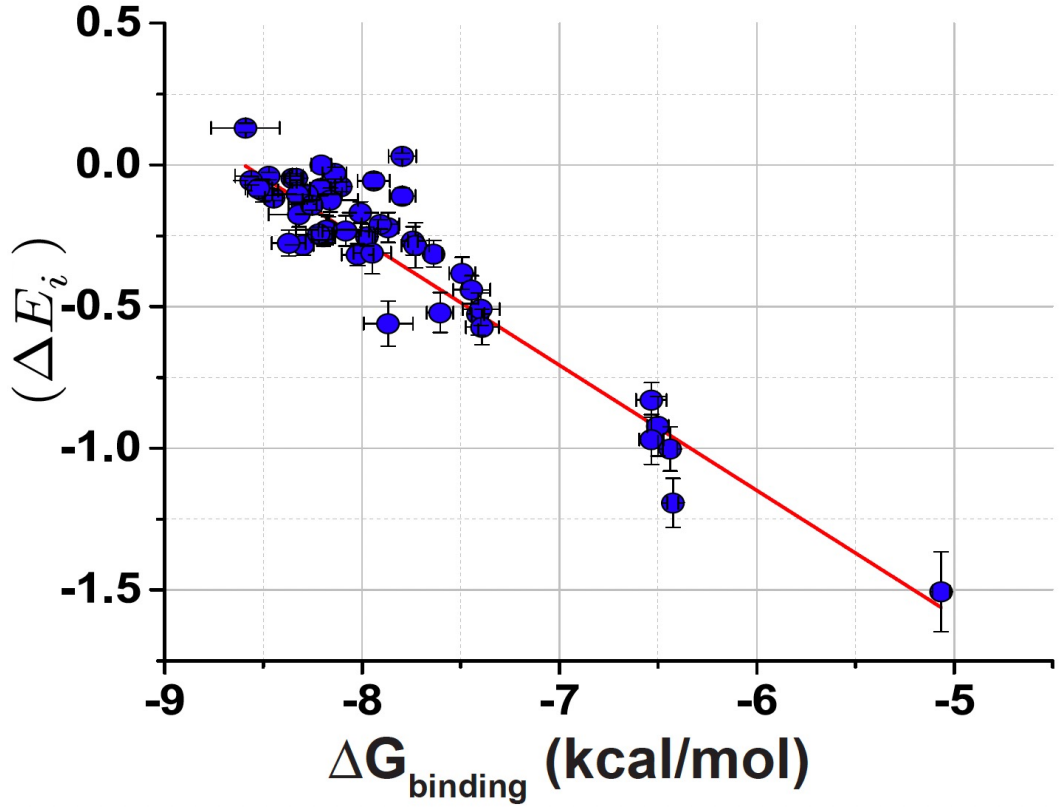

Figure 3.4. The dynamic range of the bacterial-two-hybrid, PDZ-ligand binding assays
A library of PDZ mutants with known binding energies. ΔE_i is the enrichment of PDZ variant i as defined by the natural log of the relative change in its frequency from the start of the PDZ-binding based assay to the end. Binding energies of -8.7 to -5.1 kcal/mol are tested corresponding to a K_D range of ~0.4 μM to ~183 μM. This binding assay is what is used to generate data in all DMS experiments.

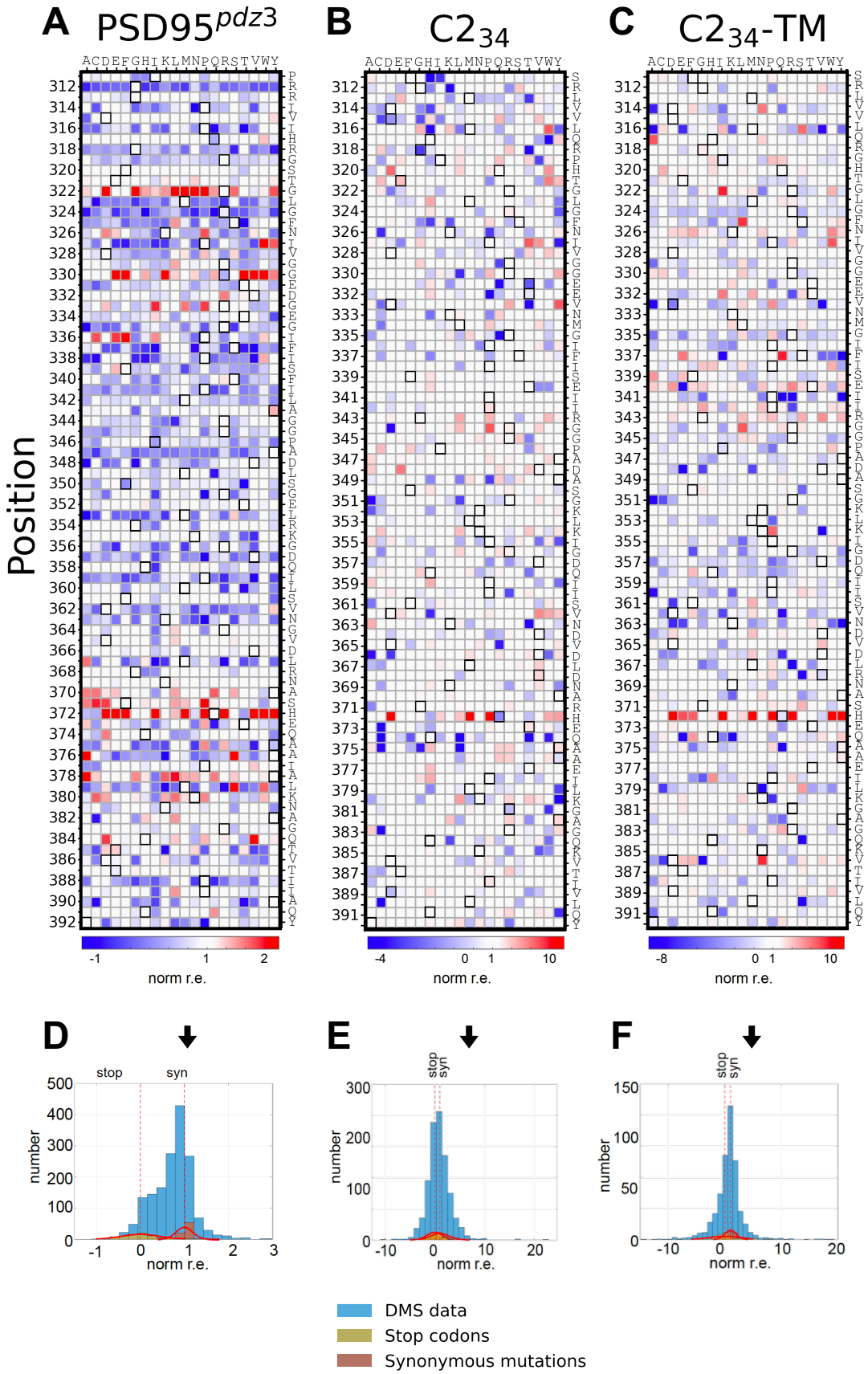
A
PSD95pdz3
B
C234
C
C234-TM
Position
norm r.e.
D
E
F
stop
syn
number
DMS data
Stop codons
Synonymous mutations

Figure 3.5. DMS for $T_{-2}F$ ligand binding

A. *DMS of $PSD95^{pdz}$ for $T_{-2}F$ ligand binding as in Figure 3.2. The position in the protein is shown on the left and the original protein sequence is shown on the right as well as a white box in that row. Function from low (blue), to neutral (white), to high (red) is shown in the form of norm r.e. Positions within the DMS which were missing from the assay are shown in grey.*
B. *As in (A) for $C2_{34}$. Note the existence of position 332.5 which is an insertion between positions 332 and 333 for $C2_{34}$ relative to $PSD95^{pdz3}$.*
C. *As in (A) for $C2_{34}$-TM.*
D. *Histograms of the DMS data from for $PSD95^{pdz3}$ from (A). Stop codons are shown in yellow and synonymous mutations, are shown in red. Norm r.e. defines 0 as the mean of the stop codon mutations and 1 as the mean of the synonymous mutations.*
E. *As in (D) for $C2_{34}$.*
F. *As in (D) for $C2_{34}$-TM.*

Ligand binding experiments, with a dynamic range of at least 0.4 μM to 183 μM, were used to test DMS libraries of each protein against the $T_{-2}F$ ligand. These experiments show stark mutational sensitivity differences when compared to the DMS binding experiments conducted against the CRIPT ligand (Compare Figure 3.2 to Figure 3.5). $PSD95^{pdz3}$ shows the least difference qualitatively. Although it is collapsed, there is still a bimodal distribution of mutational effects with most mutations clustering in a peak centered near neutrality with respect to $T_{-2}F$ ligand binding function (Figure 3.5.D). Additionally, the second peak, which includes a smaller number of mutations than the near-neutral peak, contains deleterious mutations. Unlike the $PSD95^{pdz3}$-CRIPT DMS though, there is a relatively large tail of mutations which improve function for binding the $T_{-2}F$ ligand. This makes sense as the unmutated protein initially has a poor ability to bind the $T_{-2}F$ ligand. For $C2_{34}$ and $C2_{34}$-TM the story is different (Figure 3.5.E-F). For both proteins the distribution of fitness effects is now unimodal with most mutations being approximately neutral for $T_{-2}F$ ligand binding. There is a small tail of mutations which are gain-of-function, exclusively corresponding to mutations at position 372 but this number is greatly decreased relative to $PSD95^{pdz3}$.

As PDZ domains, particularly PSD95^{pdz3}, are a well-studied system, the structural effect of single mutants is known in a variety of cases[26]. The previously mentioned position 372, which is in the $\alpha 2$ helix of the binding pocket and contains a universally conserved histidine, directly contacts the threonine hydroxyl at the -2 position in the CRIPT ligand through a hydrogen bond (Figure 3.6.A). This histidine is therefore thought to be crucial to the protein's specificity for class I ligands. Unsurprisingly, mutations at this position in PSD95^{pdz3}, such as H372A, have been shown to cause class switching behavior by eliminating the hydrogen bond interactions with the CRIPT ligand and removing the steric clash experienced in the unmutated protein when the phenylalanine of the $T_{-2}F$ ligand is bound in the active site (Figure 3.6.C-D). In general, mutations to position 372 can directly switch the ligand specificity of a PDZ domain from class I to class II ligands through a local perturbation. Consistent with this, results from the DMS ligand binding assay show all three proteins have gain-of-function mutations at position 372, allowing for improved binding to the $T_{-2}F$ ligand at the cost of decreased CRIPT ligand binding (Figure 3.2.A-C, Figure 3.5.

Another position where mutations to the natural protein are known to improve binding to the $T_{-2}F$ ligand is 330[26]. The mechanism for this phenomenon is entirely different than position 372 as position 330 is located on a surface loop between the $\beta 2$ and $\beta 3$ strands and does not contact the active site. Binding of the $T_{-2}F$ ligand, causes the $\beta 2$-$\beta 3$ loop to partially adopt an alternative confirmation which is required to stabilize the new rotamer at site 372 that is forced by the steric clash of the phenylalanine (330_2, Figure 3.6.B). Mutations at site 330, such as G330T stabilize the 330_2 confirmation. Importantly, 330_2 does not preclude the original rotamer state of position 372 and

therefore mutations at position 330 are conditionally neutral (Figure 3.6.E-F). Only PSD95^{pdz3} has conditionally neutral mutants for T$_{-2}$F ligand binding at position 330. Furthermore, only PSD95^{pdz3} has significant adaptive mutants of any kind at any position other than 372 in DMS studies.

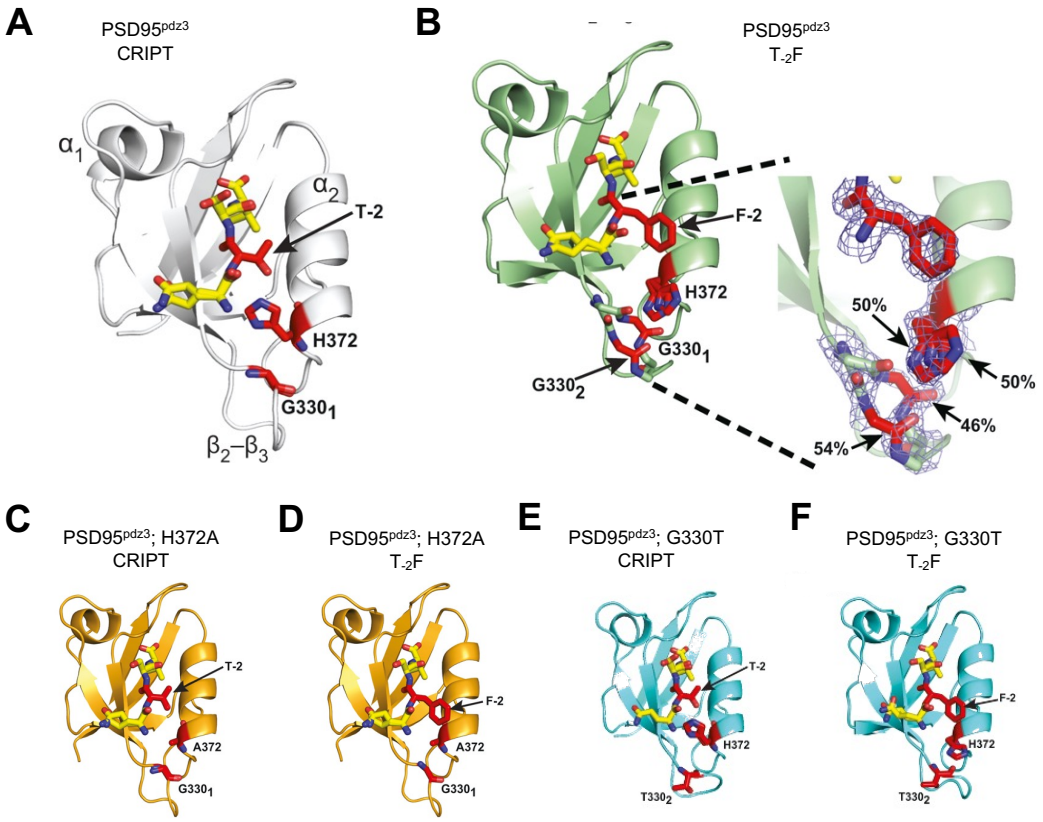

Figure 3.6. Mutations at position 330 and 372
Adapted from *Raman AS, et al, 2016*[26]

A. The PSD95^{pdz3} bound to CRIPT (PDB ID 5HEB) structure illustrates key aspects of class I ligand recognition. The hydroxyl group of threonine at position -2 in the ligand forms a hydrogen bond with histidine at position 372. The glycine at position 330 is situated within a well-ordered β2-β3 loop near His$_{372}$ but is not part of the active site (330$_1$, where the subscript represents conformation 1).
B. When T$_{-2}$F binds to wild-type PSD95^{pdz3} (PDB ID 5HED), a new partially occupied conformation of the β2–β3 loop (330$_2$) is generated. This conformation allows His$_{372}$ to accommodate a non-native rotamer which avoids steric interference.
C. Structures of PSD95^{pdz3}; H372A bound to CRIPT (PDB ID 5HFB) show truncation of the side chain at position 372 with minimal other structural changes. The loss of both size and hydrogen bonding at position 372 corresponds with the decreased ability to bind the CRIPT ligand observed in the DMS.
D. Same as in (C) for the T$_{-2}$F ligand bound to PSD95^{pdz3}; H372A (PDB ID 5HFC). The local adjustment introduced by the H372A mutation accommodates the bulky phenylalanine side chain at position -2.
E. The G330T mutation stabilizes 330$_2$ in the β2–β3 loop which allows histidine 372 to occupy either the native or a non-native rotamer depending on the ligand present. Consequently, in PSD95^{pdz3}; G330T, His$_{372}$ assumes the native rotamer when bound to CRIPT ligand (PDB ID 5HEY)
F. Same as in (E) for PSD95^{pdz3}; G330T bound to the T$_{-2}$F ligand (PDB ID 5HF1). His$_{372}$ adopts a non-native rotamer.

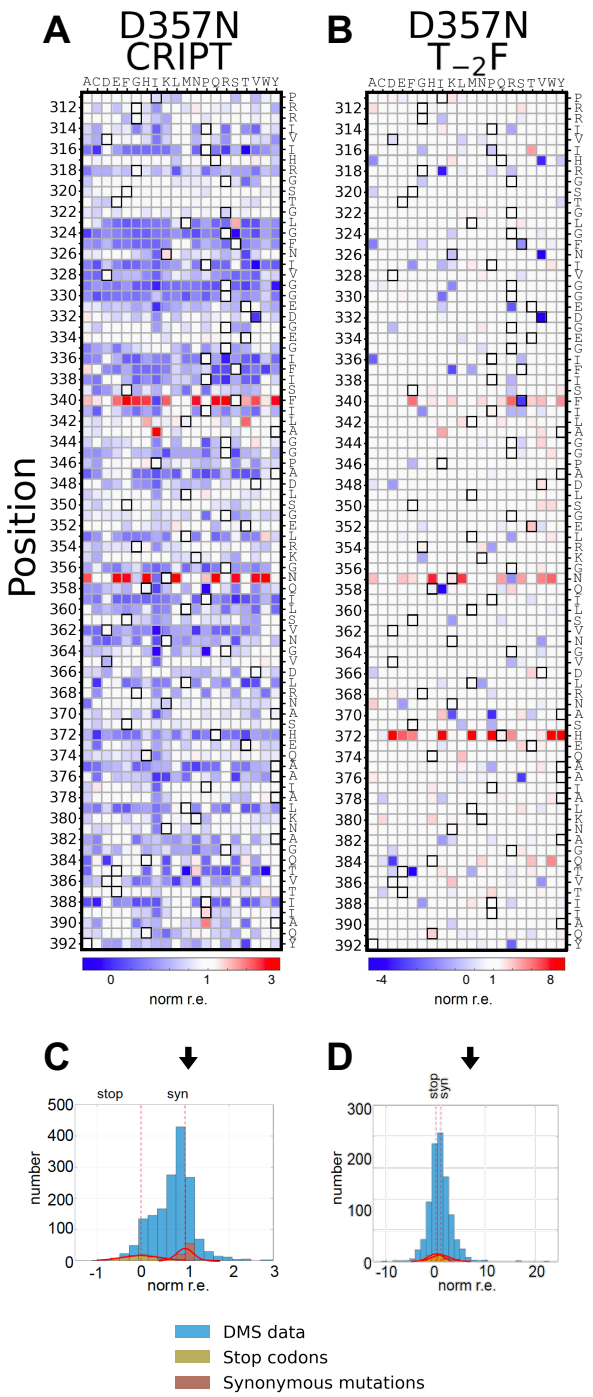

Figure 3.7. DMS of D357N binding to CRIPT and T-2F ligands

A. DMS of D357N for CRIPT ligand binding as in Figure 3.2. The position in the protein is shown on the left and the original protein sequence is shown on the right as well as a white box in that row. Function from low (blue), to neutral (white), to high (red) is shown in the form of norm r.e. Positions within the DMS which were missing from the assay are shown in grey.

B. As in (A) for D357N binding to the T-2F ligand.

C. Histograms of the DMS data from for D357N from (A). Stop codons are shown in yellow and synonymous mutations, are shown in red. Norm r.e. defines 0 as the mean of the stop codon mutations and 1 as the mean of the synonymous mutations.

D. As in (D) for D357N binding the T-2F ligand.

A destabilized version of PSD95^{pdz3}

Despite C2$_{34}$-TM's functionality, its decreased thermal stability compared to PSD95^{pdz3} could be the reason for its lack of CN and a predicted lowered exaptive capacity. There is a body of work linking evolvability – specifically robustness and exaptation – to protein stability, and, while that work does not make the claim that stability is the only factor, this effect must be examined in C2$_{34}$-TM[15,20,27,31]. An ideal test case would involve a mutant version of PSD95^{pdz3} which differs only in its lowered thermal stability.

Previous work has quantified an array of biophysical parameters for a library of single point mutants to 83 positions in PSD95^{pdz3}. The point mutants that were chosen to represent the next most common amino acid in the PDZ alignment at each position and therefore should result in a protein which behaves similarly to PSD95^{pdz3}[22]. One specific mutant, PSD95^{pdz3}; D357N, referred to as simply D357N, has the property of reducing the thermal stability of the wildtype protein by 15°C (T_m = 53°C) while still retaining CRIPT ligand binding (K_D = 1.63 μM). This thermal stability is similar to that of C2$_{34}$-TM (58°C), and when a DMS is conducted for CRIPT ligand binding the results are again qualitatively and quantitatively similar (Figure 3.7.A,C, Figure 3.2). The only slight difference is a shifting of the roughly neutral peak to just left of neutrality, indicating the average mutation is slightly deleterious compared to wildtype PSD95^{pdz3}, a fact that could be predicted from its lower thermal stability.

Unlike C2$_{34}$-TM, D357N binds the T_{-2}F ligand only 27.2 times worse than the CRIPT ligand (K_D = 65.6 ± 7.0 μM). This number is actually a slight improvement on the 37.6x decrease for PSD95^{pdz3} and significantly better than for C2$_{34}$-TM's 120x decrease.

A DMS for D357N binding to the class II ligand initially looks similar to $C2_{34}$-TM, in that its histogram of REs has a unimodal peak centered around neutrality (Figure 3.7.B,D, Figure 3.5). However, this visualization hides some key differences that become apparent when only the mean of the top five RE values are plotted for each position (Figure 3.8). First, like all other proteins tested, D357N, sees the strongest adaptation at site 372. The reasons for this are the same as those previously discussed and like all other proteins, these mutations are direct switching for the experimental conditions of the DMS assay.

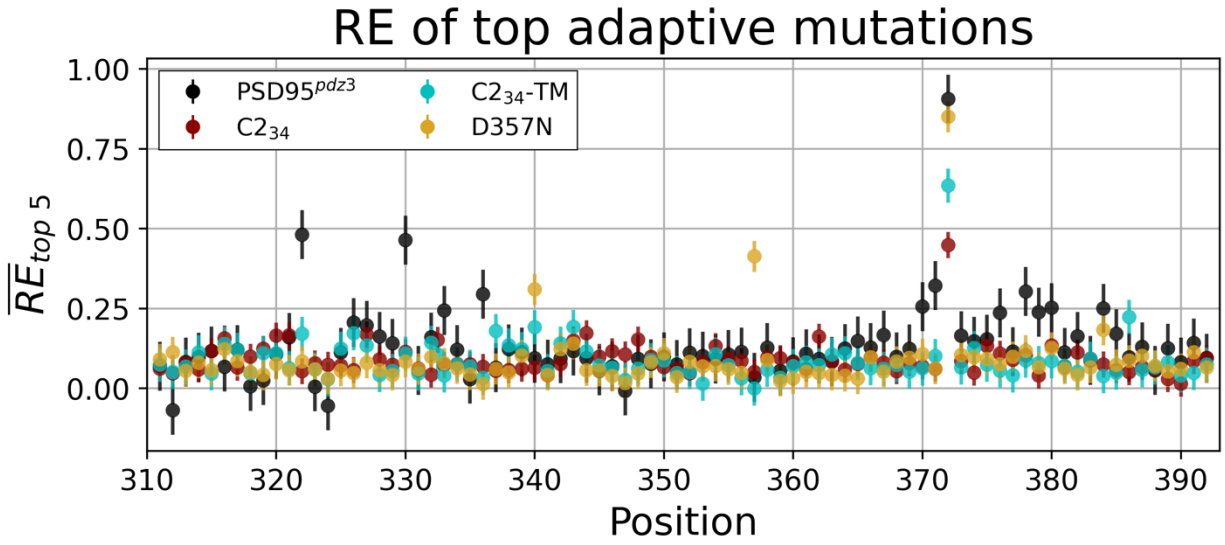

Figure 3.8. Site specific improvements to $T_{-2}F$ ligand binding
The mean and standard error of the top 5 RE scores for each protein (color coded dots) are plotted for each protein at each position. Analyzing different numbers, N, of top RE scores does not qualitatively change the results as long as N is less than 10. Beyond 10, for a site with only a few mutants which provide $T_{-2}F$ function, signal can be hard to distinguish from noise.

Position 357 also appears as a hot spot for improvement to $T_{-2}F$ ligand binding. This is not surprising as a mutation at this position is what initially destabilized $PSD95^{pdz3}$ and therefore mutations here are likely reverting the stability back to wildtype levels. Although no single mutation will similarly revert $C2_{34}$-TM to a natural protein, if, all that was needed to allow binding to the $T_{-2}F$ ligand was an increase in thermal stability, it expected that positions would exist in the $C2_{34}$-TM DMS that show a similar signal as

position 357 does in the D357N DMS. This simple fact argues that something further is wrong with $C2_{34}$-TM, likely stemming from the scrambling of its non-sector surroundings.

An even stronger argument exists at site 340 though. Site 340 is far from the active site and crucially, is not the position mutated to make D357N (Figure 3.3.D). This position allows for $T_{-2}F$ ligand binding while also increasing CRIPT ligand binding meaning it is CN (Figure 3.7.A-B). It is again entirely possible, that mutations at position 340, in the background of D357N, only function to increase stability but that further proves the point. Even if additional stability is all that is needed for D357N to bind to a class II ligand and generate CN, this same path does not exist for $C2_{34}$-TM. It is entirely unable to adapt in any capacity other than a local perturbation to the active site.

Assay for productive SGV in natural and synthetic proteins

As it has been previously shown that mutations to the active site are likely to be direct switching for all but the weakest selection pressures, a logical hypothesis from the previous data is that $C2_{34}$-TM will be deficient in maintaining the CN portion of its productive SGV. That is, it will not be able to maintain the exaptive CN mutants required to survive under conditions of changing selection. BTH-PACE provides a way to test this assertion. If proteins are evolved in BTH-PACE while requiring binding to the CRIPT ligand for a period long enough to generate a pool of SGV, then those proteins which can maintain CN mutants will have a larger fraction of their population preadapted for $T_{-2}F$ ligand binding. This is a statistical argument derived from the fact that for a given selection pressure, direct switching mutants are more likely to depleted within a population due to purifying selection[26].

Experiments, with and without selection for CRIPT ligand binding, were carried out and this is exactly what was found in a preliminary analysis (Figure 3.9). As expected PSD95^{pdz3} maintained productive SGV with and without selection. The ratio of the frequencies between the two conditions is expected reflect selection against direct switching mutants. Conversely, both $C2_{34}$ and $C2_{34}$-TM struggled to maintain any $T_{-2}F$ functional mutants, even without selection, and were completely unable to do so when selection for CRIPT ligand binding was included in BTH-PACE. Finally, D357N was able to maintain productive SGV in BTH-PACE experiments with selection for CRIPT ligand binding, though not as well as PSD95^{pdz3}. These results suggest that the defect in $C2_{34}$ and $C2_{34}$-TM goes beyond a decrease in thermal stability.

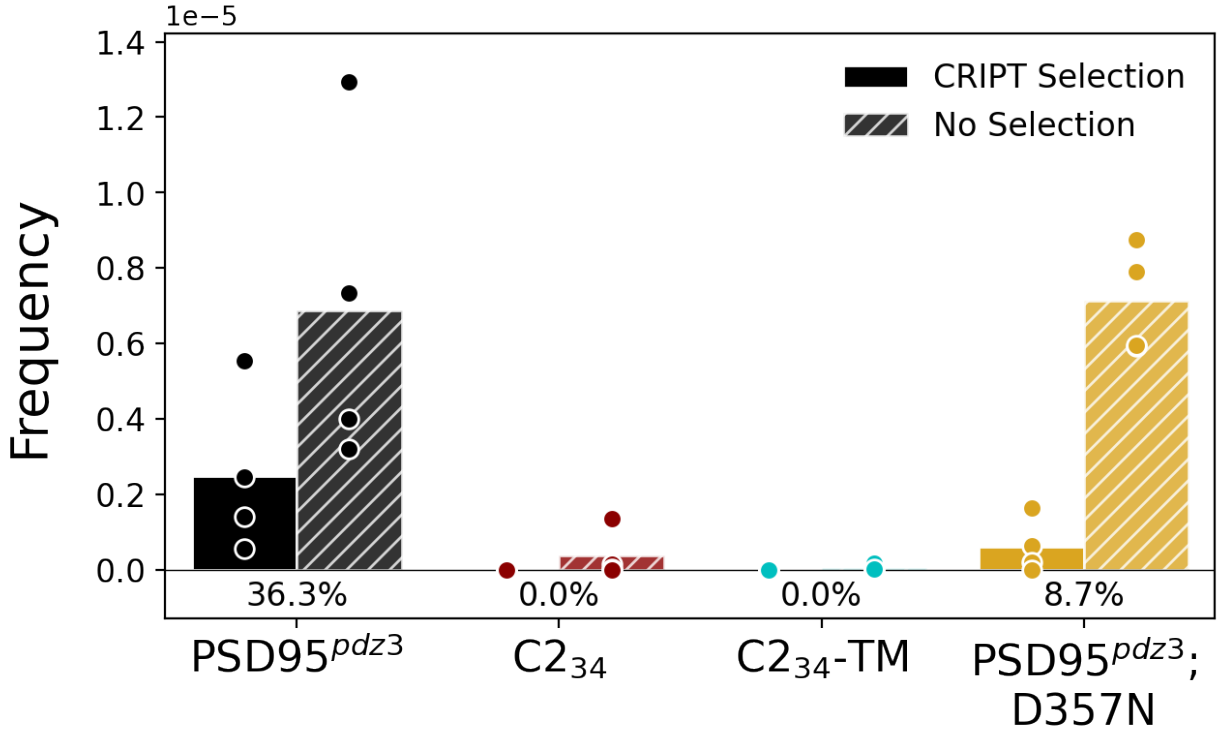

Figure 3.9. BTH-PACE assay for CN mutants
The frequency $T_{-2}F$ functional PDZ variants in a lagoon for 4 replicates (circles, bar is mean of replicates) of BTH-PACE with and without selection for CRIPT ligand binding. The percentage of $T_{-2}F$ functional mutants found in the CRIPT selection environment relative to the non-selection environment is reported as a percentage underneath the bars.

This last result can be improved upon in future work. The reporter for $T_{-2}F$ functional mutants is a plaque assay where only phage containing a PDZ domain which binds the $T_{-2}F$ ligand form plaques. The level of binding needed to form plaques is 3-5 μM and therefore only exceptionally good binders of the $T_{-2}F$ ligand are reported on. This tight binding places an incredibly high bar for $T_{-2}F$ ligand binding function; a bar that likely precludes many physiologically relevant mutations from being visualized. Sanger sequencing of individual plaques confirm this, indicating that variants detected via this assay are predominantly at position 372. A simple future experiment is planned to probe the results of the BTH-PACE experiments more thoroughly. Deep sequencing of the BTH-PACE samples before and after growth dependent on $T_{-2}F$ ligand binding will reveal the fraction of mutants in the population with $T_{-2}F$ ligand binding function regardless of binding affinity.

3.4. Discussion

Prior work has shown proteins to be inherently heterogeneous, including, but not limited to, a biphasic intramolecular structure[1]. Understanding this intrinsic heterogeneity is essential to explaining the biophysical properties and functional versatility of proteins. We exaggerate this heterogeneity by retaining only the constraints of collective epistasis defined by the sector and scrambling the rest of the protein. Previous work has shown the sector to be necessary to protein folding, function, and to contain the sites of CN mutations, but experiments testing the sufficiency of sectors have not been conducted.

Here, using a designed protein, $C2_{34}$, which has had its non-sector surroundings scrambled, we show that $C2_{34}$ can bind the native, class I ligand with natural like affinity

and specificity. Furthermore, introducing three stabilizing mutations to the non-sector region produces a protein with almost wildtype levels of thermal stability. This stabilized protein, $C2_{34}$-TM, is robust to mutation for native CRIPT ligand binding and is equally competitive when tasked with improving binding to this ligand in BTH-PACE. Up to this point, $C2_{34}$-TM appears broadly indistinguishable from the natural protein $PSD95^{pdz3}$. However, once a new functional challenge is introduced – binding the class II ligand $T_{-2}F$ – an intrinsic deficiency of $C2_{34}$-TM becomes apparent. $C2_{34}$-TM has no initial ability to bind this alternative ligand, whereas $PSD95^{pdz3}$, can do so with affinities that approach physiological relevance. Additionally, while the natural protein has no problems maintaining conditionally neutral mutants for $T_{-2}F$ ligand binding, $C2_{34}$-TM is generally unable to do so. It is only capable of local adaptation at position 372 in the active site which switches ligand specificity from CRIPT to $T_{-2}F$.

A model for the non-sector surroundings

A reasonable explanation for $C2_{34}$-TM's impaired ability to adapt to a novel ligand is the fact that its non-sector surroundings have been scrambled by the design process. We propose a conceptual model to explain this phenomenon by framing protein folding as a high-dimensional landscape, often visualized as a funnel (Figure 3.10). In this landscape, natural proteins reside at the bottom of an energy well, representing a stable, low-energy state. However, the landscape is not smooth, and other local minima also exist, allowing for the possibility of alternative conformations. When a protein binds its native ligand, it triggers the activation of an allosteric network that leads to a new energy minimum, consistent with the modeling work of Rouviere *et al*, 2023[32]. The distinction between an allosteric network and an epistatic network is essential here. While epistatic

networks are a statistical phenomenon defined by a collective cooperativity in mutational effects, allosteric networks are a physical phenomenon. They are the paths of energy propagation upon perturbation at the active site. Although the reason for the presence of an epistatic network is often the existence of an allosteric network at overlapping or nearby positions, they are not the same thing.

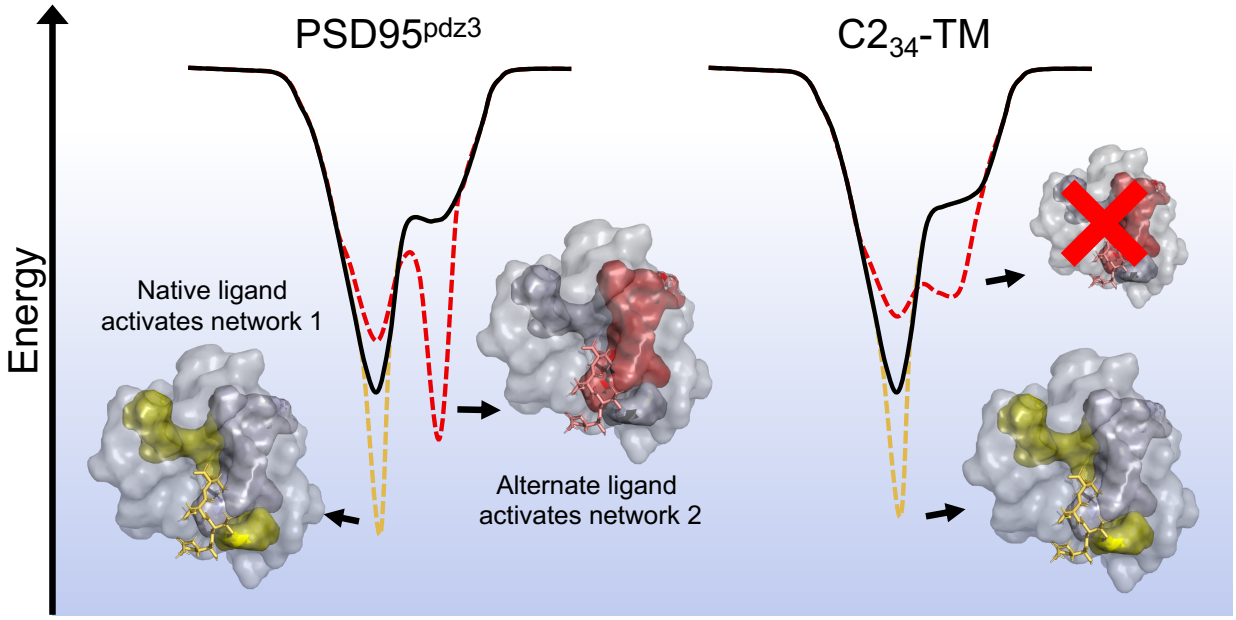

Figure 3.10. Non-sector surroundings are essential for conditional neutrality
Natural proteins, not bound to any ligand, exist in a stable native state conformation at the bottom of an energy well (left, black line). Upon binding their native ligand an allosteric network is activated (network 1) and a new local energy minimum is found (left, yellow, dashed line). When tasked with binding an alternative ligand, a second allosteric network (network 2) can be activated resulting in an alternate conformation distinct from when the native ligand is bound. Although it might not be as stable as for the native ligand, this also creates a new energy minimum, (left, red, dashed line). For proteins lacking constraints in their surroundings, the unbound (right, black, line) and native ligand bound states (right, yellow, dashed line) may look similar to the natural protein. However, they are incapable of utilizing an allosteric network second and conformational state required for binding a sufficiently different alternative ligand (right, red, dashed line).

Importantly, the native ligand is an integral part of both networks, as supported by unpublished data from the Ranganathan lab, and therefore alternative ligands might not efficiently engage the allosteric network. However, in natural proteins, when an alternative ligand binds, we propose that there is the possibility of shifting the equilibrium towards

another energy well, activating a distinct allosteric network. Preliminary data from the Ranganathan lab also supports this assertion. Importantly, this alternative ligand does not necessarily activate the original allosteric network, essentially rerouting the protein's functional pathways. Proteins like $C2_{34}$-TM, which have lost the constraints on their non-sector surroundings, can still activate the allosteric network for the native ligand but appear to be unable to engage any alternative allosteric networks. This restricts their ability to evolve, as they are limited to local alterations. Local alterations are unlikely to be CN and therefore their ability to maintain productive SGV is diminished.

This work is consistent with the constraints on the non-sector surroundings being essential for conditional neutrality – the statistically preferred method of exaptation. The sector itself defines the canonical solution for function (e.g., ligand binding), but interactions with the surrounding regions appear to allow proteins to "riff" off this primary function and explore a degeneracy of solutions in response to a new functional challenge. Interactions between the sector and its surroundings allow a protein to evolve in a conditionally neutral mediated way.

3.5. References

1 Lockless SW & Ranganathan R. Evolutionarily conserved pathways of energetic connectivity in protein families. *Science* **286,** 295-299, (1999).

2 Süel GM, *et al*. Evolutionarily conserved networks of residues mediate allosteric communication in proteins. *Nature Structural and Molecular Biology* **10**, 59-69, (2003).

3 Halabi N, *et al*. Protein sectors: evolutionary units of three-dimensional structure. *Cell* **138,** 774-786, (2009).

4 Weigt M, *et al.* Identification of direct residue contacts in protein-protein interaction by message passing. *PNAS* **106**, 67-72, (2009).

5 Morcos N, *et al.* Direct-coupling analysis of residue coevolution captures native contacts across many protein families. *PNAS Plus* **108**, E1293-E1301, (2011).

6 Payne JL & Wagner A. The causes of evolvability and their evolution. *Nature Reviews Genetics* **20**, 24-38 (2019)

7 Wagner A. Evolvability-enhancing mutations in the fitness landscapes of an RNA and a protein. *Nature Communications* **14**, 3624, (2023).

8 Paaby AB & Rockman MV. Cryptic genetic variation: evolution's hidden substrate. *Nature Reviews Genetics* **15** 247-258, (2014).

9 Luria SE & Delbrück M. Mutations of bacteria from virus sensitivity to virus resistance. *Genetics* **28**, 491-511, (1943).

10 Gould SJ & Vrba ES. Exaptation-A missing term in the science of form. *Paleobiology* **8**, 4-15, (1982).

11 Gould SJ. The exaptive excellence of spandrels as a term and prototype. *PNAS* **94**, 10750-10755, (1997).

12 Frenkel-Pinter M, *et al.* Adaptation and exaptation: from small molecules to feathers. *Journal of Molecular Evolution* **90**, 166-175, (2022).

13 Dellus-Gur E, *et al*. What makes a protein fold amendable to functional innovation? Fold polarity and stability trade-offs. *Journal of Molecular Biology* **425**, 2609-2621, (2013).

14 Draghi JA, *et al.* Epistasis Increases the Rate of Conditionally Neutral Substitution in an Adapting Population. *Genetics* **187**, 1139-1152, (2011).

15 Brosius J. Exaptation at the molecular genetic level. *Science China Life Sciences* **61**, 437-452, (2019).

16 Draghi JA & Plotkin JB. Hidden diversity sparks adaptation. *Nature* **474**, 45-46, (2011).

17 Hayden EJ, *et al*. Cryptic genetic variation promotes rapid evolutionary adaptation in an RNA enzyme. *Nature* **474**, 92-95, (2011).

18 Barve A & Wagner A. A latent capacity for evolutionary innovation through exaptation in metabolic systems. *Nature* **500**, 203-206, (2013).

19 Anderson JT, *et al*. Genetic trade-offs and conditional neutrality contribute to local adaptation. *Molecular Ecology* **22**, 699-708, (2013).

20 Wagner A. Robustness, evolvability, and neutrality. *FEBS Letters* **579**, 1873-3468, (2015).

21 Paaby AB & Rockman MV. Cryptic genetic variation: evolution's hidden substrate. *Nature Reviews Genetics* **15** 247-258, (2014).

22 McLaughlin RN, *et al*. The spatial architecture of protein function and adaptation. *Nature* **491**, 138-142, (2012).

23 Rivoire O, *et al.* Evolution-Based functional decomposition of proteins. *PLoS Computational Biology* **12**, e1004817, (2016).

24 Socolich M, *et al*. Evolutionary information for specifying a protein fold. *Nature* **437**, 512-518, (2005).

25 Russ WP, *et al.* Natural-like function in artificial WW domains. *Nature* **437**, 579-583, (2005).

26 Raman AS, *et al*. Origins of allostery and evolvability in proteins: A case study. *Cell* **166**, 468-480, (2016).

27 Bloom JD, *et al*. Protein stability promotes evolvability. *PNAS* **103**, 5869-5874, (2006).

28 Pandey P, *et al*. Predicting the effect of single mutations on protein stability and binding with respect to types of *mutations. International Journal of Molecular Sciences* **24**, 12073, (2023).

29 Han K & Kim E. Synaptic adhesion molecules and PSD-95. *Progress in Neurobiology* **84**, 263-268, (2008).

30 Lee HJ & Zheng JJ. PDZ domains and their binding partners: structure, specificity, and modification. *Cell Communication and Signaling* **8**, 8, (2010).

31 Tokuriki N & Tawfik DS. Stability effects of mutations and protein evolvability. *Current Opinion in Structural Biology* **19**, 596-604, (2009).

32 Rouviere E, *et al*. Emergence of single- versus multi-state allostery. *PRX Life* **1**, 023004, (2023).

Chapter 4. The impact of evolutionary history on conditional neutrality

4.1. Abstract

Prior simulations of evolving proteins suggest that the unique evolutionary history of a protein, particularly the rate at which environmental conditions fluctuate, plays a critical role in defining its allosteric network and, consequently, its ability to adapt to new selection pressures. To investigate this relationship further, we evolved six class I PDZ domains using the bacterial two-hybrid phage-assisted continuous evolution (BTH-PACE) system under varying switching rates of selection pressure. In the previous chapter, we demonstrated that conditional neutrality serves as a readout for changes in protein architecture and preliminary results reveal that proteins exposed to intermediate rates of environmental change exhibit a distinct increase in their ability to harbor conditional neutrality. These findings provide experimental evidence supporting the hypothesis that a protein's architecture is intricately linked to its evolutionary history.

4.2. Introduction

In the previous chapter, it was demonstrated that algorithmically designed ligand-binding proteins, which contain only the sector-defining constraints ($C2_{34}$ and $C2_{34}$-TM), were impaired in their ability to adapt to the class-switching ligand $T_{-2}F$. This is despite both $C2_{34}$ and $C2_{34}$-TM maintaining their natural-like function for binding to their native ligand which is consistent with non-sector positions (the surrounding regions) playing a role in protein evolution. The proposed function of these surrounding regions is that they are essential for activating alternative allosteric networks located within and along the sector positions of the PDZ domain. Absent the ability to activate these allosteric

networks, distant perturbations, in the form of mutations, are hypothesized to have no effect on the active site, making the maintenance of conditionally neutral (CN) mutants impossible.

A straightforward yet unsurprising conclusion from Chapter 3 is that altering the constraints on a protein, as was done with $C2_{34}$ and $C2_{34}$-TM, leads to a changed ability to support CN. This suggests that CN can be viewed as a readout for the global constraints, which are shaped by evolution, that define a natural protein. From this work, a hypothesis emerges: a protein's architecture, defined by its statistical constraints and structure, is a product of its unique evolutionary history. This hypothesis is not new; previous simulations in a simple Ising network model show that the size of the allosteric network depends on the mutation rate of the system and the fluctuation rate between environmental conditions[1-2]. Although this conclusion has not been experimentally tested, if validated in natural systems, it would support the hypothesis that a protein's allosteric network changes as a function of its selection pressure history. Moreover, since the ability to adapt (or exapt) to new selection conditions is a property of the allosteric network, measures of adaptation should also be influenced by the rate of change in selection pressures, serving as a readout for the evolving architecture of the protein[3].

An ideal system for testing this hypothesis experimentally would allow for control over mutation rates, fluctuating selection pressures, and population size in an evolving system. Thanks to advancements made in the bacterial two-hybrid phage-assisted continuous evolution (BTH-PACE) system, all of this is now possible (see chapter 2). In this study, six different class I PDZ domains were placed into phage and evolved under three distinct rates of switching between selection for class I and class II ligand binding:

constant selection for the class I ligand binding only, slow switching, and intermediate switching. The phage populations were evolved for at least 504 generations, with the PDZ proteins in each population accumulating up to seven mutations from the initial wild-type PDZ domain. While work is still ongoing, early results suggest a potential difference in the capacity to adapt between proteins evolved under intermediate selection fluctuations and those evolved under constant selection. If confirmed, this would represent the first experimental validation of the long-standing hypothesis that the constraints observed in present-day proteins depend on the specifics of their evolutionary history.

4.3. Results

Characterization of PDZ domains in BTH-PACE

The model system for these evolutionary trajectories is the PDZ domain, a small ligand-binding protein that interacts with C-terminal ligands with affinities in the range of 1 μM[4]. Previous work in this thesis has focused exclusively on the class I ligand-binding PDZ domain, $PSD95^{pdz3}$. Class I domains bind ligands with the consensus C-terminal sequence -X-**S/T**-X-ϕ-COOH (where X is any amino acid and ϕ is any hydrophobic amino acid). $PSD95^{pdz3}$ binds its canonical class I ligand, CRIPT (-TKNYKQ**T**SV-COOH, derived from the cysteine-rich interactor of $PSD95^{pdz3}$) with a K_D of 0.452 ± 0.028 μM. Other ligand types, such as class II (-X-ϕ-X-ϕ-COOH) and class III (-X-**D/E**-X-ϕ-COOH), are well-studied and are thought to present distinct binding challenges for a given PDZ domain. For instance, the CRIPT ligand can be switched to a class II ligand ($T_{-2}F$; -TKNYKQ**F**SV-COOH), which $PSD95^{pdz3}$ binds with a weaker K_D of 17.0 ± 1.1 μM.

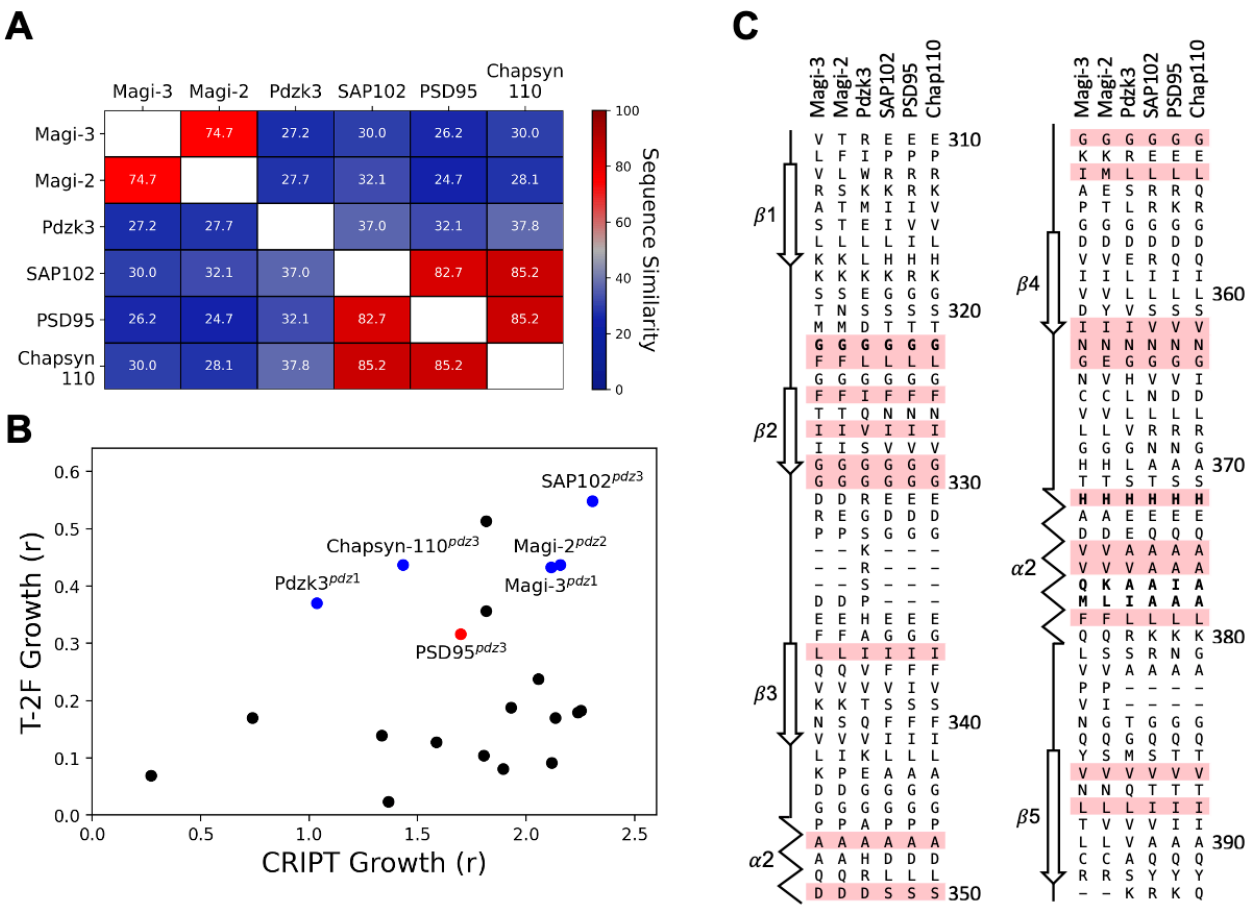

Figure 4.1. Comparison of PDZ domains in BTH-PACE

A. Percent similarity between all PDZ domains that were picked. Values for percent similarity are color coded from red (high similarity) to blue (low similarity).

B. 21 PDZ domains were assayed for BTH-PACE growth in the context of the CRIPT or $T_{-2}F$ ligand. r values are determined by fitting phage titers to a growth curve of the form $N = N_0 \times e^{rt}$. $PSD95^{pdz3}$ (red dots), the five other PDZ domains that were picked for further analysis (blue dot) and all other PDZ domains tested (black dots) are shown.

C. An alignment of the 6 PDZ domains picked. Secondary structure of the $PSD95^{pdz3}$ domain is shown. Sector positions are shown in red. Positions, 322, 372, 377, and 378 are in bold.

Alternating the ligand (CRIPT or $T_{-2}F$) required for growth therefore constitutes a changing environment that can be controlled in BTH-PACE. PDZ domains selected for evolution in BTH-PACE had to be able to bind both the CRIPT and $T_{-2}F$ ligands strongly enough to not wash out of the lagoon. 21 PDZ domains, which had previously been identified to bind the CRIPT ligand[5], were assayed for their ability to grow in BTH-PACE when challenged with either the CRIPT or $T_{-2}F$ ligand (Figure 4.1.B). Of these, six

domains, including PSD95^{pdz3}, were picked because of their ability to bind to both CRIPT and $T_{-2}F$ well enough to grow in BTH-PACE. These six domains (PSD95^{pdz3}, SAP102^{pdz3}, Magi-2^{pdz2}, Pdzk3^{pdz1}, Magi-3^{pdz1}, and Chapsyn-110^{pdz3}) have had different functional requirements and selection pressure fluctuations throughout their history. Therefore, any relationship that is found between evolutionary history and protein architecture will support the hypothesis that a proteins architecture is a consequence of its evolutionary history (Figure 4.1.A).

Evolutionary trajectories of PDZ domains in BTH-PACE

The three different fluctuation rates tested in BTH-PACE are defined by the rate at which the environment switches between requiring CRIPT ligand binding and $T_{-2}F$ ligand binding. The timescale of these environmental fluctuations is determined by the empirically measured rate of mutation in BTH-PACE. One environment, termed the medium fluctuation rate, had a ligand switching time of 72 hours, which corresponds to the average time it takes for the majority of the PDZ population to acquire a single mutation in BTH-PACE. A slow fluctuation rate was also chosen with a ligand switching time of 144 hours. In total, five lagoons containing PSD95^{pdz3}, as well as one lagoon for each of the other five PDZ domains highlighted in Figure 4.1, were evolved. Each experiment ran for 504 hours under slow and medium fluctuation regimes, as well as a constant environment where the CRIPT ligand was never changed.

In all environmental conditions and for all proteins tested, the total phage population was quantitated using plaque assays. Since population bottlenecks can complicate analysis, any lagoon where the total phage population dropped below 5×10^4 was excluded from further analysis. Using Illumina sequencing, the frequencies of specific

mutants within each lagoon were tracked throughout the experiment (Figure 4.2). Different variants of the starting PDZ domain rose and fell in frequency over time, with up to seven mutations accumulating during the experiment. The accumulation of mutations within all PDZ domains tested suggests that none of the PDZ domains are optimized for any of the selection conditions, even constant selection for CRIPT ligand binding. An example of one specific evolutionary trajectory, medium fluctuation between ligands starting with SAP102^{pdz3} is shown in Figure 4.2.

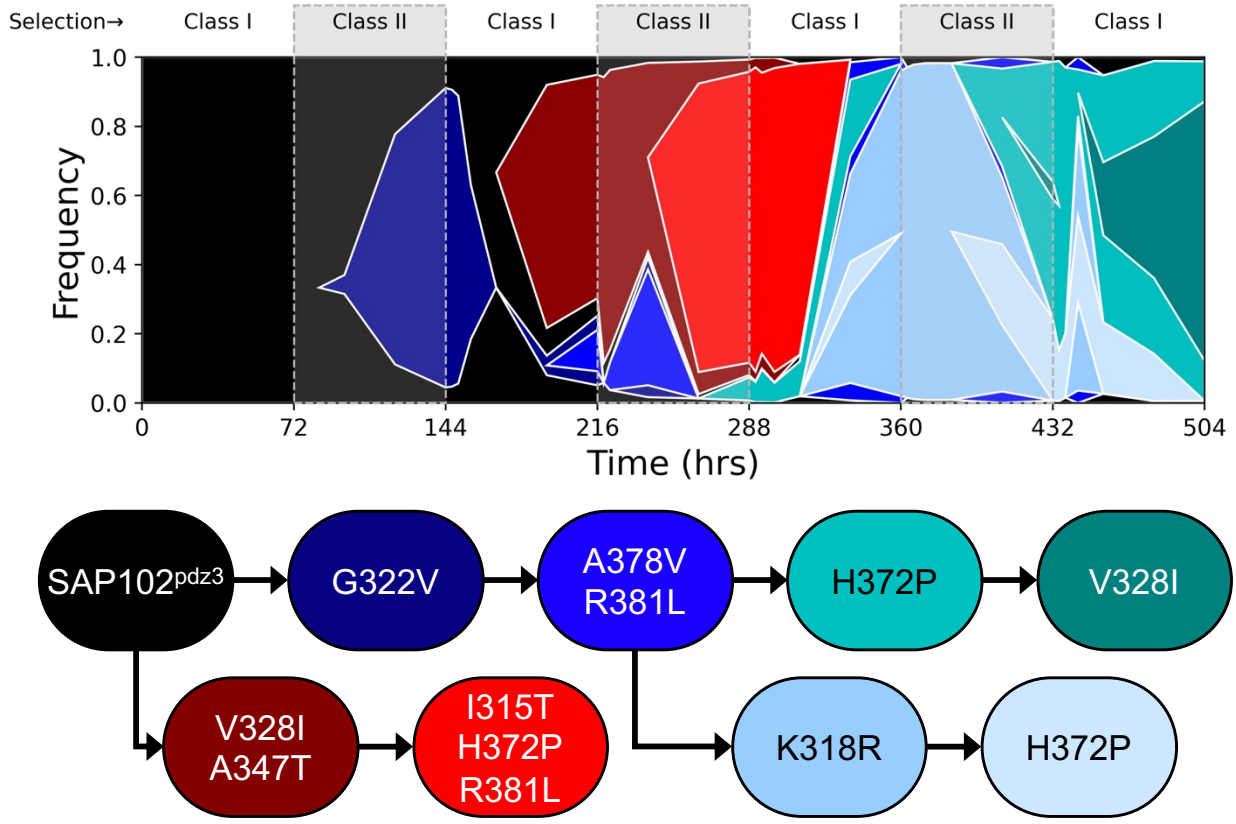

Figure 4.2. An evolutionary trajectory in BTH-PACE
A Muller plot of a BTH-PACE trajectory for the SAP102^{pdz3} protein in the medium fluctuation (72 hour) regime. All variants (colored wedges) which reached a frequency above 5% of the total PDZ population for a given time point are shown. A wedge which is started within another wedge (i.e., dark blue G322V in black SAP102^{pdz3}) indicates a mutation within the preceding (left) wedge. Wedges contain all the mutations for all wedges they are descendent from. The dominant species at the end of this trajectory is the dark green wedge, SAP102^{pdz3}; G322V; V328I; H372P; A378V; R381L.

When the mutations that arose during BTH-PACE are mapped onto the primary sequence of the PSD95^{pdz3} domain, it becomes clear that certain hotspots are particularly prone to adaptation in each set of environments (Figure 4.3.A). Positions 377 and 378 in the α2 helix, which do not directly contact the bound ligand, were found to mutate in every condition (Figure 4.4). Unfortunately, little is known about the specific roles these positions play in PDZ function. The mutations frequently observed in the PSD95^{pdz3} trajectories (I377F and A378V, Figure 4.3.B) are not individually known to be adaptive for either CRIPT or $T_{-2}F$ ligand binding (Figure 3.2.A, Figure 3.4.A). Consistent with this, the mutations I377F and A378V do not occur independently in these trajectories but exist almost exclusively as a pair, suggesting the presence of epistasis (Figure 4.5).

In contrast to the slow or constant environments, mutations at positions 372 and 322 are the most common in the medium fluctuation environment. This represents the first observed difference between proteins with varying evolutionary histories. Position 322, located in the β1-β2 carboxylate binding loop, and which always mutates before position 372, is highly conserved and contains a glycine residue in all six proteins (Figure 4.1.C). This glycine provides the loop with conformational flexibility, where, upon ligand binding, the protein incurs an entropic cost as the loop transitions from an open to a closed conformation (Figure 4.6.A). Mutations at position 322 pre-clamp the loop, reducing this entropic cost for both CRIPT and $T_{-2}F$ ligands (Figure 4.6.B, Figure 3.2.A, Figure 3.5.A). Notably, only in the medium fluctuation environment, where the environment fluctuates fastest, does this mutation, which allows for consistent binding of both ligands, consistently persist throughout the experiment. This suggests that G322V, while

improving function in general, may be near a local fitness peak and is eventually outcompeted when environmental conditions remain static for extended periods.

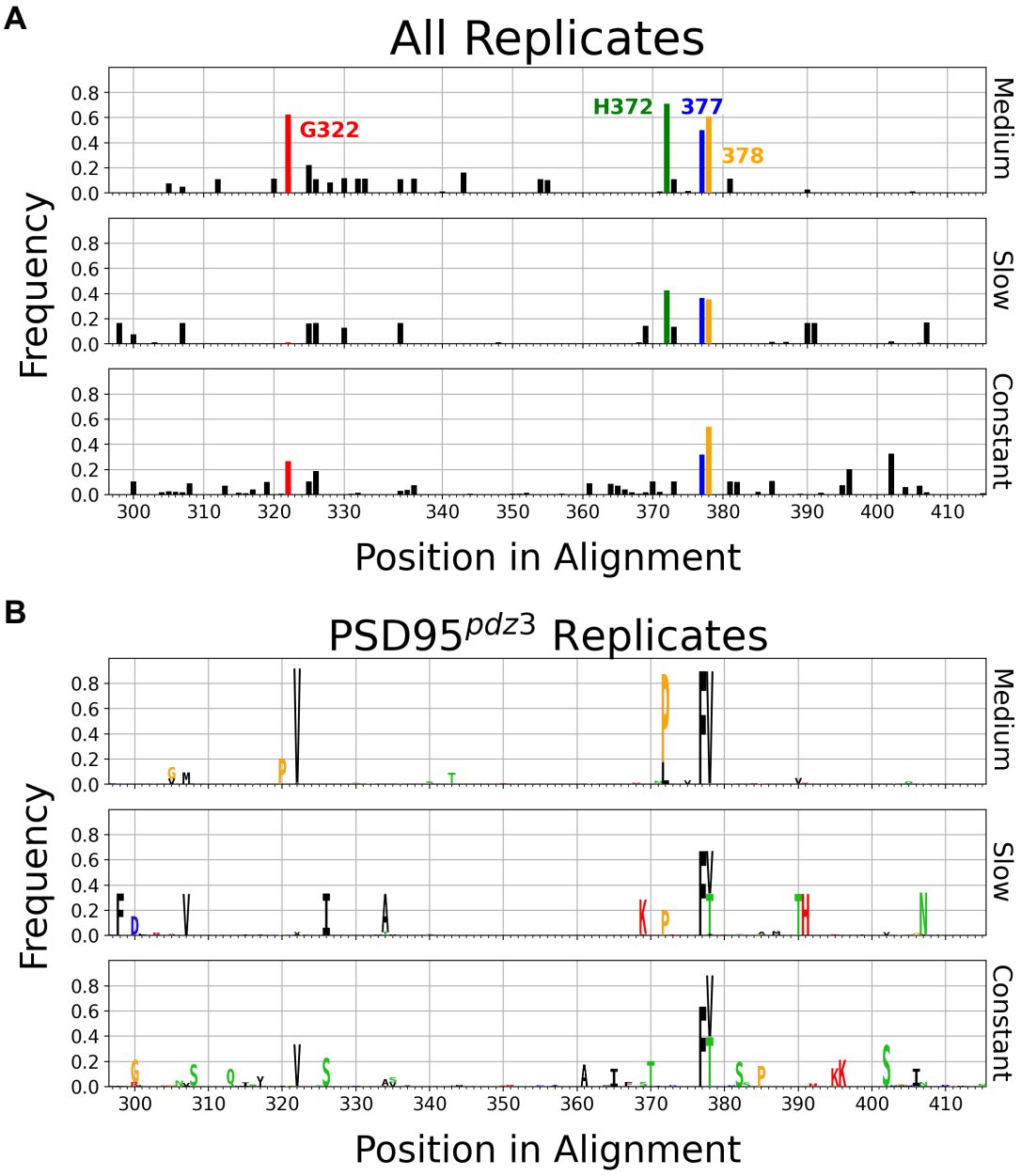

Figure 4.3. Combined mutational spectrum at the end of BTH-PACE

A. Average frequency of mutation at each site in the tested PDZ domains mapped onto its equivalent position in PSD95^{pdz3} for each environmental condition. Medium fluctuation has 9 lagoons (Pdzk3^{pdz1} was dropped), slow has 6 lagoons (2 of PSD95^{pdz3}, and 1 each of SAP102^{pdz3} and Magi-2^{pdz2} were dropped), and constant selection has 10 lagoons. Positions of interest are colored (322 – red, 372 – green, 377 – blue, and 378 – yellow).

B. Same as in B except data for only PSD95^{pdz3} lagoons are shown. Instead of a bar chart the specific mutations are shown by single letter. The height of a letter corresponds to the frequency of that mutation. Amino acids are color coded by function (red – positively charged, blue – negatively charged, green – polar, black – hydrophobic, and yellow – special).

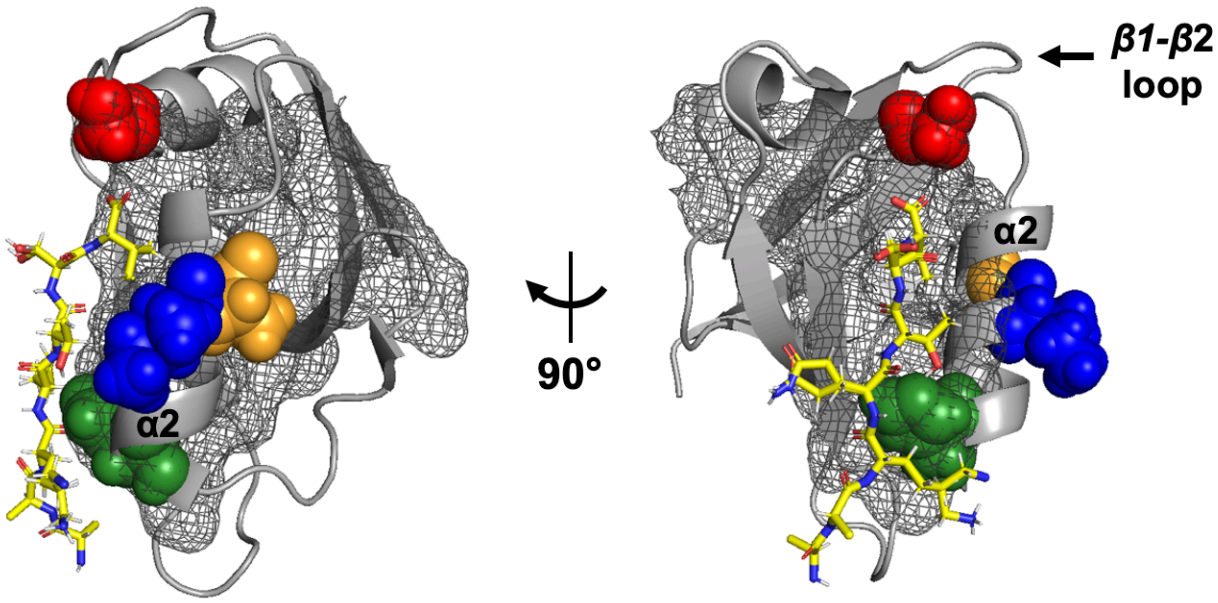

Figure 4.4. Mutational hotspots in PSD95^{pdz3}
Ribbon representation of the crystal structure of PSD95^{pdz3} (PDB ID: 5HEB) with its native CRIPT ligand (yellow). The sector is shown in grey mesh. Positions 322, 372, 377, and 378 are shown in red, green, blue, and yellow respectively.

The third, and somewhat unexpected, mutation hotspot, which is primarily observed in the medium fluctuation environment, is position 372 (Figure 4.3). In the PDZ alignment, position 372 is a highly conserved histidine residue that directly contacts the -2 position of the PDZ ligand (Figure 4.1.C)[3]. In PSD95^{pdz3}, this histidine forms a hydrogen bond with the threonine of the CRIPT ligand (Figure 4.6.C). When the -2 position in the ligand is mutated to phenylalanine, a bulky aromatic residue, His_{372} in PSD95^{pdz3} is forced to adopt an alternate conformation to prevent steric clashing, resulting in a significantly reduced binding affinity for the $T_{-2}F$ ligand (Figure 4.6.D, Figure 3.5). Although crystal structures for their binding to $T_{-2}F$ do not exist, this phenomenon is expected to be similar in the other five PDZ domains tested. Mutating position 372 to proline removes the steric clash and significantly improves binding to the $T_{-2}F$ ligand (Figure 4.6.E). However, the

H372P mutation alone, at least in PSD95^{pdz3}, drastically decreases the PDZ domain's ability to bind the CRIPT ligand (K_D = 46.1 ± 8.8 μM).

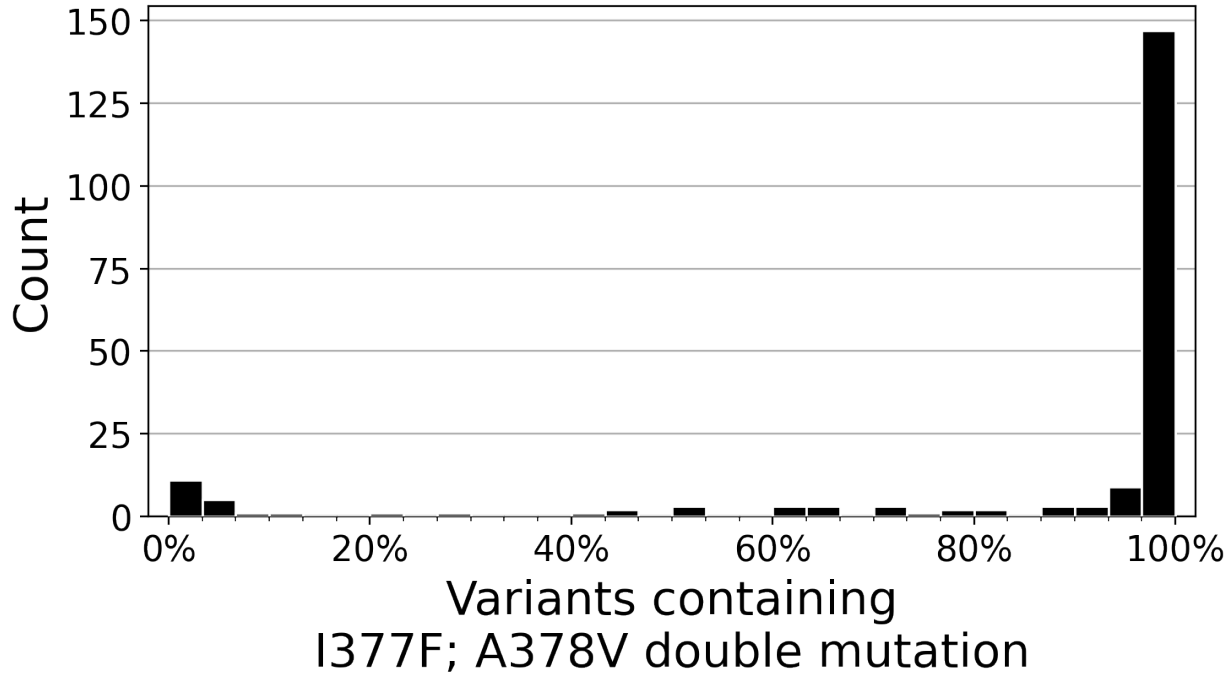

Figure 4.5. Co-mutation of A377F and A378V in PS95^{pdz3}
In PSD95^{pdz3}, for all samples and all selection regimes, the fraction of the population containing at least one of the I377F or A378V mutations is found. A histogram is made by binning samples based on the percent of this fraction that contains the double mutant (I377F; A378V) as opposed to either single mutant. It is unlikely to see single mutants of I377F or A378V in the BTH-PACE experiments started with PSD95^{pdz3}.

The H372P mutation, located near the active site and occurring exclusively alongside other mutations, can be understood through the "outside-in" hypothesis of molecular evolution. This hypothesis suggests that adaptation typically start at the periphery of a protein and gradually move toward more functionally critical regions, such as the active site[6-9]. The observation that the H372P mutation arises later in evolutionary trajectories supports this model. It implies that initial mutations, like the more distant and conditionally neutral G322V mutation in the β1-β2 loop, enable the protein to explore new functions without immediately impacting its current functionality. As evolution progresses, mutational changes in the protein outside the active site set the stage for subsequent

mutations like H372P, which now only fine-tune the domain's substrate specificity rather than drastically altering it. Moreover, since mutations at position 372 persist most often during the fastest fluctuation rate tested, a related hypothesis can be proposed: outside-in evolution, which relies on adaptation through CN mutants, occurs only when the rate of environmental change is high enough to preclude significant adaptation between fluctuations (Figure 4.7).

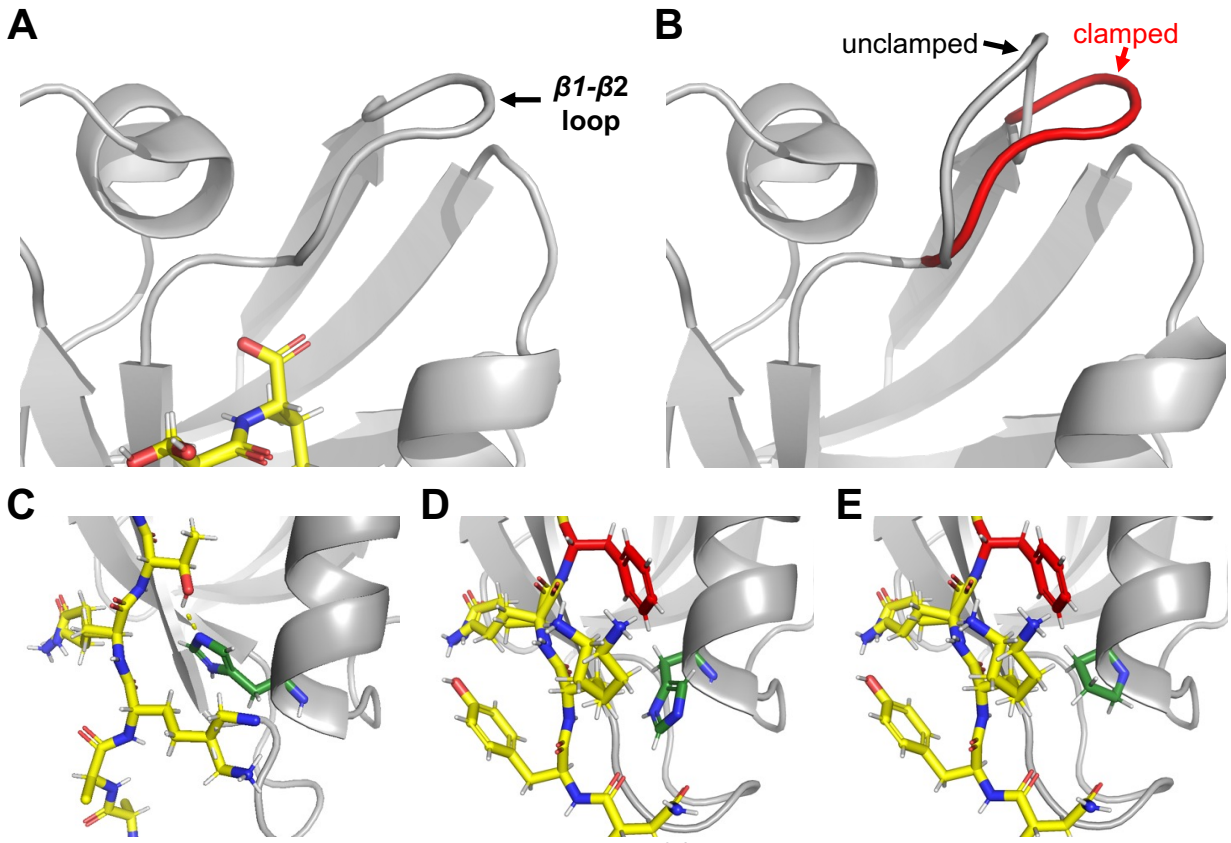

Figure 4.6. Known mutational effects in PSD95^{pdz3}

A. *Ribbon representation of the crystal structure of PSD95^{pdz3} (PDB ID: 5HEB) with its native CRIPT ligand (yellow). As the PDZ domain is binding the CIRPT ligand, the β1-β2 loop is clamped.*
B. *A G322A mutation preclamps the β1-β2 loop (red). Preclamping occurs even without ligand present and mimics the clamping seen in (A). This lowers the entropic cost of binding.*
C. *Same as in (A) with the focus now on position 372 (green sticks). The histidine at position 372 makes a hydrogen bond (yellow dashed lines) with the threonine in the -2 position of the CRIPT ligand.*
D. *Mutating the -2 position of the CRIPT ligand to a phenylalanine (red sticks) creates a steric clash with the histidine at position 372 (PDB ID: 5HED). This steric clash, which completely removes the hydrogen bond between position 372 in the PDZ domain and position -2 in the ligand, is a large factor in the lower binding affinity of PSD95^{pdz3} for the T$_{-2}$F ligand.*
E. *Same as in (D) except position 372 has been mutated to a proline (green sticks) in PyMol. This removes the steric clash and increases the binding affinity of PSD95^{pdz3} for the T$_{-2}$F ligand.*

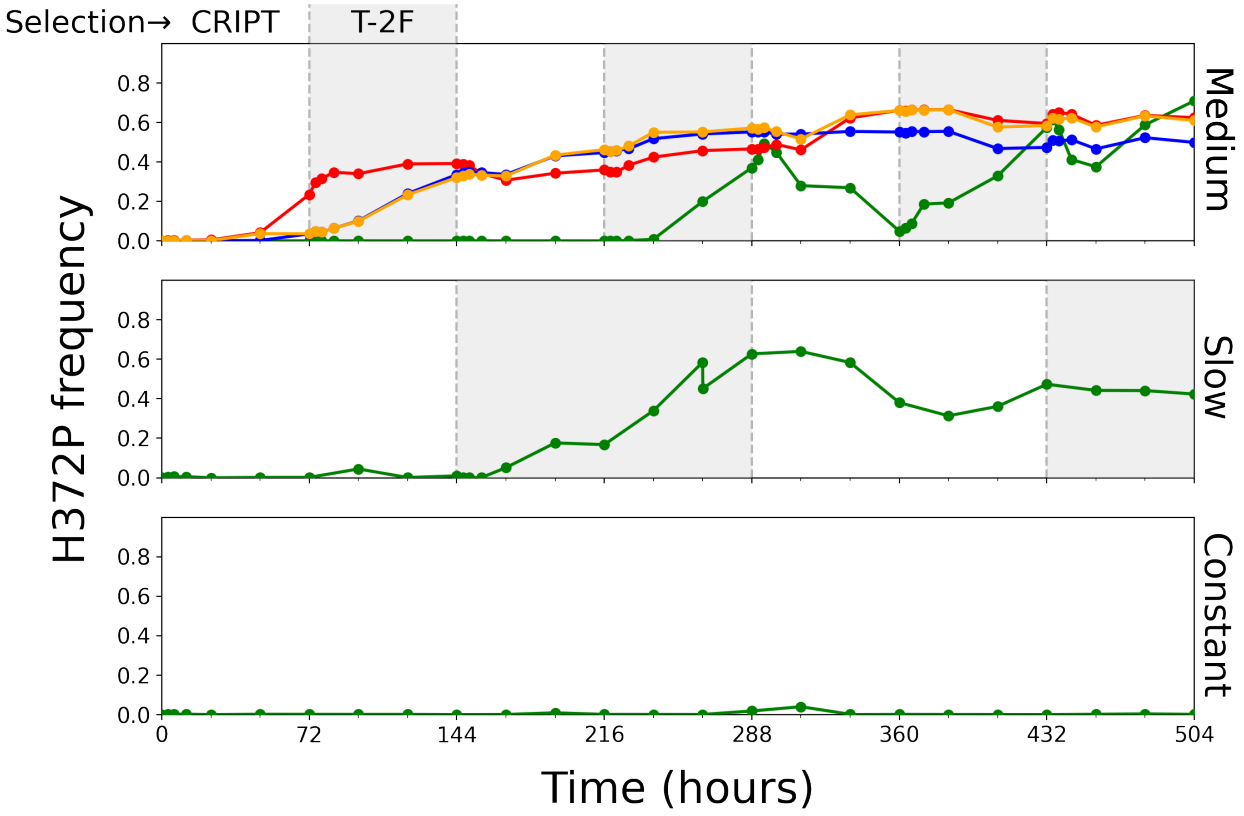

Figure 4.7. Trajectory of mutations at site 372
The average frequency of mutations at site 372 (green) is tracked over time for all three selection regimes (individual samples, black dots). The selection regime specified on the right of each plot and the condition of selection at a given time is indicated by the shading of the plot (CRIPT ligand binding – white; $T_{-2}F$ – grey). For the medium fluctuation regime, mutations at positions 322 (red), 377 (blue), and 378 (yellow) are also shown.

Latent function in evolved PDZ domains

To further investigate the impact of evolutionary history on a protein's architecture, a comprehensive analysis of the adaptability of these proteins is required. As previously mentioned, the sector of a protein – a feature of its architecture – coincides with its allosteric network, which facilitates CN mediated exaptation through the presence of pre-existing mutations distant from the active site. If a protein's architecture is influenced by the different environmental fluctuation rates tested in this study, this should be observed through a quantitative assessment of CN. One method to quantify the degree of CN is by measuring the ability of the dominant evolved PDZ domains from the ends of these

evolutionary trajectories to bind novel ligands relative to the starting condition (Table 4.1.A). In these BTH-PACE experiments, binding to both the class I ligand, CRIPT, and the class II ligand, $T_{-2}F$, was essential for survival. However, many additional PDZ ligands, both those that can be classified as class I, II, or III ligands, and those that cannot, exist.

Previous work has shown that PDZ ligands, traditionally categorized into three classes, can actually be divided into as many as 16 groups and subgroups[4,10]. To help identify novel ligands, multiple correspondence analysis (MCA) was performed on data previously collected by Stiffler *et al*, 2006[5]. MCA analysis identified five distinct dimensions in the data, from which thirteen ligands were selected to represent the full range of these dimensions. Ten ligands can be neatly classified into the three established PDZ ligand classes. The remaining three ligands, while still representing naturally occurring PDZ ligands, do not fit into these traditional classifications (Table 4.1.B).

A

PACE selection condition	PDZ domain	Accumulated mutations
slow	PSD95^{pdz3}	S298F; I307V; N326I; I377F; A378T; A390T; Q391H
slow	Pdzk3^{pdz1}	H372P
slow	Magi-3^{pdz1}	F325L; G330R; A373E
slow	Chapsyn110^{pdz3}	R368C; H372P; A377E
medium	PSD95^{pdz3}	S320P; G322V; H372P; I377F; A378V
medium	SAP102^{pdz3}	G322V; V328I; H372P; A378V; R381L
medium	Magi-3^{pdz1}	V312L; F325L; T326I; R332H; E334D; A354D; A373E
medium	Chapsyn110^{pdz3}	G333W; H372P
constant	PSD95^{pdz3}	D306N; G322V; A378T; A402S
constant	SAP102^{pdz3}	G319R; G322V; R381H
constant	Magi-3^{pdz1}	F325L; A373E; M378V
constant	Magi-2^{pdz2}	T326I; E332G; L336Q; E404X
constant	Chapsyn110^{pdz3}	T385P; D396G

B

Ligand	Ligand Sequence	Ligand Class
AcvR2	ESSL	1
Claudin 23	DSDL	1
CRIPT	QTSV	1
GluR1	ATGL	1
Kir2.2	ESEI	1
NMDAR2A	ESDV	1
TRPC5	TTRL	1
Claudin 2	TGYV	2
SSTR2	IAWV	2
T-2F	QFSV	2
Claudin 18	YDYV	3
Mel1a/b	VDSV	3
Claudin 4	SNYV	?
Kv2.1	DQSI	?
Neurexin 1/2	EYYV	?

Table 4.1. PDZ domains and ligands used in binding growth assay

A. PDZ domains picked from the endpoint of BTH-PACE evolution trajectories for growth rate analysis in a bacterial-two-hybrid binding assay. X is treated as a stop codon here.

*B. PDZ ligands picked for growth rate analysis. The ligand sequence provided is the 4 C-terminal amino acids. These amino acids are thought to be the determinants of PDZ-ligand binding. Ligand classes are determined by traditional class criteria (*Class I, -X-**S/T**-X-ϕ-COOH; class II (-X-ϕ-X-ϕ-COOH); class III, -X-**D/E**-X-ϕ-COOH; where X is any amino acid and ϕ is any hydrophobic amino acid).

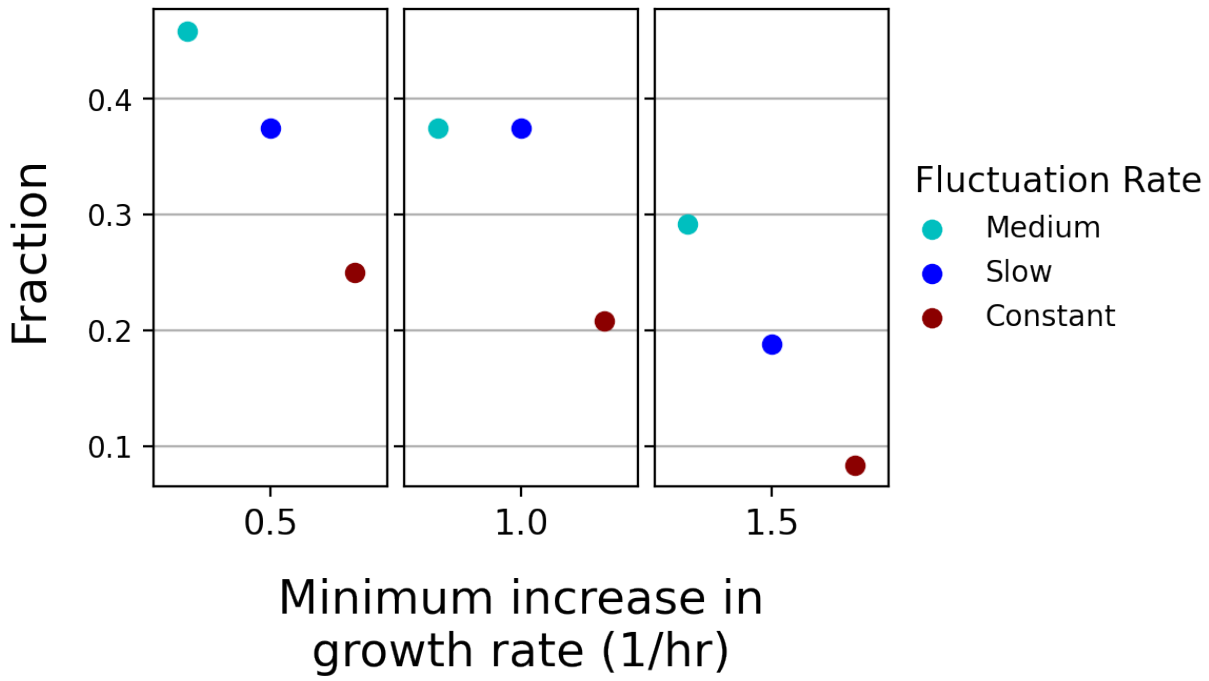

Figure 4.8. Fraction of evolved PDZ domains binding to novel ligands
Shown is the fraction of evolved PDZ-novel ligand pairs that increased the growth rate by at least 0.5, 1.0, or 1.5/hr for each BTH-PACE fluctuation rate. PDZ domains were picked from the end of each BTH-PACE experiment.

Individual PDZ domains were picked from the end of BTH-PACE trajectories for this analysis (Table 4.1.A). Individual PDZ domains are used to isolate alterations to internal architecture of single proteins from functional differences in the SGV of the population. Regardless of which selection condition that an evolved PDZ domain came from, BTH-PACE improved the fraction with which the PDZ evolved domains could bind to novel ligands in a growth rate-dependent assay. However, those domains evolved under the medium fluctuation environment exhibit a distinct advantage in their ability to bind alternative ligands (Figure 4.8). For the medium fluctuation period, 46% of the evolved PDZ-ligand pairs tested showed an increase of at least 0.5/hr in their growth rate constant, r (where $N(t) = N_0 e^{rt}$, with N as the total phage population, N_0 as the starting phage population, and t as time). In contrast, only 38% and 25% of evolved PDZ-ligand

pairs exhibited such an increase under slow fluctuation and constant CRIPT ligand conditions, respectively. This result, where the fraction of functional PDZ-ligand pairs for the fluctuating environments is higher than for the constant environment, is not dependent on the arbitrary 0.5/hr cutoff used here. Since r represents the growth rate in a bacterial two-hybrid binding assay, an increase in r implies two key points. First, these evolved proteins, while still able to survive in their original environment, have enhanced their ability to bind alternative ligands. Second, if these proteins were evolved in a new BTH-PACE experiment with a ligand other than CRIPT or $T_{-2}F$, PDZ domains evolved under the medium fluctuation regime would likely have a competitive advantage.

4.4. Discussion

Impact of Evolutionary History on Mutation Patterns

This chapter presents an early experimental assessment of how evolutionary history influences the current architecture of proteins. Six PDZ domains were evolved under three different environmental fluctuation rates. As expected, proteins with different evolutionary histories exhibited distinct patterns of accumulated mutations, exemplified by position 372 and specifically the H372P mutation. Since the H372P mutation switches ligand binding preference from the CRIPT ligand to the $T_{-2}F$ ligand, constant selection for CRIPT ligand binding resulted in minimal accumulation of this mutation. Under slow fluctuation, the H372P mutation appeared but was partially depleted as the ligand binding requirement shifted back from CRIPT to $T_{-2}F$. It was only in the medium fluctuation environment that the H372P mutation persisted in the majority of the evolutionary trajectories after 504 generations (Figure 4.7).

Interestingly, the H372P mutation was only acquired after the PDZ domains had already accumulated peripheral mutations (Figure 4.2, Figure 4.7). This suggests that prior changes to the active site configuration, likely mediated through the protein's allosteric network, were necessary before H372P could be tolerated. This observation supports the theory that evolution, at least under certain conditions, proceeds in an "outside-in" manner. Prior work has proposed this as a common evolutionary mechanism, and while this study supports that hypothesis, additional evidence is needed to fully understand the extent and conditions under which “outside-in” evolution is possible.

Changes in protein architecture as a function of evolutionary history

The extent of latent ligand binding function, or the level of CN function, was used to evaluate changes in protein architecture. Individual proteins evolved in an environment where selection conditions changed every 72 hours (medium), while still maintaining fitness in their version of BTH-PACE, demonstrated the greatest ability to bind alternative ligands compared to those evolved under other conditions (Figure 4.8). This increased level of CN function indicates that these proteins have experienced alterations in their internal architecture, shaped by the specific patterns of selection pressures they encountered during evolution. The data here provide early evidence supporting the hypothesis of a causal relationship between environmental history and protein architecture. This highlights how evolutionary pressures can fundamentally reshape protein constraints, altering their ability to adapt to novel conditions.

While this work is already suggestive of evolutionary history influencing protein architecture, further research is needed. The current quantification of conditional neutrality, while useful at this initial stage, does not differentiate between an expansion of

the ligand-binding space of a PDZ domain (a “generalist” for ligand binding function) and alterations within that space (alterations to ligand binding specificity). Previous studies have shown that single mutations alter the ligand-binding space but do not necessarily create generalists for ligand binding[3]. However, this does not guarantee that subsequent mutations, as observed in BTH-PACE, will continue to have the same effect. This should be investigated further. Moreover, the quantification of conditional neutrality shown in Figure 4.8 is averaged across multiple condition. While this already suggests there is a relationship between the exaptive capacity of individual proteins – due to alterations in their internal architecture – and their statistical history of selection, to achieve a condition-specific quantification of conditional neutrality, a different approach is required. Minor modifications to a novel method, High-Density Luria-Delbrück by Sequencing (HiDenSeq), discussed in Chapter 5 of this thesis, could enable this quantification (see Chapter 6.2.2 for proposed modifications to HiDenSeq).

The preliminary work presented here offer a guide as to whether different evolutionary histories result in differences in protein architectures. If subsequent data continue to support this, it will raise several new questions for the field of molecular evolution. One key question is where in the protein the constraints imposed by evolutionary history are stored. Given that CN is encoded along the epistatic network defined by the sector, it is tempting to propose that the sector serves as this repository of information[3,11]. Supporting this view, simulation work has shown that the size of an allosteric network in Ising models of proteins correlates with the fluctuation rate of its environment. However, Chapter 3 of this thesis demonstrates that the connection between the sector and its surroundings is crucial for an evolvable protein. Understanding

the extent to which the sector, its surroundings, or other components of the protein are constrained by evolutionary history will be important for fully appreciating the design of natural proteins.

A second, potentially more challenging problem is to learn how specific environmental histories are encoded in protein sequences. Just as proteins are optimized for their functions within a cell, they are also likely optimized for expression and biosynthesis within the environmental history in which they evolved. As synthetic and designed proteins continue to advance and gain therapeutic relevance, an important consideration will be to specify constraints that align with the expected environment in which they will function[12-13]. This is particularly relevant for applications where the genetic information of a synthetic gene is passed through multiple generations[14-15].

4.5. References

1 Hemery M & Rivoire O. Evolution of sparsity and modularity in a model of protein allostery. *Physical Review E* **91**, 042704, (2015).

2 Murugan A, *et al*. Roadmap on biology in time varying environments. *Physical Biology* **18**, 041502, (2021).

3 Raman AS, *et al*. Origins of allostery and evolvability in proteins: A case study. *Cell* **166**, 468-480, (2016).

4 Lee HJ & Zheng JJ. PDZ domains and their binding partners: structure, specificity, and modification. *Cell Communication and Signaling* **8**, 8, (2010).

5 Stiffler MA, *et al*. PDZ domain binding selectivity is optimized across the mouse proteome. *Science* **20**, 364-369, (2007).

6 Hedstrom L, *et al*. Converting trypsin to chymotrypsin: the role of surface loops. *Science* **255**, 1249-1255, (1992).

7 Hedstrom L, *et al*. Converting trypsin to chymotrypsin: ground-state binding does not determine substrate specificity. *Biochemistry* **33**, 8764-8769, (1994).

8 McLaughlin RN, *et al*. The spatial architecture of protein function and adaptation. *Nature* **491**, 138-142, (2012).

9 Rivoire O. Parsimonious evolutionary scenario for the origin of allostery and coevolution patterns in proteins. *Physical Review E* **100**, 032411, (2019).

10 Tonikian R, *et al*. A specificity map for the PDZ domain family. *PLoS Biology* **6**, e239*, (2008).*

11 Süel GM, *et al*. Evolutionarily conserved networks of residues mediate allosteric communication in proteins. *Nature Structural and Molecular Biology* **10**, 59-69, (2003).

12 Silva DA, *et al.* De novo design of potent and selective mimics of IL-2 and IL-15. *Nature* **565**, 186-191, (2019).

13 Ebrahimi SB & Samanta D. Engineering protein-based therapeutics through structural and chemical design. *Nature Communications* **14**, 2411, (2023).

14 Wright O, *et al*. Building-in biosafety for synthetic biology. *Microbiology Society* **159**, 1221-1235, (2013).

15 Schleidgen S, *et al*. Human germline editing in the era of CRISP-Cas: risk and uncertainty, inter-generational responsibility, therapeutic legitimacy. *BMC Medical Ethics* **21**, 87, (2020).

Chapter 5. Development of a high throughput Luria-Delbrück assay

5.1. Abstract

The parameters defining the distribution of mutants in the classic Luria-Delbrück experiment offer the potential for directly quantifying a system's ability to generate conditional neutrality (CN) and other evolutionary factors. Historically, accurately parameterizing this distribution required hundreds to thousands of replicates, a task that was logistically infeasible until recently. However, through a novel experimental procedure called high-density Luria-Delbrück by sequencing (HiDenSeq), initially developed by Kabir Husain, we have achieved this goal. HiDenSeq utilizes advanced sequencing techniques to efficiently analyze a vast number of samples, thus overcoming previous limitations. Additionally, using HiDenSeq, we provide the first experimental validation of an extension to the fundamental Luria-Delbrück distribution theory, demonstrating that the system's selection pressure affects the distribution's shape. Although further extensions to the theory remain to be tested, this finding opens new avenues for quantifying and understanding evolutionary processes.

5.2. Introduction

As molecular biology research continues to explore the genetic underpinnings of evolution, there is a growing need for sophisticated methodologies that can accurately detect, analyze, and interpret evolutionary processes[1-3]. The previous sections of this thesis also emphasize the necessity of experimental tools to assay molecular evolution. These tools are essential for uncovering the intricate mechanisms, as well as where they are encoded within proteins, that govern genetic variation, adaptation, and exaptation at

the molecular level[4]. By developing and refining these tools, we can enhance our understanding of molecular evolution, leading to insights into genetic diversity, the evolution of genomes, and the molecular basis of adaptation and exaptation. As more advanced design algorithms are developed the need for assays testing these evolutionary parameters will become increasingly important to quantify any differences that remain relative to natural proteins[5-7].

One of the first experiments used to assay evolutionary parameters was the famed Luria-Delbrück experiment, conducted in 1943, now more commonly known as the fluctuation assay[8-9]. This assay, which led to the formulation of the Luria-Delbrück distribution, was used to determine the mutation rate of the system and the number of phage-resistant mutants arising in a population that initially had none. With these data, they inferred the presence of pre-existing genetic variation. Over the past 80 years, this distribution has been the focus of extensive research, and, when fully parameterized, it can also be used to quantify parameters such as the level of CN or even the full distribution of fitness effects (DFE) experienced by a protein under specific environmental conditions[10-13]. However, fully parameterizing the Luria-Delbrück distribution requires thousands, if not tens to hundreds of thousands, of independent trials, a task that has historically been unfeasible due to logistical limitations[14-15].

With the help of post-doc Kabir Husain, from the lab of Arvind Murugan, we are able to overcome these logistical limitations by using a novel experimental procedure called High-Density Luria-Delbrück by Sequencing (HiDenSeq). This method allows the number of independent experiments to be limited only by transformation efficiency, enabling the execution of hundreds of thousands of replicates. We demonstrate that

HiDenSeq not only fully recapitulates the original Luria-Delbrück distribution but also quantifies an extension to this basic theory by accounting for environmental selection pressure. Additionally, we propose that other simple extensions, such as a bimodal distribution of fitness effects (DFE) and the existence of CN, should also be quantifiable through HiDenSeq.

5.3. Results

A theoretical examination of the Luria-Delbrück distribution

In the classic Luria-Delbrück experiment, a small number of bacterial cells were used to inoculate a culture and allowed to grow until their population size reached between 10^8 and 5×10^9 bacteria per mL[8]. At this point, they were plated onto agar containing T1 phage. The number of colonies, originating from bacteria that had acquired resistance to T1 phage through mutation, was counted and a distribution of these counts was plotted. For this seminal study, 280 replicates were completed; a number that, even with the use of robots, has not been significantly surpassed to this day. This work confirmed the presence of pre-existing genetic variation and identified the phenomenon of “jackpot” mutations. Jackpot mutations, which occur after only a few generations of growth in the Luria-Delbrück assay, give the distribution its characteristic power-law tail (Figure 5.1.A). Specifically, in the original formulation of this experiment, where there is no selection pressure during the initial growth period, the complementary cumulative distribution function (CCDF or 1-CDF) of the number of mutants, m, decays with a slope approaching -1 in log-log space until m begins to approach the final population size, N (Figure 5.1.B, Methods 7.9).

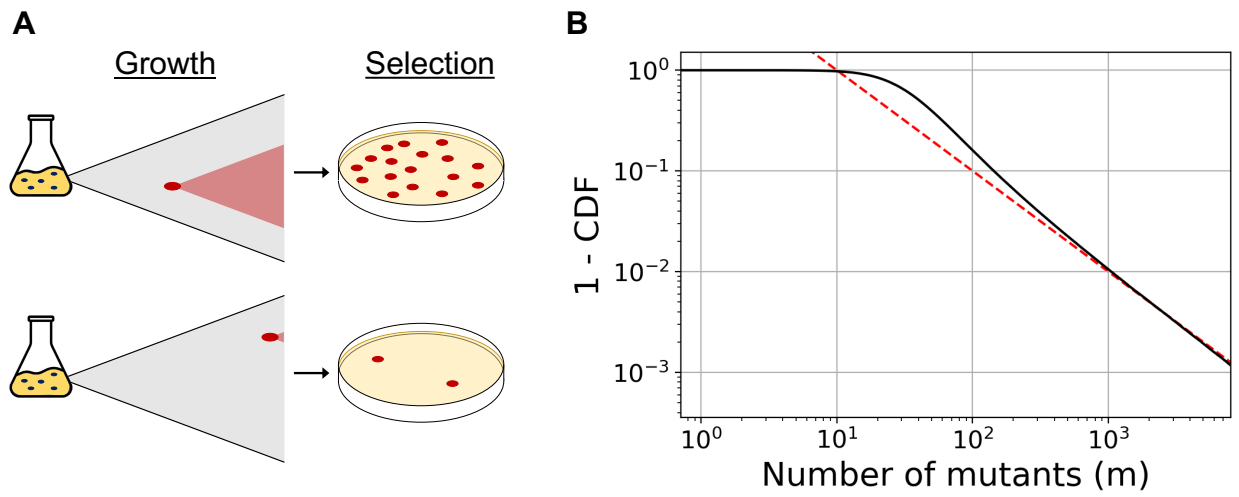

Figure 5.1. The Luria-Delbrück experiment and distribution

A. Shown here are two replicates of the Luria-Delbrück experiment. In both cases, a small number of bacteria, which lack a specific selectable function are used to inoculate a culture. The culture is allowed to grow until the total population reaches N. At this point, selection is applied, and the number of functional mutations, which now have the selectable function and therefor form colonies on a plate, is counted. In the top row, a jackpot mutation occurs, resulting in many functional colonies. Jackpot mutations occur early on in the growth phase and, as a consequence of the low population size at this time, are rare. In the second row, the only functional mutation occurs during the last few generations, which is more likely.

B. Simulation of the Luria-Delbrück distribution (Methods 7.8.2). The simulation (black line; $\sigma = 1$, $N = 10^5$, $\mu = 10^{-4}$) is done without selection pressure during the growth phase. The red-dashed line indicates a slope of -1.

Examination of the power-law tail of the distribution reveals its sensitivity to changing conditions in the experiment. In its original conception the Luria-Delbrück experiment was designed to exert very little selection pressure during the growth phase. However, this does not have to be the case. If some selection pressure is introduced, the resulting fitness effect of functional mutations, that is those that enable the original organism to survive selection, will alter the distribution of m values. This is also influenced by the mutation rate and the timing of individual mutations. We will define a parameter, σ, that is dependent on the selection pressure in the system, as the growth rate of a mutant relative to the growth rate of the wild-type genome.

$$\sigma \equiv \frac{\lambda_{mut}}{\lambda_{WT}} \qquad \text{(Eq. 5.1)}$$

In this case, if mutants grow slower than the wild-type genome σ will be less than 1. This will lead to a truncated distribution, since even jackpot mutations, which occur early in the growth phase, will not achieve a high count (Figure 5.2.A,C). This formulation also allows for the intriguing scenario where mutants grow faster than the wild type (σ > 1, Figure 5.2.B,C). Mathematically, the tail of the distribution will have a slope inversely proportional to σ, as long as m does not approach N (Eq. 5.2, Derivations 5.5.1).

$$Prob(\#\, mutants > m) \approx \mu N m^{-1/\sigma} \qquad \text{(Eq. 5.2)}$$

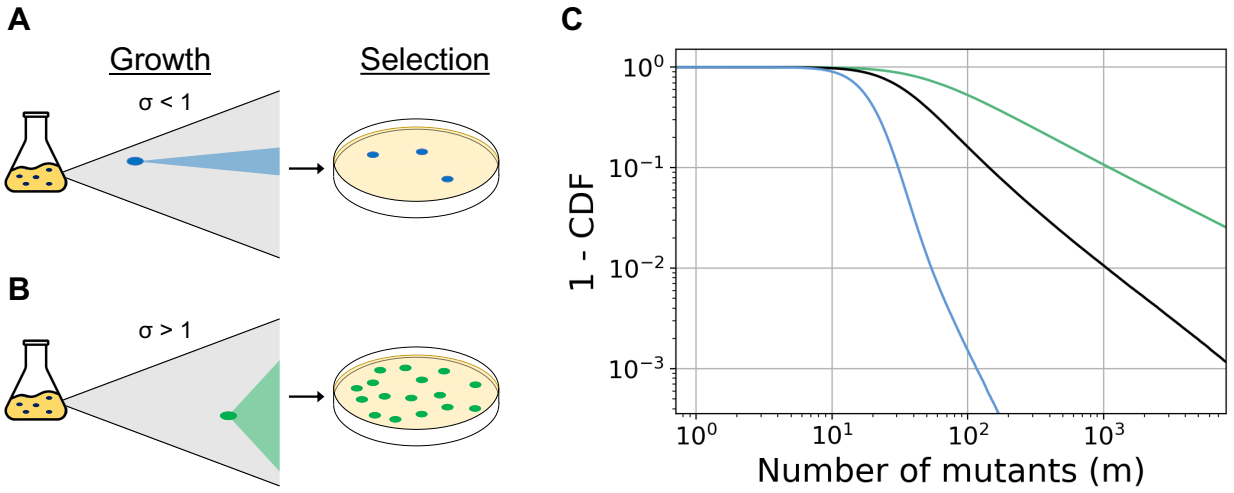

Figure 5.2. The effect of differential growth rates on the Luria-Delbrück distribution

A. Shown is a replicate of the Luria-Delbrück experiment as in 5.1.A. In this replicate, the mutation that arises has a slower growth rate than wildtype during the growth phase of the experiment. In this scenario, even jackpot mutations have few mutants present at the end.

B. Same as 5.2.A except the mutation that arises has a faster growth rate than wildtype during the growth phase of the experiment.

C. Three simulations of the Luria-Delbrück distribution with varying values of σ ($N = 10^5$, $\mu = 10^{-4}$ for all simulations). The blue line (σ = 0.5) represents a scenario like 5.2.A. The green line (σ = 1.5) represents a scenario like 5.2.B. The black line has σ = 1.

The full distribution is also dependent on μ, the rate of mutation to functional genotypes which can survive the selective challenge. This rate reflects a subset of all mutations generated during the growth phase is lower than the generic mutation rate as Luria-Delbrück experiments are blind to non-functional mutants. In all cases, deviations from a σ value equal to 1 will be exaggerated by increases in selection pressure in the

system. These three parameters, σ, μ, and N, are all that is needed to describe the full distribution.

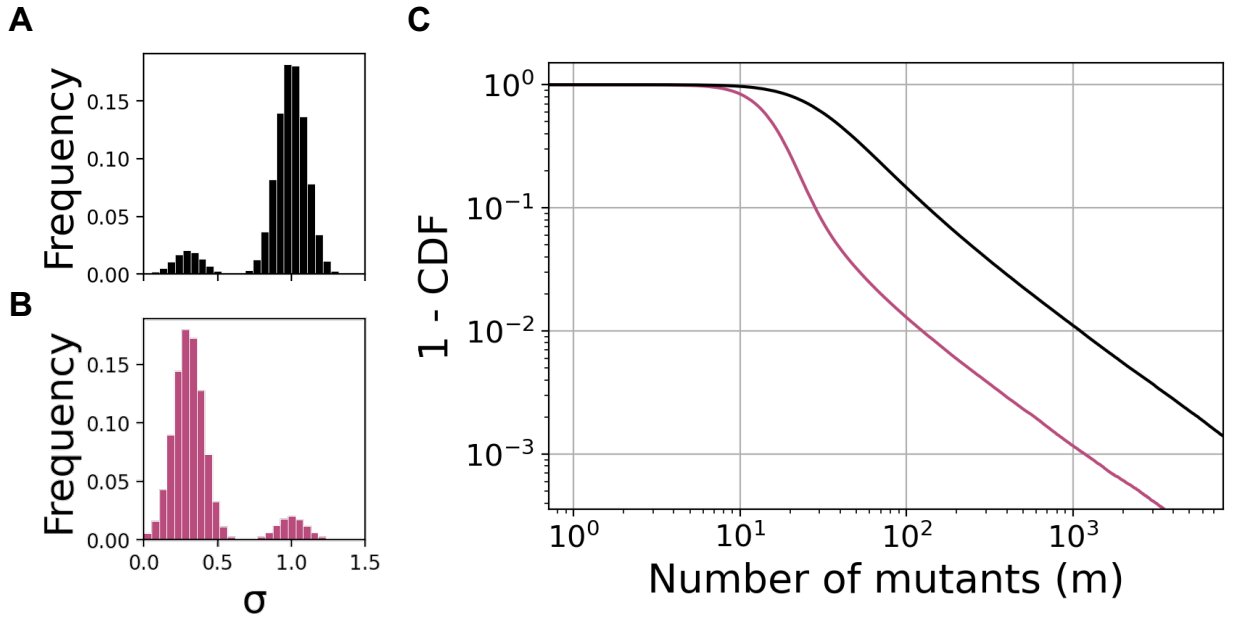

Figure 5.3. Luria-Delbrück distribution with a bimodal DFE

A. A gaussian mixture model of DFE where 90% of fitness effects are conditionally neutral (σ_{CN} = 1 ± 0.1) and 10% are direct switching (σ_{DS} = 0.3 ± 0.1).

B. Same as in 5.2A except 90% of mutants are direct switching and 10% are conditionally neutral.

C. Simulated distributions resulting from the DFEs shown in 5.3.A (black line) and 5.3.B (maroon line). A higher percentage of mutants belonging to the lower of two σ values in a bimodal DFE (more direct switching mutants) results in a biphasic Luria-Delbrück distribution with a steep initial drop off followed by a long tail sloping as -1/σ_{CN}.

Natural organisms do not produce mutations represented by a single value for σ; instead, there is a DFE covering a range of values for σ [16]. DMS experiments presented in this thesis and elsewhere have shown that a common form of the DFE for natural proteins is bimodal, with some mutations being approximately neutral and others being deleterious (Figure 3.2.D)[17-19]. Additionally, since only mutants capable of surviving under new selection are examined in the Luria-Delbrück experiment, these bimodal peaks would be classified as conditionally neutral and direct switching, respectively. Although the Luria-Delbrück distribution does not have a closed form, it can be simulated[12-13,20].

Simulations that bias the DFE toward conditionally neutral mutants (most σ = 1) or direct switching mutations (most σ < 1) result in different distributions, which could be used to quantify a protein's ability to generate conditional neutrality (Figure 5.3).

Experimental validation of Luria-Delbrück theory with HiDenSeq

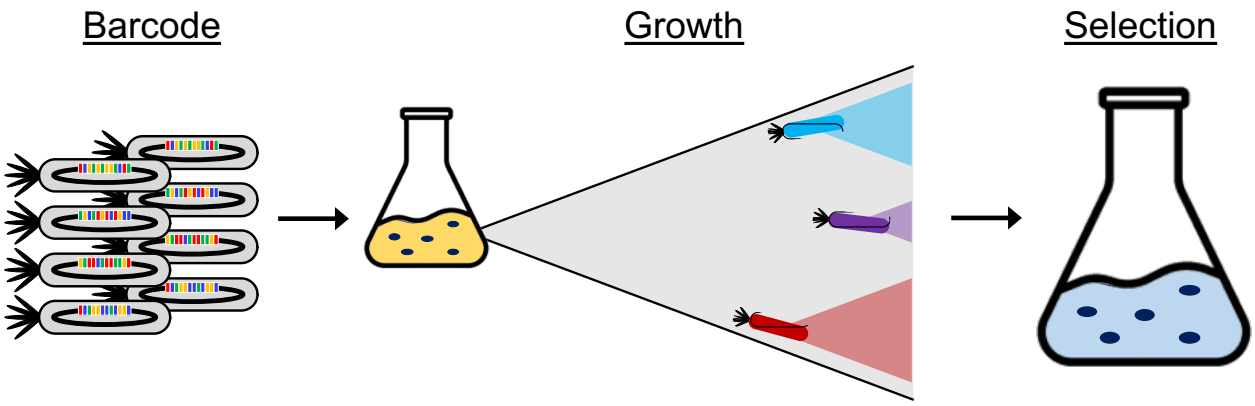

Figure 5.4. High Density Luria-Delbrück by Sequencing (HiDenSeq)
In contrast to a classic Luria-Delbrück experiment, the organism, in this case, an M13 bacteriophage, is barcoded with DNA. Barcoded phages are grown in culture with host bacterial cells, where each barcode represents a unique Luria-Delbrück experiment. Mutations arise during this growth phase, which are then selected for in a subsequent selection growth phase. Importantly, the original phage cannot grow during the selection phase of the experiment. After the selection growth, no phage containing the original genotype will persist at detectable frequencies, and the distribution of frequencies for the barcodes, as determined by Illumina sequencing, is equivalent to the Luria-Delbrück distribution, scaled along the x-axis.

These theories have yet to be tested in an experimental setting, as they require thousands to hundreds of thousands of replicates of the Luria-Delbrück assay; something that is not feasible using traditional approaches. To overcome this limitation, a novel experimental technique called HiDenSeq was developed. The key advancement of HiDenSeq is that, instead of each replicate requiring an independent flask, individual experiments are tagged with a unique DNA barcode. This allows for thousands of replicates to be conducted within a single flask, and after the experiment is completed, the count for the number of functional mutants can be obtained using next-generation sequencing (Figure 5.4). Although this technique can be implemented in various

organisms and systems, this work will focus on an application using the previously described RSP10-PDZ3 phage and its infection of a selection bacterial strain from BTH-PACE (S2060 + AP-CRIPT + DP6). This system is ideal for experimentally verifying the impact of selection pressure on the Luria-Delbrück distribution, as the presence of the DP6 plasmid allows for modulation of selection pressure through the doxycycline-inducible $P_{psp\text{-}tet}$ promoter.

When doxycycline is absent from the system, selection pressure is at its highest, and the distribution should be modeled with $\sigma < 1$. Conversely, as the concentration of doxycycline rises, selection pressure decreases, and σ is expected to approach 1. HiDenSeq experiments were conducted across a range of selection pressures (Figure 5.5 shows two conditions).

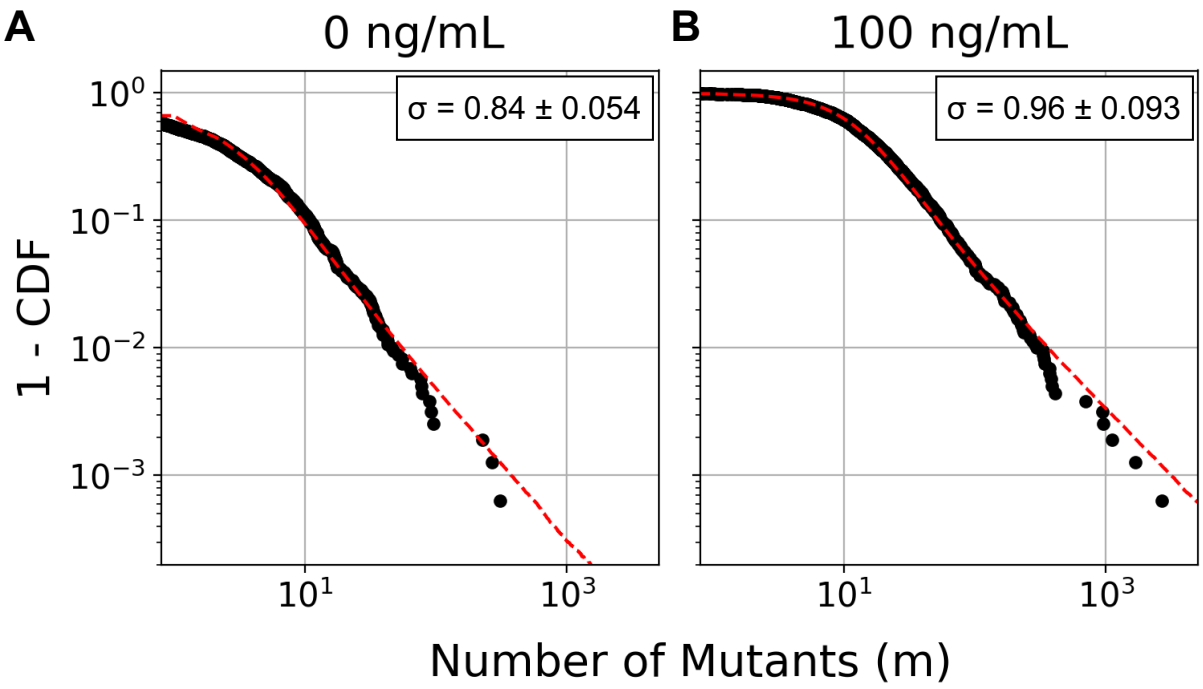

Figure 5.5. HiDenSeq experiments support Luria-Delbrück theory
HiDenSeq experiments were done at a range of selection pressures, two of which, high (A, 0 ng/mL Doxycycline) and low (B, 100 ng/mL Doxycycline) are shown here. Experimental data (black dots) and the simulation data which come from the best fit parameters in Laplace space (red dashed line), are shown. The σ which corresponds to this fit and represents a readout of the selection pressure is given.

To fit the data to a Luria-Delbrück model, two transformations were applied. First, the Luria-Delbrück theory assumes all replicates have the same population size, *N*, at the end of the growth period and before selection[12-13]. In an experimental setting this is impossible due to the distribution of barcode counts in the initial phage population (Figure 5.6). However, the Luria-Delbrück distribution has the convenient property of Lévy-stability[13]. Lévy-stability means for any distribution, *M*, which is defined as $M = m_1 + m_2$, when m_1 and m_2 are both sampled from some distribution, *P(m)*, *M* will be the same distribution as *P(m)*, with only changes to the scaling and location parameters. This means, for the Luria-Delbrück distribution, independent barcodes can have their data combined into a grouped barcode without affecting fitting of σ or μ. The population size, *N*, will be impacted but only because combining two barcodes together combines their populations. We will use a longest-processing-time-first (LPT) algorithm to group barcodes into a predefined number of buckets, each bucket containing 1 or more barcodes which, in total, have an approximately equal total count prior to selection.

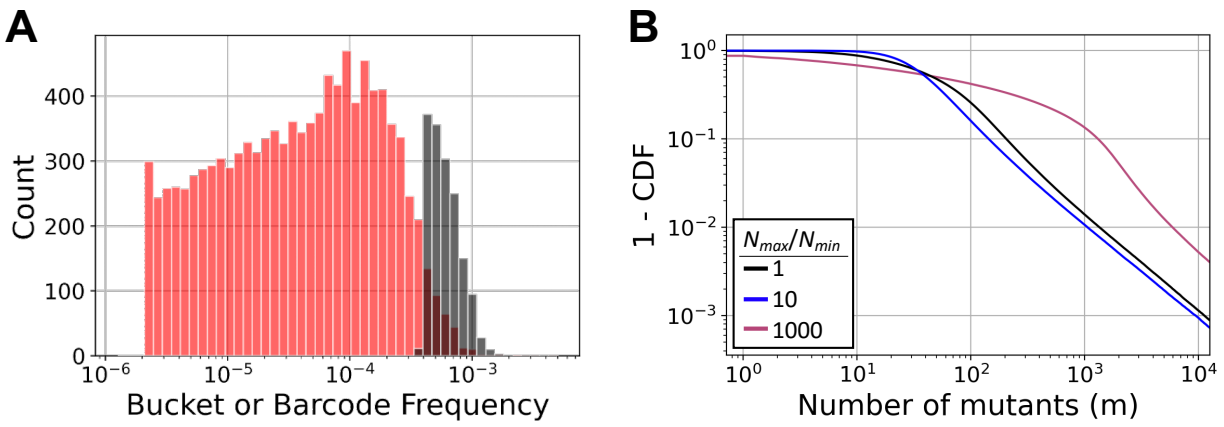

Figure 5.6. Importance of bucketing in interpreting HiDenSeq data

A. The original distribution of barcodes (red) and the distribution of bucketed barcodes using a LPT algorithm (black). Bucket or barcode frequencies are directly proportional to the population size for a HiDenSeq replicate. The LPT algorithm reduces the spread of the input data by at least 100 times.

B. Simulations of a Luria-Delbrück system where σ is 1.0 and the mean of μN is 10. Three different uniform spreads of N are shown. The larger the spread in N as defined by the ratio of the maximum N value to the minimum N value, the less likely the data behave like the expected Luria-Delbrück distribution (black). Spreads of one order of magnitude or less (blue) do not meaningfully affect the data.

The second modification to the data is to transform it into Laplace space. As previously mentioned, the Luria-Delbrück distribution does not have a known closed form. This is not true after a Laplace transform, $ln\tilde{P}(s)$, of the data where *s* is a new complex variable used in Laplace space (Eq. 5.3, Derivations 5.5.2).

$$ln\tilde{P}(s) \approx -\mu N\left(1-\frac{1}{\sigma}E_{1+1/\sigma}(s)\right) \quad \text{(Eq. 5.3)}$$

Equation 5.3 simplifies the fitting complexity by combining the mutation rate, *μ*, and population size, *N*, into a single quantity, *μN*, which can be fit as one independent parameter. This remains valid until m approaches N. A Laplace transform of the data utilizes the special function, $E_p(z)$, known as the generalized exponential integral. For more information on Equation 5.3 see Derivations 5.5.2.

Doxycycline Concentration	σ ± SD (fit)	μN ± SD (fit)	N (experimental)	μ (fspg/gen) (estimated)
0 ng/mL	0.84 ± 0.054	1.09 ± 0.19	4.3×10^5	$2.5 \pm 4 \times 10^{-6}$
10 ng/mL	0.88 ± 0.049	1.02 ± 0.20	4.4×10^5	$2.3 \pm 5 \times 10^{-6}$
25 ng/mL	0.94 ± 0.057	1.55 ± 0.30	6.6×10^5	$2.3 \pm 4 \times 10^{-6}$
50 ng/mL	0.84 ± 0.11	1.27 ± 0.38	5.5×10^5	$2.3 \pm 7 \times 10^{-6}$
75 ng/mL	0.97 ± 0.065	1.62 ± 0.37	6.1×10^5	$2.7 \pm 6 \times 10^{-6}$
100 ng/mL	0.96 ± 0.093	4.44 ± 1.28	2.6×10^6	$1.7 \pm 5 \times 10^{-6}$
150 ng/mL	0.99 ± 0.095	3.76 ± 1.12	2.7×10^6	$1.4 \pm 7 \times 10^{-6}$

Table 5.1. Fit parameters for HiDenSeq experiments.
Values for the fits of σ and μN are determined as outlined after the data are divided into 1,538 buckets. μ is in units of functional mutations per gene per generation (fspg/gen). The experimentally determined N values were obtained by dividing a plaque assay for the total phage culture by the number of buckets.

Least squares fitting was used to find the parameters for the distribution, with initial estimates of *μN* derived from experimental data (Table 5.1). Additionally, since all these experiments were conducted at the same arabinose concentration, dividing *μN* by the experimentally determined N yields an estimate for μ that is consistent across conditions

(Table 5.1). For a true Luria-Delbrück distribution, Lévy-stability also means that fitting with one set of LPT buckets should suffice for modeling data from a different number of buckets. The only alteration to the fitting parameters is that the population size must be adjusted according to the ratio of the change in the number of LPT buckets (Figure 5.7). Both verifications support the notion that the modeled data originate from a Luria-Delbrück process and that the doxycycline-dependent change in σ reflect the changing selection pressure in the system (Figure 5.8).

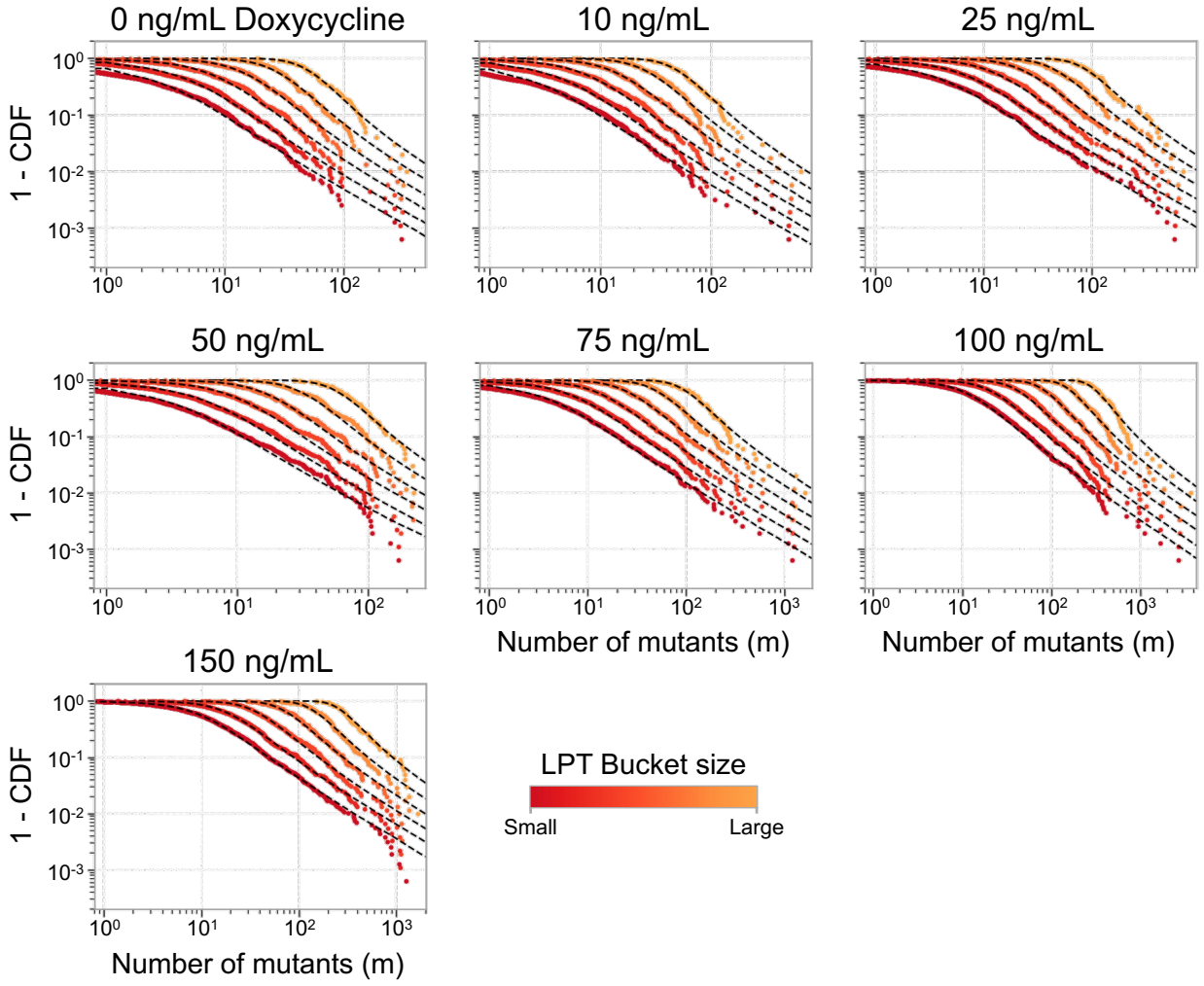

Figure 5.7. HiDenSeq data is Lévy-stable.
HiDenSeq data for all seven doxycycline conditions tested. Plotted are data with different amounts of buckets ranging from 100 (small, red) to 1584 (large, orange). As expected, a single fit of the data corresponding to 1584 buckets generates accurate simulation data for all other buckets if the value for μN is altered by the same factor as the number of buckets.

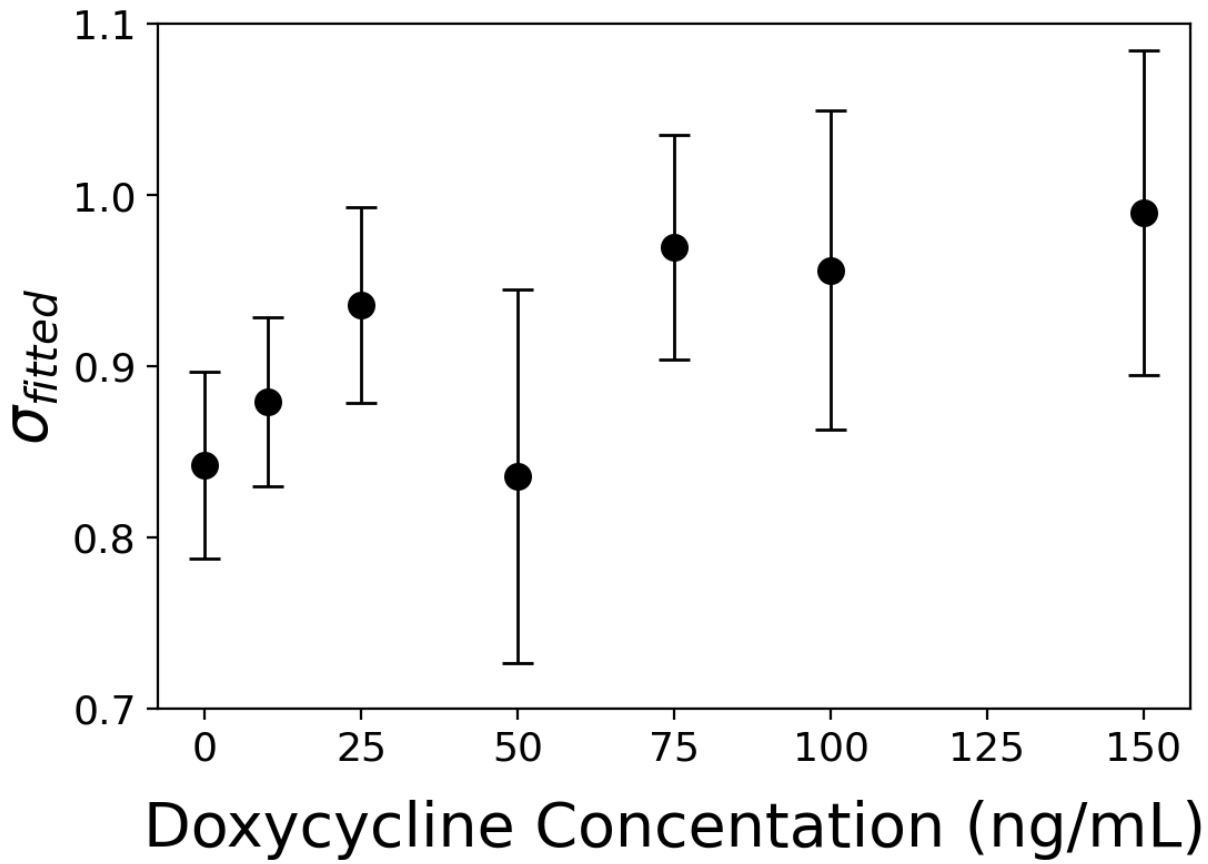

Figure 5.8. Selection pressure alters Luria-Delbrück fit parameters
Shown are the fitted σ values for HiDenSeq experiments run at seven different doxycycline concentrations ranging from high (0 ng/mL) to low (150 ng/mL). The estimated uncertainties are provided.

5.4. Discussion

This work introduces a novel assay for measuring evolutionary parameters. While numerous methods exist for quantifying fundamental protein characteristics such as folding and function, assays specifically designed to assess evolutionary processes are scarce. In this study, we demonstrate that HiDenSeq can effectively quantify selection pressure by evaluating the fitness effects of mutations on a wild-type protein. Increased selection pressure results in a steeper slope in the tail of the Luria-Delbrück distribution (Figure 5.8). Importantly, this finding, which has the benefit of being tailored to the specific system under investigation, is still generalizable. In principle, any microbial system that

experiences a selection condition where the original organism cannot currently grow but has the potential to adapt, can be utilized as a platform for HiDenSeq.

There are also many extensions to be made with this assay that have the potential to greatly increasing the quantitative nature of molecular evolution research. Many of these will be expanded on in Chapter 6 but the one that is the most relevant to this thesis is conditional neutrality. As mentioned, HiDenSeq has the potential to differentiate between conditionally neutral and direct switching mutations (Figure 5.2). If two systems are subjected to the same environment and one protein fits best to a Luria-Delbrück model with a σ value near 1 while another has a sigma value less than 1, it could be said that the first system is better at generating conditional neutrality in the given context.

Furthermore, as protein design algorithms continue to progress, such as those based on evolution like direct coupling analysis (DCA), neural networks like variational autoencoders (VAEs), or and natural language processing (NLP), it becomes crucial to evaluate the adaptive capacity of proteins produced by these processes[22-24]. HiDenSeq, through its ability to quantify CN can be this tool. Without this additional assay it will be difficult to know if the synthetics proteins have fully incorporated all the information contained in natural sequences.

5.5. Derivations

These derivations are the work of Kabir Husain.

5.5.1. Power-law tail slope

The tail of the Luria-Delbrück distribution is dominated by rare events, termed jackpot mutations, that result in a large number of organisms carrying the same functional

mutation. These jackpot mutations must occur early and therefore, they are assumed to be single mutations. This single mutation occurs at time τ, where 0 < τ < T, and T is the total time of the initial growth phase of the experiment. At time τ, the population size, n(τ), is

$$n(\tau) = e^{\lambda_0 \tau} \quad \text{(Eq. 5.4)}$$

Where λ_0 is the growth rate of the wildtype or original variant and μ is the functional mutation rate per organism per generation. And since, the probability that no mutation occurs by some early time, τ, is determined by a Poisson process ($P(m = 0) = e^{-\mu N}$), the probability that a mutation has occurred by time τ is

$$1 - e^{-\mu n(\tau)} \approx \mu e^{\lambda_0 \tau} \quad \text{(Eq. 5.5)}$$

Since τ is monotonically related to the number of functional mutants at the end of the initial growth phase, m, this is also equivalent to the probability that you observe at least m mutants. To express τ in terms of m, we need to introduce another quantity, λ, which is the growth rate of a mutant assuming all mutants have the same growth rate.

$$\lambda(T - \tau) = \ln m \implies \tau = T - \frac{1}{\lambda} \ln m \quad \text{(Eq. 5.6)}$$

And plugging this into equation 5.4

$$\begin{aligned} \text{Prob}(\#\text{ of mutants} > m) &\approx \mu N \exp\left(-\frac{\lambda_0}{\lambda} \ln m\right) \\ &= \frac{\mu N}{m^{\lambda_0/\lambda}} \\ &= \mu N m^{-1/\sigma} \end{aligned} \quad \text{(Eq. 5.7)}$$

Where σ is defined as in equation 5.1 ($\sigma = \lambda/\lambda_0$) and N is defined as the total population of the experiment after the initial growth phase. In log-log space this power-law tail has a slope of -1/σ. The expression is only true for the large m values at the tail of the

distribution, where jackpot events are revealed. However, it only holds until m approaches N as the number of mutants cannot exceed the total population of the experiment.

5.5.2. LaPlace transform of the Luria-Delbrück distribution

Assume all mutants have fitness σ relative to the wildtype population. Additionally, assume an individual experiment starts with n_0 wildtype cells and grows to a population of N cells. Functional mutations (mutations to genotypes which can survive the selective challenge, and which are a subset of all mutations seen during the first growth phase) occur with a probability $\mu \ll 1$ per cell division. X_n, an indicator function, tracks whether a mutation has occurred during each cell division (population goes from n to n+1)

$$X_n = \begin{cases} 0 \text{ with probability } 1-\mu \\ 1 \text{ with probability } \mu \end{cases} \quad \text{(Eq. 5.8)}$$

Each mutation gives rise to a mutant lineage of size Y_n

$$Y_n = \left(\frac{N}{n}\right)^{\sigma} X_n \quad \text{(Eq. 5.9)}$$

Therefore, the total number of mutants, m, produced is

$$m = \sum_{n=n_0}^{N} Y_n \quad \text{(Eq. 5.10)}$$

Then the log of the Laplace transform *P(m)* (equivalent to the negative cumulative generating function) is computed

$$\ln \tilde{P}(s) \equiv \ln \int_0^{\infty} dm\, e^{-sm}\, P(m) = \ln\langle e^{-sm} \rangle$$

$$= \sum_{n=n_0}^{n=N} \ln \left\langle e\, xp\left(-s\left(\frac{N}{n}\right)^{\sigma} X_n\right)\right\rangle$$

$$= \sum_{n=n_0}^{n=N} \ln\left(1-\mu+\mu e^{-s(N/n)^{\sigma}}\right) \quad \text{Eq. (5.11)}$$

Now, a change of variables is done to $x = n/N$ and approximate the sum by an integral, where $x_0 = n_0/N$

$$\ln \tilde{P}(s) \approx N \int_{x_0}^{1} dx \ln\left(1 - \mu + \mu e^{-s/x^{\sigma}}\right) \qquad \text{Eq. (5.12)}$$

If μ ≪ 1, and if *Re(s)* > 0, which they are, the logarithm can be expanded to the first order in μ and the integral can be evaluated

$$\begin{aligned} \ln \tilde{P}(s) &\approx \mu N \int_{x_0}^{1} dx \left(e^{-s/x^{\sigma}} - 1\right) \\ &= -\mu N(1 - x_0) + \mu N \int_{x_0}^{1} dx\, e^{-s/x^{\sigma}} \\ &= -\mu N(1 - x_0) + \frac{\mu N}{\sigma} E_{1+1/\sigma}(s) - \frac{\mu N}{\sigma} E_{1+1/\sigma}(s x_0^{-\sigma}) \qquad \text{Eq. (5.13)} \end{aligned}$$

Where $E_p(z)$ is the special function known as the generalized exponential integral

$$E_p(z) = \frac{1}{z^{p-1}} \int_{z}^{\infty} dt \frac{e^{-t}}{t^p} \qquad \text{Eq. (5.14)}$$

Equation 5.14 is the end of the calculation. However, a few simplifying assumptions can be made. In the first term of equation 5.14, $1 - x_0$ can be approximated as 1. Second, the third term is only relevant when m is on the order of $N^{\sigma} \approx N$. It is responsible for truncating the power tail of the distribution and keeping m less than N. In practice though, m never approaches N and therefore we can simplify equation 5.14 to the following expression.

$$ln\tilde{P}(s) \approx -\mu N \left(1 - \frac{1}{\sigma} E_{1+1/\sigma}(s)\right) \qquad \text{(Eq. 5.15)}$$

5.6. References

1 Payne JL & Wagner A. The causes of evolvability and their evolution. *Nature Reviews Genetics* **20**, 24-38, (2019)

2 Murugan A, *et al*. Roadmap on biology in time varying environments. *Physical Biology* **18**, 041502, (2021).

3 Castle SD, *et al*. Towards an engineering theory of evolution. *Nature Communications* **12**, 3326, (2021).

4 Wagner A. Evolvability-enhancing mutations in the fitness landscapes of an RNA and a protein. *Nature Communications* **14**, 3624, (2023).

5 Wright O, *et al*. Building-in biosafety for synthetic biology. *Microbiology Society* **159**, 1221-1235, (2013).

6 Wright O, *et al*. Building-in biosafety for synthetic biology. *Microbiology Society* **159**, 1221-1235, (2013).

7 Ebrahimi SB & Samanta D. Engineering protein-based therapeutics through structural and chemical design. *Nature Communications* **14**, 2411, (2023).

8 Luria SE & Delbrück M. Mutations of bacteria from virus sensitivity to virus resistance. *Genetics* **28**, 491-511, (1943).

9 Lang GI. Measuring mutation rates using the Luria-Delbrück fluctuation assay. *Methods in Molecular Biology* **1672**, 21-31, (2017).

10 Lea DE & Coulson CA. The distribution of the numbers of mutants in bacterial populations. *Journal of Genetics* **49**, 264-285, (1949).

11 Zheng Q. Progress of a half century in the study of the Luria-Delbrück distribution. *Mathematical Biosciences* **162**, 1-32, (1999).

12 Houchmandzadeh B. General formulation of Luria-Delbrück distribution of the number of mutants. *Physical Review E* **92**, 012719, (2015).

13 Kessler DA & Levine H. Scaling solution in the large population limit of the general asymmetric stochastic Luria-Delbrück evolution process. *Journal of Statistical Physics* **158**, 783-805, (2014).

14 Lang GI & Murray AW. Estimating the per-base-pair mutation rate in Yeast Saccharomyces cerevisiae. *Genetics* **178**, 67-82, (2008).

15 Gou L, *et al*. The genetic basis of mutation rate variation in yeast. *Genetics* **211**, 731-740, (2019).

16 Eyre-Walker A & Keightley PD. The distribution of fitness effects of new mutations. *Nature Reviews Genetics* **8**, 610-618, (2007).

17 Stiffler MA, *et al*. Evolvability as a function of purifying selection in TEM-1 ß-Lactamase. *Cell* **160**, 882-892, (2015).

18 Ogden PJ, *et al*. Comprehensive AAV capsid fitness landscape reveals a viral gene and enables machine-guided design. *Science* **6469**, 1130-1143, (2019).

19 Julia Zinkus-Boltz, *et al*. A phage-assisted continuous selectin approach for deep mutational scanning of protein-protein interactions. *ACS Chemical Biology* **14**, 2757-2767, (2019).

20 Frank SA. The number of neutral mutants in an expanding Luria-Delbrück population is approximately Fréchet. *F1000Research* **4**, 1254, (2022).

21 Nolan, J.P. Univariate stable distributions - models for heavy tailed data. *New York:Springer Verlag*. Chapter 1. (2020).

22 Russ WP, *et al*. An evolution-based model for designing chorismite mutase enzymes. *Science* **369**, 440-445, (2020).

23 Lian X, *et al*. Deep-learning-based design oof synthetic orthologs of SH3 signaling domains. *Cell Systems* **15**, 725-737, (2024).

24 Madani A, *et al*. Large language models generate functional protein sequences across diverse families. *Nature Biotechnology* **41**, 1099-1106, (2023).

Chapter 6. Conclusions

6.1. Discussion

The overall goal of my thesis has been to further understand the relationship between conditional neutrality (CN), its encoding within protein sequences, and its dependence within a protein on a history of changing selection pressures.

In Chapter 3 of my thesis, I took advantage of a of synthetic version of the ligand binding protein $PSD95^{pdz3}$, which had the constraints defining a sector algorithmically separated from the non-sector surroundings. Prior work in the Ranganathan Lab has shown that a stabilized version of this protein, $C2_{34}$-TM, was functional for native ligand binding, thermally stable enough to allow for mutational robustness in the context of binding its native ligand, and able to engage an allosteric network in the form of long-range effects on the active site. However, once a non-native ligand was introduced, $C2_{34}$-TM was shown to be deficient. It bound the non-native ligand with an affinity nearly 10 times worse than the natural protein and, when required to maintain productive genetic variation in BTH-PACE, was completely unable to do so. A prior theory posits that thermal stability buffers evolvability and therefore the issue with $C2_{34}$-TM could be its slightly lowered thermal stability relative to $PSD95^{pdz3,1-3}$. However, $PSD95^{pdz3}$; D357N, which has a similar but slightly lower melting temperature than $C2_{34}$-TM (53°C vs. 58°C), behaves similarly to $PSD95^{pdz3}$ in the context of a non-native ligand. Instead, we propose that $C2_{34}$-TM, which can engage its allosteric network in the context of the native ligand, is not able to access an alternative allosteric network in the context of an alternative ligand (Figure 3.9). The information for this ability is encoded in the non-sector surroundings and

therefore, although thermal stability is necessary for adaptation, further constraints are required to recapitulate natural proteins.

In Chapter 4 of my thesis, I developed a system for continuous evolution, termed BTH-PACE, to investigate the impact of evolutionary history on protein architecture. Six different PDZ domains were evolved in environments where the ligands fluctuated every 72 hours (medium), 144 hours (slow), or remained constant. These varying environments led to distinct mutation patterns, exemplified by the H372P mutation. In the constant environment, where significant alterations to the active site were neither necessary nor tolerated, the H372P mutation was absent. In the slow environment, H372P appeared only when the alternate ligand was present in BTH-PACE but was partially selected against when the native ligand returned. In the medium environment, after sufficient background alterations were acquired, H372P became an adaptation that enabled survival with either ligand. These varying environments induced constraints resulted in proteins with different abilities to bind a panel of novel PDZ ligands, serving as a test of conditional neutrality. Proteins from the medium environment exhibited the most CN, indicating that their internal architectures were altered from the other two environments. While more research is needed, this work offers an early assessment of the impact of evolutionary history on protein design.

In chapter 5 of my thesis, with the assistance of post-doc Kabir Husain, I tested a novel method for quantifying evolutionary parameters, called HiDenSeq. This method leverages technological and experimental advancements to reveal the full Luria-Delbrück distribution generated from the Luria-Delbrück assay. Theoretical extensions to this distribution provide valuable insights, including the ability to quantify the full distribution

of fitness effects (DFE), CN, and the selection pressure within a system. These traits have received little attention in the context of the Luria-Delbrück distribution due to experimental limitations. In our work, we demonstrate that the measured Luria-Delbrück distribution is sensitive to the selection pressure in the system and that HiDenSeq can detect these deviations. While research on the Luria-Delbrück distribution will undoubtedly continue, HiDenSeq offers an exciting and novel opportunity to quantify specific evolutionary parameters in a single experiment.

In summary, this thesis reveals how the interplay between thermal stability, allosteric networks, and evolutionary history have had an essential role in shaping CN. The research also underscores how evolutionary history influences protein architecture, with environments that switch at the pace of new gene fixation producing the most exaptive proteins. Additionally, the development of HiDenSeq introduces a new method for quantifying evolutionary parameters, offering the potential for deeper insights into the dynamics of protein evolution. These findings contribute to a more comprehensive understanding of the constraints encoding evolution within proteins.

6.2. Future Work

6.2.1. EcORep HiDenSeq

The current implementation of HiDenSeq builds on the established BTH-PACE system. In BTH-PACE, as in PACE more broadly, the key to maintaining a well-understood selection condition is the separation of the evolution of the M13 bacteriophage from the host bacterial cells[4]. This is achieved by raising the mutation rate of the bacteria in the lagoon and setting the lagoon's dilution rate high enough so that the bacteria, with

their slower replication rate than the phage, are on average not replicating while in the lagoon. Without this balancing act the evolution of the phage and bacteria become intertwined and straightforward interpretation of data becomes impossible. In HiDenSeq, the phage and bacteria are allowed to grow together for many generations, and evolution where the PDZ domain encoded in the phage evolves for binding to an unchanging and homogeneous ligand is lost.

Perhaps the clearest example of this comes from a deeper sequencing of the HiDenSeq data from Chapter 5. If, instead of just sequencing barcodes, you look at both the barcodes and the full $PSD95^{pdz3}$ PDZ domain linked to each barcode, some surprising mutations begin to appear. Specifically, the H372P mutation in the active site of the protein becomes dominant. Furthermore, if you isolate those barcodes with single mutants for the H372P mutation and look at their HiDenSeq distribution, the estimated σ value which best fits the data is 0.89 (Figure 6.1). If only the CRIPT ligand was present in the bacterial cells this finding, where σ approaches 1, should not be possible and that is backed up by both measurements of the K_D for $PSD95^{pdz3}$; H372P to the CRIPT ligand (K_D = 46.1 ± 8.8 μM) and the fact that H372P mutations do not appear as single mutants in BTH-PACE when selecting for CRIPT ligand binding (Figure 3.2.A, Figure 4.7). I hypothesize that this difference in HiDenSeq comes from the fact that, due to the long incubation times of bacteria and phage without dilution, the bacteria encoded ligand is also allowed to mutate. Although the examination of selection pressure in HiDenSeq that was tested in this work is not affected by this complication, statements about conditional neutrality become impossible as the selection condition is not well defined.

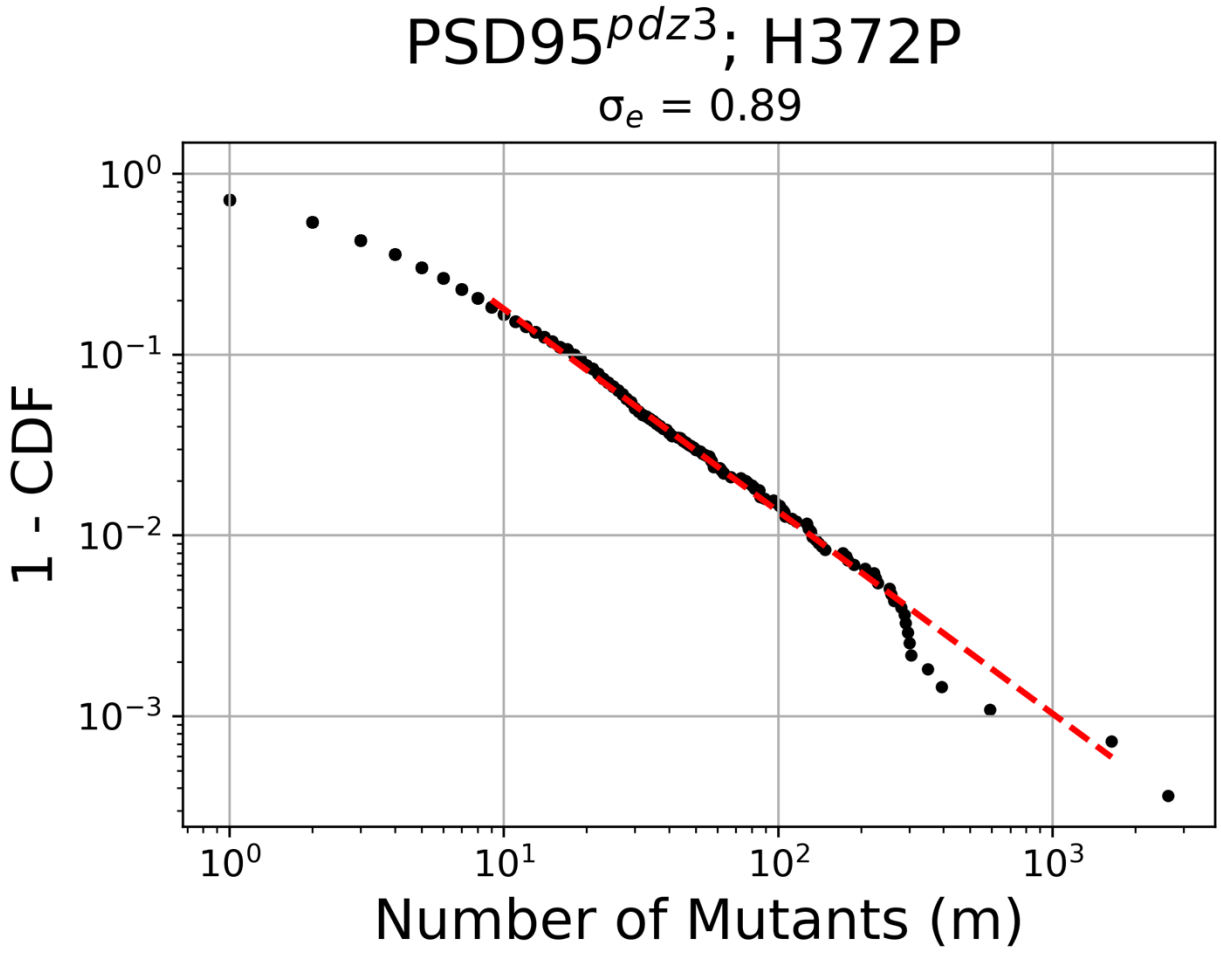

Figure 6.1. PSD95^{pdz3}; H372P in HiDenSeq
A subset of the data from a HiDenSeq experiment of PSD95^{pdz3} assaying for CRIPT to T-2F ligand binding at 100 ng/mL doxycycline restricted to the H372P mutation. Data (black dots) are overlayed with an estimate of σ (σ_e = 0.89) that comes from a linear fitting of the tail of the distribution (red dashed line). See Derivations 5.5.1 for an explanation of this approximation.

A solution to this issue is to perform HiDenSeq such that the evolving PDZ domain is truly separated from its ligand. A recently developed semi-continuous evolution system called *E. coli* orthogonal replication system (EcORep) allows for this[5]. Simply, EcORep utilizes of a novel orthogonal replication system in *E. coli* where an orthogonal replicon is replicated by a DNA Polymerase that does not replicate the bacterial genome. This DNA polymerase has a much higher mutation rate (7.6 x10^{-6} spb/gen) than the genomic DNA polymerase. The replicon can contain cargo as large as 16.5 kilobases which is more

than sufficient to encode the half of the bacterial-two-hybrid containing the PDZ domain. The rest of the bacterial-two-hybrid system, including a ligand of interest, and some means of selection, could then be transformed into *E. coli* on another plasmid that is replicated by the normal genomic DNA polymerase. To date, a bacterial-two-hybrid system has not been modified for use in EcORep, but its implementation should be straightforward given the extensive work with similar assays in the Ranganathan lab and others.

As with any new experimental system, some care must be taken to account for the intricacies of EcORep relative to BTH-PACE. Most importantly, the mutation rate in EcORep is, presently, about an order of magnitude lower than what is currently used in HiDenSeq (~8.7 x 10^{-5} spb/gen). The Luria-Delbrück distribution is sensitive to the mutation rate and a lower rate will shift the entire distribution to the left on the x-axis, decreasing the chance of seeing larger m values. This, combined with the low transformation efficiency of the orthogonal replicon (1,824 transformants per µg DNA), lowers the number of independent replicates of HiDenSeq. Consequently, capturing the tail of the distribution could be challenging. Either increasing the mutation rate of the orthogonal DNA polymerase or increasing the transformation efficiency of the orthogonal replicon could solve this limitation.

EcORep-HiDenSeq presents another minor concern, there is currently no method for modulating the selection pressure in EcORep. Altering the copy number of the plasmid can be used to approximate changes in selection pressure, but for HiDenSeq analysis is simplest if the copy number is kept as low as possible. Instead, like PACE, a plasmid would likely need to be introduced to supplement *E. coli* growth independently of PDZ-

ligand binding. Once these issues are addressed, EcORep-HiDenSeq could proceed similarly to what is described in the Methods Section 7.4, with only small changes due to the differing biology of the two systems.

6.2.2. Theoretical Extensions to the Luria-Delbrück Distribution

Presented in this work was a verification to the extension of the classic Luria-Delbrück theory that allowed selection pressure during the growth phase of a Luria-Delbrück experiment. However, other data were also collected that, presently, cannot be interpreted without additional extensions. As shown in Chapter 3, the $C2_{34}$ and $C2_{34}$-TM PDZ domains are unable to adapt to binding a non-native ligand, likely because of an inability to engage their allosteric network. Furthermore, current data suggests that these proteins, when assayed for $T_{-2}F$ ligand binding, are unable to maintain productive SGV which is statistically likely to take the form of CN mutants. A HiDenSeq experiment was conducted to see if the Luria-Delbrück distribution generated by these proteins would be modeled by a σ less than 1 as would be consistent if they could not produce CN mutants.

The experimental data have three unexpected results (Figure 6.2). First, as explained in 6.2.1, interpretations of CN are impossible given the ill defined selection function that exists in this implementation of HiDenSeq. Secondly, and unlike results for $PSD95^{pdz3}$, the majority of the $C2_{34}$ and $C2_{34}$-TM domains had more than one mutation. This is not accounted for in any current Luria-Delbrück theory. Third, the distribution did not fit simulations for a single mode of mutational effects (see Figure 5.2.C for an example). Instead, and again in contrast to $PSD95^{pdz3}$, there is a distinct biphasic nature

to the distribution, indicative of a more complicated DFE (see Figure 5.3.C for an example).

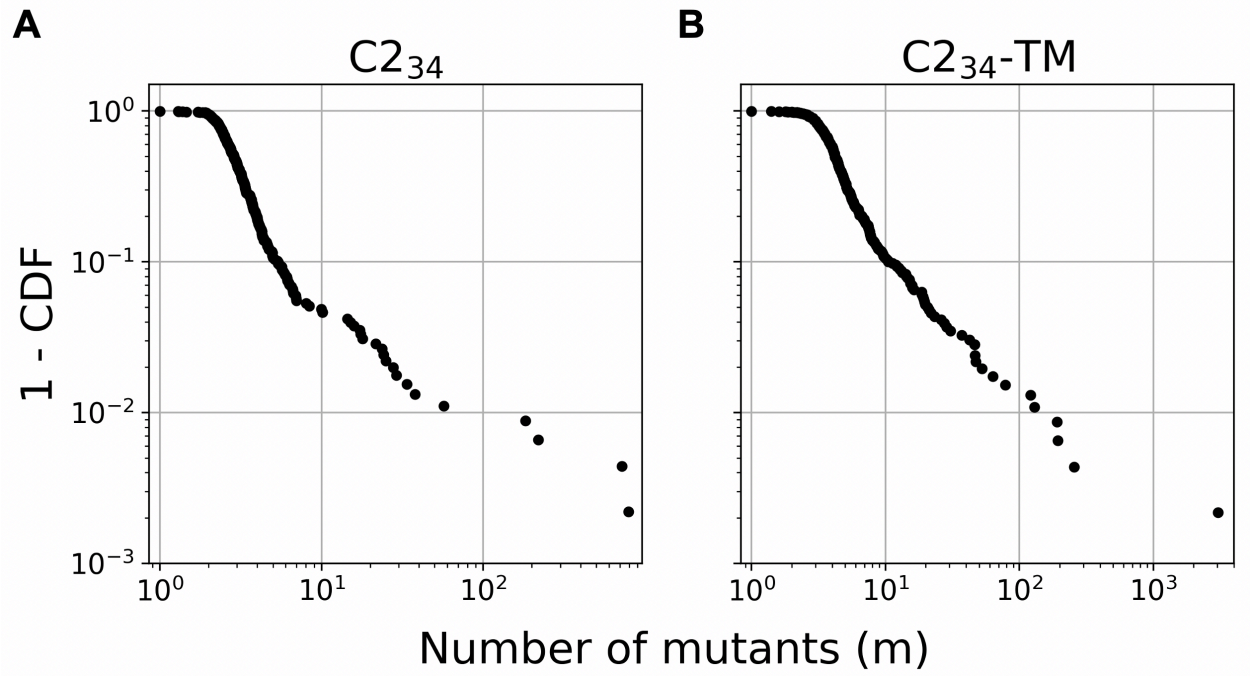

Figure 6.2. HiDenSeq of $C2_{34}$ and $C2_{34}$-TM
HiDenSeq of $C2_{34}$ (A) and $C2_{34}$-TM (B) assaying for CRIPT to $T_{-2}F$ ligand binding at 100 ng/mL doxycycline. Plotted are all mutants, including those with more than 1 mutant to the wildtype protein.

These unexpected results, although intertwined, need to be understood before a complete interpretation of the data from a HiDenSeq experiment done on $C2_{34}$ or $C2_{34}$-TM is possible. The result without a clear path to a solution is finding number two, the existence of multiple mutations in a single PDZ domain. One possible route to modeling this phenomenon is to rewrite the negative cumulative generating function (Laplace transform) to account for additional acquired mutations after the original mutation. Logically, if a mutation occurs at time τ, this is equivalent to starting a new Luria-Delbrück experiment at time τ which then grows until the total population of it, its subsequent mutants, and the original variant reach a total population size of N. Therefore, it seems worthwhile to develop a Laplace transform where the Luria-Delbrück distribution is nested

within the parent distribution for each new mutation. Similarly, simulations could be written and tested where a nested simulation is started each time a new mutation occurs.

This strategy, for it to prove useful requires simplifying assumptions to be placed on the σ values used to fit data. In its purest form, each new mutation introduces a whole new set of potential σ values for the new Luria-Delbrück distribution and fitting in this manor will quickly prove intractable. One possible simplification assumes σ can only take three values across all mutations – one for direct switching ($\sigma < 1$), one for perfectly conditionally neutral mutations ($\sigma = 1$), and one for gain-of-function mutations ($\sigma > 1$). Another idea sets σ relative to the parent strain instead of the original wildtype strain. This simplifies the number of parameters that need to be fit if the relative σ is assumed to be constant. These ideas are untested and not exhaustive but should provide a jumping off point for future work.

6.2.3. Further characterization of the impact of evolutionary history

This work has tentatively shown that even a brief period of evolution can leave an impact on the architecture of a protein. Mutations that are relevant and beneficial in one environment, such as H372P in the intermediate fluctuation regime, become conditionally viable in a slow environment, or non-viable when the environment is static. This changing pattern of mutations leads to a difference in the ability to bind novel ligands. However, a more comprehensive analysis should be done to confirm this finding. In this work, only three environmental conditions were tested, including the extreme of no environmental fluctuation. Two additional fluctuation rates should assayed: 24 hour fluctuation (environmental fluctuation faster than the time to new mutant fixation), and the constant

presence of both the CRIPT and T-2F ligand (a control for the upper limit of “fluctuation”). With these two additional conditions, the work would sufficiently span the entire scope of fluctuation regimes possible.

With these environments rates tested, additional post BTH-PACE work should also be done. Assuming EcORep-HiDenSeq is viable, quantification of the ability to produce CN mutants should be correlated to any changes in the architecture of a protein. Successful EcORep-HiDenSeq experiments require a ligand that the proteins have not been exposed to and cannot currently bind but also where subsequent point mutations to the protein allow for binding to that ligand. Work which resulted in Figure 4.8 indicates it is unlikely that a single novel ligand will be both unable to bind PDZ domains generated from all conditions and still have functional single mutants. These are the two criteria for a working HiDenSeq assay. Instead, unique ligands will need to be identified for each protein evolved in BTH-PACE that can work in HiDenSeq. After EcORep-HiDenSeq is completed, the distributions should be parametrized and experimentally derived σ values compared. It is my hypothesis that proteins from a medium fluctuation environment, which are required to continually adapt to new conditions but still able to alter their genotypes between fluctuations, will have σ values closest to perfect conditional neutrality (σ = 1). Furthermore, if the data fit a bimodal DFE, then I would expect the medium fluctuation environment to have the largest fraction of its mutants approximately conditionally neutral.

6.2.4. Evolutionary viability of a synthetic variant of PDZ

This work tested the capacity for evolution using synthetic variants of $PSD95^{pdz3}$ in the form of $C2_{34}$ and $C2_{34}$-TM. These proteins were designed by systematically altering

the existing architecture by retaining only sector constraints while scrambling the non-sector surroundings. Consequently, they do not behave like natural proteins when subjected to non-native functional challenges. More complex algorithms for protein design exist such as those that rely on the statistics of direct coupling analysis (DCA) or neural networks (variational auto encoders amongst others)[6,7]. These approaches aim to capture the totality of constraints placed upon proteins and recapitulate them in completely novel protein sequences. Recent work by Xinran Lian and Nikša Praljak demonstrated that by using one such method for protein design, termed InfoVAE, synthetic orthologs of the small ligand binding protein, SH3, could be made (Figure 6.3.A)[7]. Furthermore, when these orthologs were mapped into a low-dimensional latent space, those proteins which localized near the SH3 domain Sho1 had ligand binding function for the Sho1 ligand 43.9% of the time; a number comparable to natural orthologs (Figure 6.3.B).

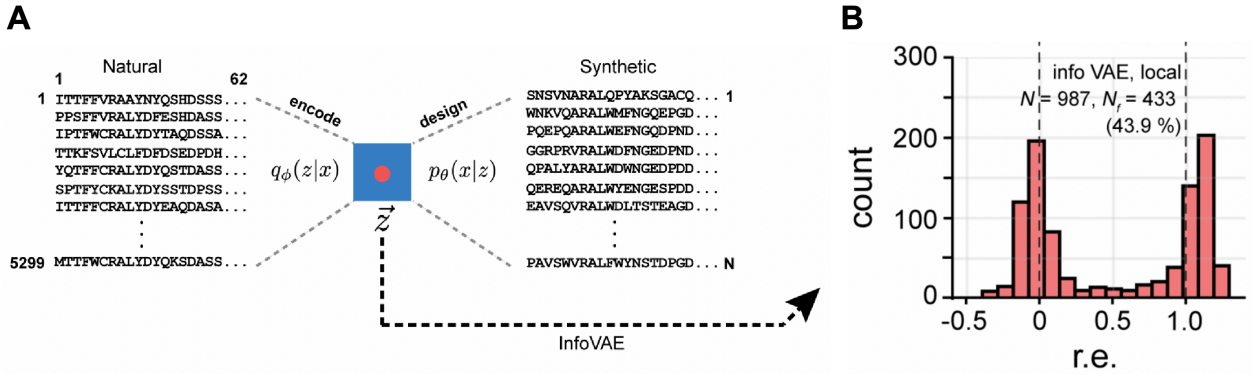

Figure 6.3. InfoVAE design of SH3 orthologs

A. *Adapted from Lian X, et al., 2024[7], Figure 1C. Schematic of evolutionary-based data-driven generative models, consisting of a compression step (encoding) that maps a sequence alignment of natural homologs to a low-dimensional Gaussian latent space (blue box), defined by vector $\vec{z}$ for each sequence, and a decoder that converts latent space coordinates to synthetic protein sequences. By definition, a VAE is trained to reproduce its inputs; thus, decoded sequences represent hypotheses for synthetic members of the protein family.*

B. *Adapted from Lian X et al., 2024, Figure 4F. Distribution of normalized relative enrichment scores measured by a high-throughput selection assay for the 987 local InfoVAE designed SH3 domains. Locality is defined by a sampling of the InfoVAE latent space near to $SHO1^{SH3}$ and its orthologs. 43.9% of these designed sequences showed natural like function, a number comparable to natural proteins localized to the same region in the InfoVAE latent space.*

What remains to be tested for these, or any, synthetic proteins, is if they contain any capacity for adaptation. The structure of the latent space for InfoVAE SH3 domains is organized by both functionality and phylogeny. This indicates that some knowledge of evolutionary history has been retained by the model but does not necessarily mean the proteins are adaptable. Proteins very near to each other in latent space can still be tens of mutations away in sequence space, a gap that may not be traversable in evolution. To date testing these and other proteins for adaptability has been difficult due to a lack of experimental tools and metrics for quantifying and conducting evolution. With the advancements made in BTH-PACE it should be possible to compete natural SH3 domains with their synthetic orthologs for binding to a variety of ligands, both constant and changing. As was seen in Chapter 4 of this work, natural PDZ domains can adapt to changing conditions of selection; whether or not synthetic proteins have a similar capacity remains an open question. Lastly, once EcORep-HiDenSeq is working, the presence of conditional neutrality in synthetic variants for various ligands can be quantified and compared to their natural counterparts. These two experiments would provide an essential test of how far the biological community has come in designing truly natural-like proteins.

6.3. References

1 Bloom JD, *et al*. Protein stability promotes evolvability. *PNAS* **103**, 5869-5874, (2006).

2 Tokuriki N & Tawfik DS. Stability effects of mutations and protein evolvability. *Current Opinion in Structural Biology* **19**, 596-604, (2009).

3 Wagner A. Robustness, evolvability, and neutrality. *FEBS Letters* **579**, 1873-3468, (2015).

4 Esvelt KM, *et al*. A system for the continuous directed evolution of biomolecules. *Nature* **28**, 499-503, (2011).

5 Tian R, *et al*. Establishing a synthetic orthogonal replication system enables accelerated evolution in *E. coli*. *Science* **383**, 421-426, (2024).

6 Russ WP, *et al*. An evolution-based model for designing chorismite mutase enzymes. *Science* **369**, 440-445, (2020).

7 Lian X, *et al*. Deep-learning-based design oof synthetic orthologs of SH3 signaling domains. *Cell Systems* **15**, 725-737, (2024).

Chapter 7. Methods

7.1. Bacterial-two-hybrid phage assisted continuous evolution (B2H-PACE)

This is a general method for B2H-PACE which is used many times throughout this work. For specific variants of B2H-PACE, see section 7.2.

M13 bacteriophage strains

All viruses are derived from the PDZ3-RSP10 strain made by BoRam Lee. RSP10 is a modification of the SP098 M13 bacteriophage. RSP10 has the essential gene, gene III knocked out. The gene III promoter however, is still used to express the RNA Polymerase ω subunit linked to a $PSD95^{pdz3}$ domain. Gibson assembly was used to replace the $PSD95^{pdz3}$ protein with other PDZ domains ($C2_{34}$, $C2_{34}$-TM, $SAP102^{pdz3}$, $Magi\text{-}3^{pdz1}$, $Pdzk3^{pdz1}$, $Magi\text{-}2^{pdz2}$, and $Chapsyin\text{-}110^{pdz3}$). Transformation of assembled phage genomes was done in electrocompetent S2208 cells (described below) with a two hour recovery time in 2xYT media. Phage were then purified by centrifugation at 11,000xg for 4 minutes and passage of the supernatant through a 0.2 μM filter. Supernatant containing isolated phage was stored at 4°C.

Bacterial strains

All bacterial cells are derived from the *E. coli* strain S2060. Non-selection bacterial cells (S2208) carry the plasmid $P_{psp\text{-}tet}$-gene III (pJC175h) which initiates gene III production upon any phage infection. The original selection strain (S2060 + AP-CRIPT) as developed by Boram Lee contained a version of the accessory plasmid (AP-CRIPT; pAB076i3-CRIPT) which expresses the PDZ ligand, CRIPT, linked to a DNA binding

domain (bacteriophage 434 repressor protein CI). QuikChange mutagenesis was used to generate the AP-$T_{-2}F$ from AP-CRIPT. Accessory plasmids with other ligands were created through overlap extension PCR. Cells that allowed control over the mutation rate and the selection pressure contained the DP6 plasmid in addition to the plasmids previously described (selection; S2060 + AP-CRIPT + DP6, non-selection; S2208 + DP6). DP6 contains a constitutive promoter for titratable arabinose induction of the P_{BAD} promoter expressing mutator genes and a doxycycline inducible promoter, $P_{psp\text{-}tet}$, for gene III production.

Generalized BTH-PACE

Turbidostatic growth at 37°C, in 200 mLs Davis Rich Media + 1 mL/L Tween-80, is initiated from 500 µLs bacterial cells of a 3 mL overnight culture grown in 2xYT media at 37°C. Liquid from the turbidostatic culture is flowed to lagoons at a rate of 12 mL/hr. Lagoons are where phage evolve during BTH-PACE and have a volume of 6 mLs meaning there is a dilution rate of 2 times per hour. The lagoons are also supplied with a constant flow of arabinose and doxycycline from an external source when required. These control the mutation rate and selection pressure respectively. If there is increased mutagenesis in the system the concentration of arabinose in the lagoon is 25 mM. The concentration of doxycycline varies from 0 ng/mL to 150 ng/mL in the lagoon. When doxycycline is present, the whole system is shielded from light to protect from degradation.

After an overnight equilibration of the system, 5×10^8 phage are added to each lagoon. Samples are taken at predefined times and the phage are purified through

centrifugation as described previously. The turbidostat, lagoons, and tubing connecting the turbidostat to the lagoons is exchanged every two days to prevent biofilm formation. A change to a fresh turbidostat, started from a new overnight culture of bacteria grown at 37°C, is done every 4 days for the same reason. Waste from the turbidostat and lagoons flows into a bucket containing 10% bleach. Phage titers throughout the experiment are monitored through colorimetric plaque assay for infection of S2208 bacterial cells using 0.001 g/mL X-gal in 2xYT + 4 g/L top agar and 2xYT + 18 g/L bottom agar.

7.2. Variants of B2H-PACE seen throughout thesis

Unless otherwise stated, all versions of BTH-PACE described in this section are exactly as described above. Variations in length, initial condition, mutation rate, and selection are described.

Competitive stress BTH-PACE

Each lagoon is initiated with an expected equal proportion of M13 bacteriophage containing one of the $PSD95^{pdz3}$, $C2_{34}$, or $C2_{34}$-TM PDZ domains. In total the starting phage population is 5×10^8. The bacterial strain used AP-CRIPT + DP6. The mutation rate is set to 8.7×10^{-5} substitutions per basepair per generation by 25 mM arabinose in the lagoons. To not wash out the phage, the selection pressure is lowered through the addition of 10 ng/mL doxycycline. The experiment was conducted for 120 hours corresponding to roughly 120 generations.

BTH-PACE to assay standing genetic variation

Each lagoon is initiated with 5×10^8 M13 bacteriophage containing only one of $PSD95^{pdz3}$, $PSD95^{pdz3}$; D357N, $C2_{34}$, or $C2_{34}$-TM PDZ domains. The bacterial strain used

is used is AP-CRIPT + DP6 for selection and S2208 + DP6 for non-selection. In both cases the mutation rate is set to 8.7 x 10^{-5} substitutions per basepair per generation by 25 mM arabinose in the lagoons. There is no doxycycline. The experiment was conducted for 6 hours corresponding to roughly 6 generations. Total phage was quantified using a colorimetric plaque assay where S2208 bacterial cells were infected. The frequency of $T_{-2}F$ functional phage was quantified by first propagating the sample for 8 hours in S2208 bacteria. Then the propagated sample is assayed for its total phage and $T_{-2}F$ functional phage using a colorimetric plaque assay of S2208 and AP-$T_{-2}F$ bacterial cells respectively. The propagation step is needed to increase the resolution of the experiment.

Fluctuating environment BTH-PACE

In every condition, lagoons are initiated with 5 x 10^8 M13 bacteriophage containing a single PDZ domain (PSD95^{pdz3}, SAP102^{pdz3}, Magi-3^{pdz1}, Pdzk3^{pdz1}, Magi-2^{pdz2}, or Chapsyin-110^{pdz3}). The initial bacterial strain used is used is AP-CRIPT + DP6 and this is switched to and from AP-$T_{-2}F$ + DP6 at the defined rate. The mutation rate is set to 8.7 x 10^{-5} substitutions per basepair per generation by 25 mM arabinose in the lagoons. To not wash out the phage, the selection pressure is lowered through the addition of 10 ng/mL doxycycline. The experiment was conducted for 504 hours corresponding to roughly 504 generations.

7.3. Growth rate determination of PDZ variants

Phage containing PDZ domains of interest were isolated from colonies seen on a plaque assay and verified by Sanger sequencing. The colonies were propagated for 8

hours in 2xYT containing a 1:1000 dilution of an overnight culture (37°C) of S2208 bacterial cells that were allowed to recover for 1 hour at 37°C before phage were added. After 8 hours of propagation the phage was purified and their PDZ domains verified with Sanger sequencing. The phage titer was then assayed with a colorimetric plaque assay using S2208 cells.

Overnight cultures (37°C) of bacteria were diluted 1:1000 in 5 mL 2xYT and allowed to recover for 1 hour at 37°C. For each phage containing a unique PDZ domain, 15 different AP bacteria were used each containing a different ligand. After 1 hour of recovery, phage was added to create a population of 200 phage/μL. The phage and bacteria were incubated for 3.5 further hours. 1 mL of phage was purified after 2 and 3.5 hours. The phage titer at each timepoint was determined with a colorimetric plaque assay using S2208 cells. The growth rate was determined by calculating for r in the equation $N(t) = N_0 e^{rt}$, where *N* is the population size and *t* is time.

7.4. High Density Luria-Delbrück by Sequencing (HiDenSeq)

For HiDenSeq, all overnight cultures, bacterial growth, and phage propagation was done at 37°C. A 16-nucleotide barcode was added to PDZ3-RSP10 phage strains using overlap extension PCR. A 1 mL overnight culture of AP-CRIPT+DP6 cells was diluted 1:200 and allowed to recover for 2 hours in 2xYT media. After 2 hours 10^6 phage, corresponding to 10^5 unique barcodes were added. Additionally, the culture was now made to have 25 mM arabinose and a doxycycline concentration ranging from 0 to 150 ng/mL. After 2.5 hours of growth the entire phage sample was purified as previously described. This process was repeated 2-3 times, until the total phage population reaches

10^{10}, where the phage added are the purified phage from the previous growth. The volume of bacteria for each growth will increase to equal the volume of phage previously purified. A 60 µL sample is withheld after each purification to monitor phage titers.

Once the phage population is 10^{10}, the entire population is added to a 100 mL culture of AP-T_{-2}F bacterial cells that have been allowed to recover for 2 hours in 2xYT after a 1:250 dilution from an overnight culture. The phage is propagated for 3 hours, a 500 µL sample of phage is purified, and the 100 mL culture is spun down at 4000 xg for 15 minutes at room temperature. The supernatant is retained and immediately added to a second 100 mL culture of AP-T_{-2}F bacterial cells that have been allowed to recover for 2 hours in 2xYT after a 1:250 dilution from an overnight culture creating a 200 mL culture. This is repeated two more times resulting in 800 mLs of phage and bacteria and 12 hours of total phage propagation time. At this point, a 1 mL phage sample is taken. This 1 mL sample is added to a 1 L culture of AP-T_{-2}F cells that are diluted 1:10,000 in 2xYT and allowed to recover for 1 hour. The phage is propagated for 12 hours with 1 mL samples being purified every 3 hours.

7.5. Illumina sequencing preparation

Samples used for next generation sequencing on Illumina MiSeq, NextSeq 550, NextSeq 2000, or NovaSeq 6000 machines were prepared in the same way. An initial PCR with Q5 DNA polymerase and a 2.5% DMSO spike-in involving 10 cycles (5 min|98°C hot start; 10 cycles of 18 sec|98°C, 18 sec|58°C, and 18 sec|72 °C; 2 min|72°C; hold|10°C) was used to excise the PDZ domain, and barcode when relevant, from the phage. This step also introduced nucleotide diversity and added on the read 1 and read

2 sequencing primer binding sites. A second round of PCR with Q5 DNA polymerase involving 25 cycles (30 sec|98°C hot start; 25 cycles of 18 sec|98°C, 18 sec|58°C, and 18 sec|72 °C; 2 min|72°C; hold|10°C) was used to add indices for multiplexed samples and add on the P5/P7 clustering sequences. For information on the specifics of the sequences, see Illumina's website. Samples were quantified with Qubit assay and then loaded at concentrations instructed by Illumina documentation. Paired end sequencing runs were used and a 25% PhiX spike-in was included.

7.6. Processing of Illumina sequencing data

FASTQ data from each Illumina run was processed with custom Julia code (Methods 7.8.1). FLASh was first used to merge paired end reads. Merged sequencing files were then trimmed to examine only the region of interest. Trimmed reads were then assessed for data quality using a rolling window average of the Phred scores within a given read. If at any point the average Phred score in a rolling window (length of 4) dropped below 20 for a given read, that read was removed from future analysis. In addition, for every read that passed this check, an error rate was found corresponding to the percent chance of at least one error in the sequence. All reads were then grouped by their sequence and stored as a dictionary of counts. Using the previously stored error rates, and a false discovery rate of 1%, sequences of count K were accepted or rejected based on if K was significant in a binomial distribution with a probability equal to the determined error rate and the number of trials equaling the total number of reads for the most prevalent sequence in the sample.

7.7. Luria-Delbrück data processing

Data processing for Luria-Delbrück sequencing were analyzed as described above through the step of sequence trimming. Sequence trimming leaves only a 16-nucleotide barcode and an associated PDZ domain per read. Barcodes are removed from further analysis if a single nucleotide has a Phred score below 30. All barcodes were then grouped by their sequence and stored as a dictionary of counts. Barcodes with a count below 3 are dropped. Using a LPT algorithm, barcodes from before selection are grouped into a predefined number of "machines" to standardize the count per machine prior to selection. This same grouping is used post selection. However, current theory only allowed for single mutations to the starting PDZ domain and therefore barcodes associated with a PDZ domain which had more than one mutation were not included in the counts after selection. Data are fit in Laplace space by least squares fitting to a model which has the free parameters μN (functional mutation rate times population size), σ (differential growth rate), and α (x-axis scaling factor). Uncertainty in μN and σ is determined by finding the curvature around the minimized value, MSD_0, of the mean squared deviations for either parameter, p, through fitting of a second order Taylor expansion.

$$\frac{1}{MSD_0} MSD(p) \approx 1 + \frac{H}{MSD_0}(p - p_0)^2 \qquad \text{(Eq. 7.1)}$$

Then, as the covariance matrix can be approximated by the inverse of the Hessian curvature matrix, the following quantity should approximate the standard deviation of the measurement.

$$uncertainty\ in\ p = \sqrt{\frac{MSD_0}{H}} \qquad \text{(Eq. 7.2)}$$

7.8. Code of significant importance to this work

7.8.1. Conversion of Illumina .fastq files to dictionaries of nucleotide counts

This code was run with Julia version 1.7.2. It will turn .fastq files from an Illumina sequencing run into .csv files which contain dictionaries of counts for unique nucleotide sequences. It also trims reads down to a region of interest and does error checking and removes untrustworthy reads. The entire code was saved into a Julia file named NGS_part_1.jl and is run from the terminal. For this code to work the following file structure must be in place.

```
> Home directory
      > NGS_part_1.jl
      > flash
      > FASTQ_files
            > #ID_FwdPrimer_RevPrimer_S#_Lane#_Read#_001.fastq
```

NGS_part_1.jl is the Julia code to be executed. “flash” is a Unix executable from John Hopkins University Center for Computational Biology (v1.2.11). It is used to merge read 1 and read 2 files for paired-end Illumina sequencing runs. FASTQ_files, is where the unzipped .fastq files from the Illumina sequencing run are stored. Names for the .fastq files must follow the naming structure above with special care given to the placement of the underscores (“_”). With the exception of “#ID”, which is a unique identification number (or a number that is shared between a set of paired-end sequencing files) that must be added to each file, all of this should be in place in the raw Illumina output files. NGS_part_1.jl uses regular expressions to split the filename based on the underscores to properly run.

The output data will be stored in a folder that is specified when the code is run. It has the following structure.

```
> Output folder
```

```
> Flash_Output
> Merged
> nt_seqs
> QS_scores
> Raw_nt_counts
```

All files will begin with the #ID at the beginning of the .fastq files. Flash_output is a folder that stores all the output from the flash executable that is not the merged file. Merged is a folder that stores the merged sequence files (ex: 1_merged.fastq). nt_seqs is a folder which stores only reads which can be trimmed and pass error checking. If the analysis is being done in chunks, there may be multiple files per unique #ID separated by ~K, where K is Kth chunk (ex: 1~4_nt_seqs.fastq, where the #ID is 1, and this is the 4th chunk of analysis for this #ID). QS_scores is a folder which stores the probability of that a read has no errors for each read that can be trimmed and pass error checking. File naming is the same as for the nt_seqs folder (ex: 1~4_QS_scores.fastq, where the #ID is 1, and this is the 4th chunk of analysis for this #ID). Raw_nt_counts contains two file types. The first file type is a .csv file which, per #ID, contains all the nucleotide sequences for all chunks of that #ID (ex: 1_raw.csv). The second file type is a .csv file which, per #ID, is a dictionary where the entries are unique nucleotide sequences, and the counts are the number of times that sequence is in the data (ex: 1.csv).

An example of a terminal command entry is shown below.

```
julia ./NGS_part_1.jl FASTQ_files nextseq -m -M 210 -b CTTCTAGA -e CAAGTCCT -Q
trimmomatic -c 0.99 -H 1000000 -o NGS_Analysis_Output
```

In order of flags and options, this code will:

- FASTQ_files: analyze .fastq files from the FASTQ_files folder.
- nextseq: analyze the data from a Nextseq2000 Illumina sequencing run.

- -m: merge files from a paired end sequencing run.
- -M 210: set the maximum overlap of merged sequences for the flash Unix executable to be 210 nucleotides.
- -b CTTCTAGA: look for a region of interest, after merging, beginning just after the nucleotide sequence specified.
- -e CAAGTCCT: look for a region of interest, after merging, ending just before the nucleotide sequence specified.
- -Q trimmomatic: use the trimmomatic method for determining if a read should be kept for further analysis.
- -c 0.99: The sliding window in the trimmomatic error checking method cannot fall below a 99% confidence that it is correct or the whole read is removed from the analysis pipeline.
- -H 1000000: For large files, the workflow is split into chunks of 1 million reads. This speeds up the analysis.
- -o NGS_Analysis_Output: Where the output data is stored.

The code for the NGS_part_1.jl file.

```
using ArgParse
using DataFrames
using CSV
using StringDistances
using StatsBase

function parse_commandline()
    s = ArgParseSettings()

    @add_arg_table s begin
        "folder"
            help = "Folder containging all FastQ Files"
            required = true
            arg_type = String
        "machine"
            help = "Illumina machine used to generate FASTQ files"
            required = true
            arg_type = String
            range_tester = in(["novaseq", "miseq", "nextseq"])
        "--cutoff", "-c"
            help = "Cutoff for quality score.  For probabilistic method this is probability
                    the sequence is correct.  For minQ and trimmomatic methods this is a
```

```
                    phred score (if a proability is given it will be converted to a phred
                    score"
            arg_type = Float64
            default = 0.8
        "--output", "-o"
            help = "Location of output folder"
            arg_type = String
            default = "NGS Analysis Output"
        "--flip", "-f"
            help = "Flip nt sequence if needed"
            action = :store_true
        "--merge", "-m"
            help = "Merge files if paired end reads"
            action = :store_true
        "--beginROI", "-b"
            help = "NT sequence just upstream ROI, use 'N' to specify no upstream sequence"
            arg_type = String
            default = "AGGACTTG"
        "--endROI", "-e"
            help = "NT sequence just downstream ROI, use 'N' to specify no downstream sequence"
            arg_type = String
            default = "TCTAGAAG"
        "--minOverlapFLASH", "-O"
            help = "Minimum overlap for flash to merge sequences"
            arg_type = Int
            default = 10
        "--maxOverlapFLASH", "-M"
            help = "Maximum overlap for flash to merge sequences"
            arg_type = Int
            default = 70
        "--QTrimMethod", "-Q"
            help = "Method to use for quality trimming of data. Can be either:
                    'probabilistic' {reads kept based on total Phred score calculation}
                    'minQ' {reads kept based on lowest probability base in read}
                    'trimmomatic' {reads kept based on sliding window average}"
            arg_type = String
            default = "probabilistic"
        "--window", "-w"
            help = "Only necessary when 'trimmomatic' method is used.  Sets the size of
                    the sliding window"
            arg_type = Int
            default = 4
        "--IgnoreR1", "-i"
            help = "Ignore read 1 files, superceded by -m"
            action = :store_true
        "--IgnoreR2", "-I"
            help = "Ignore read 2 files, superceded by -m"
            action = :store_true
        "--Chunksize", "-H"
            help = "If this value is set to a number greater than 0, nt_seq and QS_score
                    files will be generated in chunks where this value indicates the
                    number of lines that are gone through before a new file is generated.
                    A value of 1,000,000 should speed up processing times."
            arg_type = Int
            default = 0
    end

    return parse_args(s)
end

#Find the input files and merge files if necessary
function get_input_files(folder::String,
                         merge::Bool,
                         minOverlap::Int,
                         maxOverlap::Int,
                         output::String,
                         machine::String,
                         IgnoreR1::Bool,
                         IgnoreR2::Bool)
    files = [i for i in readdir(folder) if i[1] != '.']
    mkpath(output)
```

```
    #Merge files if requested
    if merge
        #Create folder for merged files and find file pairs
        mFolder, flFolder = output * "/Merged", output * "/Flash_Output"
        mkpath(mFolder)
        mkpath(flFolder)
        paired_files = pair_files(files, machine)

        #Go through each pair and merge with flash
        for i in 1:size(paired_files, 1)
            #Find the files and the identifier
            file1, file2 = paired_files[i, :]
            if machine == "novaseq"
                identifier = parse(Int64, last(split(file1, "_")[1], 3))
            elseif machine == "miseq"
                identifier = parse(Int64, split(file1, "_")[1])
            elseif machine == "nextseq"
                identifier = parse(Int64, split(file1, "_")[1])
            end

            #Don't do flash if file already exists
            if string(identifier) * "_merged.fastq" in readdir(mFolder)
                continue
            end

            #identifier = split(file1, "_")[1]
            file1 = folder * '/' * file1
            file2 = folder * '/' * file2

            #Determine the options for flash and run it
            opt1 = "--output-prefix=1"
            opt2 = output
            opt3 = "-m " * string(minOverlap)
            opt4 = "-M " * string(maxOverlap)
            run(pipeline(`./flash $opt1 $file1 $file2 -d $opt2 $opt3 $opt4`, stdout = "devnull"))

            #Move the merged file
            destination = mFolder * "/" * string(identifier) * "_merged.fastq"
            mv(output * "/1.extendedFrags.fastq", destination)

            #Move the other files
            others = [j for j in readdir(output) if (isfile(output * "/" * j)) && (j[1] != '.')]
            destination = flFolder * "/" * string(identifier) * "_"
            for j in others
                mv(output * "/" * j, destination * j)
            end
        end

        #Get and return the merged files
        return [mFolder * "/" * j for j in readdir(mFolder)]

    #If ignoring read 1 or 2 return only the read 2 or 1 files
    elseif IgnoreR1
        return [folder * "/"* i for i in readdir(folder) if (i[1] != '.') && (occursin("R2", i))]
    elseif IgnoreR2
        return [folder * "/"* i for i in readdir(folder) if (i[1] != '.') && (occursin("R1", i))]
    end

    #If not merging return the already good files
    return [folder * "/" * i for i in readdir(folder) if i[1] != '.']
end

#Finds the proper pairs of files and returns them
function pair_files(files::Vector{String},
                    machine::String)
    #Find identifiers and run (i5 or i7)
    IDs = DataFrame(name = String[], identifier = Int[], Run = Int[])
    for i in files
        if i == ".DS_Store"
            continue
```

```
        end

        #Each machine has a slightly different file structure
        if machine == "novaseq"
            id = parse(Int64, last(split(split(i, "_")[1], "-")[3], 3))
            run = parse(Int64, split(i, "_")[3][2])
        elseif machine == "miseq"
            id = parse(Int64, split(i, "_")[1])
            run = parse(Int64, split(i, "_")[5][2])
        elseif machine == "nextseq"
            id = parse(Int64, split(i, "_")[1])
            run = parse(Int64, split(i, "_")[5][2])
        end
        push!(IDs, [i, id, run])
    end

    #Pair up files and return the (Nx2) array
    paired_files=Array{Union{Nothing, String}}(nothing,length(unique(IDs, 2)[!,"identifier"]), 2)
    for (a, i) in enumerate(unique(IDs, 2)[!, "identifier"])
        pair = IDs[isequal.(IDs.identifier, i), :].name
        paired_files[a, :] = pair
    end
    return paired_files
end

#Finds the reverse compliment sequence for a string of nucleotides
function flip_seq(seq::String)
    compliment = Dict('A' => 'T', 'T' => 'A', 'C' => 'G', 'G' => 'C', 'N' => 'N')
    return reverse(join([compliment[i] for i in seq]))
end

#Finds the region of interest in a given sequence
function find_ROI(seq::String, start::String, final::String)
    #Create key but only specify start and final if needed
    if start != "N" key = start * "(.+)"
    else key = "(.+)" end

    if final != "N" key *= final end

    if key == "(.+)"
        return seq, 1, length(seq)
    end
    key = Regex(key)

    #Find the region of interest
    roi = findall(key, seq)

    #If the match does not work do not continue examining the seq
    if length(roi) != 1
        return "", -1, -1
    end

    #Find the start and final index
    sInd, fInd = roi[1][1], last(roi[1])
    if start != "N" sInd += length(start) end
    if final != "N" fInd -= length(final) end

    #Return roi and indices
    roi = seq[sInd:fInd]
    return roi, sInd, fInd
end

#Method for finding Qscore of a given read
function find_Q(qroi::String)
    qroi = [Int(only(i)) - 33 for i in split(qroi, "")]
    qroi = 1 .- 10 .^ (qroi/-10)

    Qscore = 1
    for i in qroi Qscore*= i end
    return Qscore
end
```

```
#Method for finding minimum phred score in a read
function find_Qmin(qroi::String)
    Qmin = minimum([Int(only(i)) - 33 for i in split(qroi, "")])
    return Qmin
end

#Method for determining if window ever drops below cutoff in a read
function find_Qslide(qroi, window, cutoff)
    qroiINT = [Int(only(i)) - 33 for i in split(qroi, "")]

    for i in 1:length(qroiINT)-3
        if sum(qroiINT[i:i+3])/window < cutoff
            return true
        end
    end
    return false
end

#Read in a file for sequence analysis and convert them to fastq files containing
#just the nucleotides from the region of interest (roi)
function read_file(file::String,
                   start::String,
                   final::String,
                   cutoff::Float64,
                   flip::Bool,
                   code_length::Int,
                   output::String,
                   QTrimMethod::String,
                   window::Int,
                   machine::String,
                   chunksize::Int)
    seqs_folder = "nt_seqs"
    QS_folder = "QS_scores"

    #Open the file
    open(file) do F
        #Find the file identifier
        identifier = split(last(split(file, "/")), "_")[1]
        line = 1
        read = ""
        roi, sInd, fInd, = "", -1, -1

        #Make specific filename adendums for breaking up the file into chunks
        chunk = 1
        if chunksize > 0
            chunkname = "~" * string(chunk)
        else
            chunkname = ""
        end

        #Some variables that need to be accessed outside the while loop
        output_nt = ""
        output_QS = ""
        count = 1

        #Go through each line (each entry for a fastq files is 4 lines)
        while ! eof(F)
            l = readline(F)

            #First line is the identifier, just reset some of the variables
            if line % 4 == 1
                read = l * "\n"
                roi, sInd, fInd, = "", -1, -1

            #Second line is the sequence
            elseif line % 4 == 2
                #Flip seq if necessary
                if flip
                    l = flip_seq(l)
                end
```

```
    #Find ROI
    roi, sInd, fInd = find_ROI(l, start, final)

    read *= roi * "\n"

#Third line is a "+" sign
elseif line % 4 == 3
    nothing

#Fourth line is the phred scores
else
    #Do not add sequences
    if sInd == -1
        line += 1
        continue
    end

    #Find Q-score ROI and determine Q score
    if flip
        l = reverse(l)
    end
    qroi = l[sInd:fInd]
    Qscore = find_Q(qroi)

    #Based on quality trimming method decide if read meats minimum standard
    #Create alternate Q20 dataset
    if QTrimMethod == "probabilistic"
        if Qscore < cutoff
            line += 1
            continue
        end
    elseif QTrimMethod == "minQ"
        Qmin = find_Qmin(qroi)
        if cutoff < 1
            cutoff = -log10(1 - cutoff) * 10
        end
        if Qmin < cutoff
            line += 1
            continue
        end
    elseif QTrimMethod == "trimmomatic"
        if cutoff < 1
            cutoff = -log10(1 - cutoff) * 10
        end
        if find_Qslide(qroi, window, cutoff)
            line += 1
            continue
        end
    end

    #Create folders/files and append data
    output_nt *= read * "+\n" * qroi * "\n"
    output_QS *= string(Qscore) * ","
    count += 1

    #For speed, files are only written every 300 reads
    if count % 300 == 0
        filename1 = output * "/" * seqs_folder * "/" * identifier * chunkname * "_"
        filename2 = output * "/" * QS_folder * "/" * identifier * chunkname * "_"
        filename1 *= "nt_seqs.fastq"
        filename2 *= "QS_scores.fastq"

        mkpath(output * "/" * seqs_folder)
        mkpath(output * "/" * QS_folder)

        #Write sequence file
        file1 = open(filename1, "a")
        write(file1, output_nt)
        close(file1)
        output_nt = ""
```

```
                    #Write quality score file
                    file2 = open(filename2, "a")
                    write(file2, output_QS)
                    close(file2)
                    output_QS = ""
                end

                #If the files are large and need to be broken up into chunks
                if chunksize > 0
                    #Check if the chunksize (i.e. the number of reads) of the file
                    #has been reached
                    if count % chunksize == 0
                        #Make a new file and write the remainder of the data to the
                        #old output files
                        filename1 = output * "/" * seqs_folder * "/" * identifier * chunkname*"_"
                        filename2 = output * "/" * QS_folder * "/" * identifier * chunkname * "_"
                        filename1 *= "nt_seqs.fastq"
                        filename2 *= "QS_scores.fastq"

                        mkpath(output * "/" * seqs_folder)
                        mkpath(output * "/" * QS_folder)

                        file1 = open(filename1, "a")
                        write(file1, output_nt)
                        close(file1)
                        output_nt = ""

                        file2 = open(filename2, "a")
                        write(file2, output_QS)
                        close(file2)
                        output_QS = ""

                        chunk += 1
                        chunkname = "~" * string(chunk)
                    end
                end
            end

            line += 1
        end

        #Write the remainder of the data to the output files
        filename1 = output * "/" * seqs_folder * "/" * identifier * chunkname * "_"
        filename2 = output * "/" * QS_folder * "/" * identifier * chunkname * "_"
        filename1 *= "nt_seqs.fastq"
        filename2 *= "QS_scores.fastq"

        mkpath(output * "/" * seqs_folder)
        mkpath(output * "/" * QS_folder)

        file1 = open(filename1, "a")
        write(file1, output_nt)
        close(file1)

        file2 = open(filename2, "a")
        write(file2, output_QS)
        close(file2)
    end
end

#Method to convert FastQ files into dictionary with counts for each unique sequence
function compress_to_dict(files::Vector{String}, output::String, machine::String)
    data = Dict()
    identifier = split(split(last(split(files[1], "/")), "_")[1], "~")[1]
    println(identifier)
    folder = output * "/Raw_nt_counts/"
    mkpath(folder)

    #Go through each file and convert to a redundent array of sequences with a count of 1
    for file in files
```

```
        println("  ", file)
        #For large files larger than 100GB do this line by line version which is slower
        #but will not crash
        if filesize(file) > 100000000000
            #Open the file
            open(file) do F
                line = 1

                #Go through each line
                while ! eof(F)
                    l = readline(F)

                    #Find the nucleotide sequences and add them to the array
                    if length(l) != 0 && length(findall(r"[AGTC-]", l)) == length(l)
                        if l in collect(keys(data))
                            data[l] += 1
                        else
                            data[l] = 1
                        end
                    end

                    #This section is for tracking the progress in the output window
                    if line % 10^7 == 0
                        println("   ", line)
                        CSV.write(folder * identifier * ".csv", data)
                    end
                    line += 1
                end
            end
        #Otherwise read into a redundent array of sequences with a count of 1
        else
            println("  All at once")
            dict_data = CSV.read(file, DataFrame, delim = 'a', header = 0)
            dict_data = dict_data[dict_data.Column1 .!= "+", :]
            dict_data = dict_data[(!).(startswith.(dict_data.Column1, "@") .&
                                contains.(dict_data.Column1, " ")), :]
            dict_data = dict_data[contains.(dict_data.Column1, "T"), :]
            seqs = unique(dict_data.Column1)

            println("  ", length(seqs), " unique seqs")
            CSV.write(folder * identifier * "_raw.csv", dict_data, append = true)
            dict_data = nothing
        end
    end

    #Convert the array into a dictionary with a count associated with each unique sequence
    matrix_data = CSV.read(folder * identifier * "_raw.csv", DataFrame, header = 0)
    matrix_data = convert(Matrix, matrix_data)
    counts = countmap(matrix_data)
    matrix_data = nothing
    CSV.write(folder * identifier * ".csv", counts, append = true)
    counts = nothing
end

function main()
    parsed_args = parse_commandline()

    #Get the FASTQ files and flash to merge if necessary
    files = get_input_files(parsed_args["folder"],
                            parsed_args["merge"],
                            parsed_args["minOverlapFLASH"],
                            parsed_args["maxOverlapFLASH"],
                            parsed_args["output"],
                            parsed_args["machine"],
                            parsed_args["IgnoreR1"],
                            parsed_args["IgnoreR2"])

    #Get some variables from the inputs
    start = parsed_args["beginROI"]
    final = parsed_args["endROI"]
    cutoff = parsed_args["cutoff"]
```

```
    code_length = 0

    #Open Files one at a time and read through them line by line
    flip = parsed_args["flip"]
    QTrimMethod = parsed_args["QTrimMethod"]
    if !(QTrimMethod in ["probabilistic", "minQ", "trimmomatic"])
        println("ERROR: Quality trimming method not found.  Please try ",
                "probabilistic, minQ, or trimmomatic")
        return
    end
    output = parsed_args["output"]
    window = parsed_args["window"]
    machine = parsed_args["machine"]
    chunksize = parsed_args["Chunksize"]
    for i in files
        println(i)
        read_file(i, start, final, cutoff, flip, code_length, output, QTrimMethod,
                  window, machine, chunksize)
    end

    #Get the nt created sequence files
    nt_files = get_input_files(output * "/nt_seqs",
                               false,
                               parsed_args["minOverlapFLASH"],
                               parsed_args["maxOverlapFLASH"],
                               output,
                               parsed_args["machine"],
                               false,
                               false)

    #Convert nucleotide sequences to dictionary
    if chunksize > 0
        nt_file_groups = Set([split(last(split(i, "/")), "~")[1] for i in nt_files])
        for nfg in nt_file_groups
            nfg_files = [file for file in nt_files if occursin("/" * nfg * "~", file)]
            compress_to_dict(nfg_files, output, machine)
        end
    else
        for i in nt_files
            compress_to_dict([i], output, machine)
        end
    end
end

main()
println("Done.")
```

7.8.2. Luria-Delbrück Distribution Code

This code was used from a Jupyter notebook running Julia version 1.7.2.

```
using Distributions
using Random

function simulateLD_DFE(mu::Float64,
                        mut_GrowthRate_vec::Vector{Float64},
                        mut_GrowthRate_sd_vec::Vector{Float64},
                        mut_GrowthRate_freq_vec::Vector{Float64},
                        nRep::Int,
                        nPop::Int;
                        seed::Int = 102594)
    Random.seed!(seed)
    # Starting from one wt cell, how long does it take to get to nPop cells?
    finalT = log(nPop)

    #Create the distribution for finding the DFE
    mGRv = hcat(mut_GrowthRate_vec, mut_GrowthRate_sd_vec)
```

```
    DFE = MixtureModel([Normal(mGRv[i, 1], mGRv[i, 2]) for i in 1:size(mGRv)[1]],
                       mut_GrowthRate_freq_vec)

    # Compute the number of mutations that occur per replicate
    dist = Poisson(mu * nPop)
    r = rand(dist, nRep)

    # At what times does each mutation occur?
    tv = Dict()
    rmax = maximum(r)
    tv = rand(nRep, Int(rmax))
    tv = log.(1 .+ (nPop - 1) * tv)

    # What are all the growth rates?
    GR = rand(DFE, (nRep, rmax))
    GR[GR .< 0] .= 0

    # What is the final number of mutants?
    mutNum = zeros(nRep)
    for rep in 1:nRep
        val = finalT .- tv[rep, 1:Int(r[rep])]
        mutNum[rep] += sum(exp.(GR[rep, 1:length(val)] .* val))
    end

    return mutNum
end
```