Distinguishing between structural models for the *Escherichia Coli* RNA binding protein ProQ

by

Katherine Dailey

A Paper Presented to the Faculty of Mount Holyoke College

In Partial Fulfillment of the Requirements

For the Degree of Bachelors of Arts with Honor

Program in Biochemistry,

Department of Chemistry

South Hadley, MA 01075

May 2022

This paper was prepared

under the direction of

Professor Katie Berry

for eight credits.

Acknowledgements

There is no way that I can thank everyone who assisted me on this journey. Please note this is just a small fraction of those who have helped me over the years.

First, I would like to thank the members of my committee. Thank you to Dr. Katie Berry for giving me the opportunity to pursue research. Your lab gave me the space to cultivate a love for science, even in the middle of a pandemic and hundreds of miles from the bacteria I was studying. I am grateful for the experiences I have already had doing research, as well as the doors that opened during my time in your lab. Thank you to Dr. Amy Camp and Dr. Timothy Farnham. Amy, your courses allowed my love for biochemistry and microbiology to grow in such an authentic, genuine way. Tim, Environmental Issues put the experiments in the lab in a broader societal context, allowing me to understand the connections between my work and the outside world. My scientific writing would not be possible without your course and my other Environmental Studies courses. Many thanks to all of you for allowing me to share my work on this one particular protein with you.

Thank you to past and current members of the Berry Lab. This work would not be possible without the support of all of you, nor without the impressive science done by Smriti Pandey, Oliver Stockert, Chandra Gravel, Linda Wang, and Amy Wang.

Thank you to my coaches. I never thought I would be a college athlete, much less be able to balance athletics and independent research. You both have created an environment that allows for simultaneous athletic and academic growth. To all of my teammates over the last four years: I have learned so much from each and every one of you, and for that, I am beyond grateful. I especially want to acknowledge Jaya, Simran, Jocelyn, Claire, and Amy. It has been a joy to grow as students, rowers, and people with the five of you over this chapter of our lives. I am so impressed by each of you. It is an honor to call you teammates and friends.

Thank you to my friends and family. Thank you to my roommate Amelia, who has supported me throughout this process even though I did not listen to her suggestions for “biochemistry” thesis topics over the last four years. In living with you, you have shown me that you don’t have to be stressed to be successful. Thank you to Halee, for being my biggest fan and very best friend, and for supporting my love for biochemistry even as you could not understand why anyone would spend so many hours on PyMol. Thank you to my parents and sister, for supporting my journey to Mount Holyoke and bringing me flowers and blueberries from the coast of Maine as I spent my first summer away, doing research here. Thank you also to Buttercup and Vincent, for the laughs and cuddles and endless sources of joy. Finally, thank you to my therapist, without whom this work (or any of my research, for that matter) would not be possible.

Table of Contents

List of Figures

List of Tables

List of Abbreviations

Aa	*Aquifex aeolicus*
Ago	Argonaute
α-	N-terminal domain of alpha subunit of RNA polymerase
β-gal	Beta-galactosidase
B3H	bacterial three-hybrid
CASP	Critical Assessment of protein Structure Prediction
CTD	C-terminal domain
DNA	deoxyribonucleic acid
ds	double-stranded
Ec	*Escherichia coli*
gelFRET	gel-based fluorescence resonance energy transfer
IPTG	isopropyl-ß-D-thiogalactoside
λCI	bacteriophage λCI protein
LB	lysogeny broth
Lp	*Legionella pneumophila*
mRNA	messenger RNA
MSA	Multiple sequence alignment
$MS2^{CP}$	bacteriophage MS2 coat protein
$MS2^{hp}$	21-nucleotide cognate RNA hairpin of $MS2^{CP}$
nc	non-coding
NTD	N-terminal domain
Nm	*Neisseria meningitidis*
NMR	nuclear magnetic resonance
PDB	Worldwide Protein Data Bank
RBP	RNA binding protein
RBS	ribosome binding site
RMSD	root mean square deviation
RNA	ribonucleic acid
RNAP	RNA polymerase
rRNA	ribosomal RNA
sRNA	small RNA
SARS-CoV-2	severe acute respiratory syndrome coronavirus 2
ss	single-stranded
TPEG	phenylethyl-β-D-thiogalactopyranoside
tRNA	transfer RNA
UTR	untranslated region
X-gal	5-Bromo-4-Chloro-3-Indolyl β-D-Galactopyranoside

Table 1: Amino acid residues, codes, and ionization states at physiological pH

Amino Acid	Three Letter Code	One Letter Code	Side Chain Charge at Physiological pH (7.4)
Alanine	Ala	A	Neutral
Arginine	Arg	R	Positive
Asparagine	Asn	N	Neutral
Aspartic Acid	Asp	D	Negative
Cysteine	Cys	C	Neutral
Glutamine	Glu	Q	Neutral
Glutamic Acid	Gln	E	Negative
Glycine	Gly	G	Neutral
Histidine	His	H	Neutral
Isoleucine	Ile	I	Neutral
Leucine	Leu	L	Neutral
Lysine	Lys	K	Positive
Methionine	Met	M	Neutral
Phenylalanine	Phe	F	Neutral
Proline	Pro	P	Neutral
Serine	Ser	S	Neutral
Threonine	Thr	T	Neutral
Tryptophan	Trp	W	Neutral
Tyrosine	Tyr	Y	Neutral
Valine	Val	V	Neutral

Abstract

Evolving research on small RNAs (sRNAs) in bacteria implicates sRNAs as a key effector of gene regulation, influencing expression for genes involved in processes from basic bacterial biology to serious public health issues such as virulence and antibiotic resistance. While some sRNAs are able to act independently, many are dependent on an RNA-binding protein, such as the well-established Hfq in *Escherichia coli.* Another family of RNA-binding proteins is the FinO family, including ProQ and FinO in *E. coli*, NMB1681 in *Neisseria meningitidis*, and Lpp1663 in *Legionella pneumophila*. Previous work on ProQ has not supplied a satisfying answer on how ProQ binds to RNA, despite an available NMR structure. In July of 2021, the AlphaFold database was released, which included an alternate structure for ProQ. In order to critically evaluate both of these structures, I compared the structures of FinO domain proteins, examined highly conserved residues Y70 and R80 through the use of a forward genetic screen with our laboratory's bacterial three-hybrid assay, and used the same assay to probe predicted interactions from the structural models with the use of site-directed mutagenesis. The available structures of FinO domains were found to vary from the NMR structure of ProQ in both quality and chemical properties. Two key residues on ProQ, Y70 and R80, were extremely sensitive to mutation. It is possible that these residues are directly involved in RNA binding by ProQ, a hypothesis supported by the structures and research on other FinO domain proteins. This work suggests that the NMR structure of ProQ should be examined more critically, and it is possible that the AlphaFold structure provides an alternate model for this protein. Through this, I hope to generate insights into the most relevant structural conformations for *in vivo* RNA binding by FinO proteins and the ways in which the structure of *E. coli* ProQ is both similar and distinct from orthologous FinO domain proteins.

Chapter I: Introduction

I-1. The central dogma of molecular biology

The central dogma of molecular biology outlines the flow of genetic information in a living organism. This idea states that genetic information is stored in DNA, which is transcribed into RNA, which is then translated into protein (Figure 1). In transcription, RNA polymerase (RNAP) binds to double-stranded DNA, unwinds it, and synthesizes a single-stranded RNA molecule using the rules of complementary base pairing. In translation, a newly formed messenger RNA (mRNA) molecule is translated into a protein with the help of ribosomes (Figure 1). The first step of translation is initiation, in which the mRNA bearing the code for a polypeptide binds to the small subunit of the ribosome. The coding portion of mRNA is defined by a specific AUG start codon. In prokaryotes, this initiation codon is guided into the correct position on the ribosome by the Shine Dalgarno sequence on the mRNA which interacts with ribosomal RNA (rRNA). The Shine Dalgarno sequence is a consistent sequence 8-13 base pairs to the 5' end of the initiation codon which base pairs to a sequence on the small subunit of the ribosome (Nelson & Cox, 2012). In all organisms, this area where the ribosome binds to begin translation is referred to as the "ribosomal binding site" (RBS).

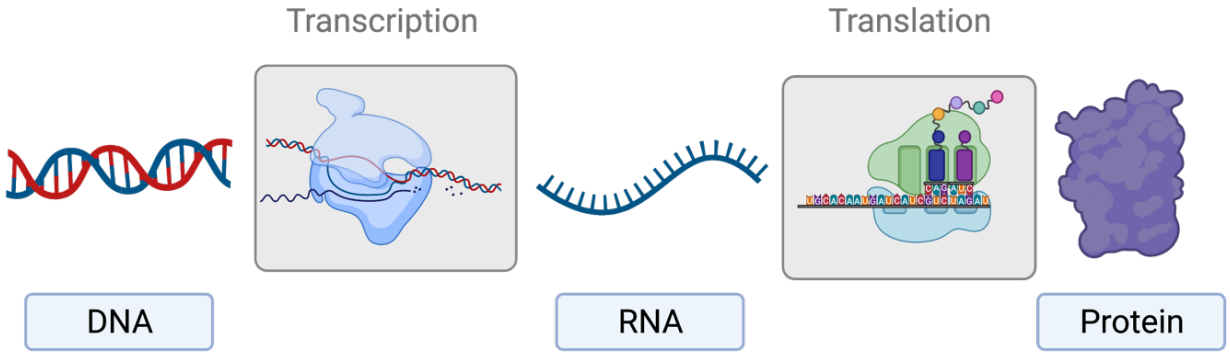


Figure 1: The central dogma of molecular biology. Genetic information is stored in cells in the form of DNA. DNA is transcribed into RNA by RNA polymerase. mRNA is then translated into protein by ribosomes, which

decode the RNA codons and translate it into a polypeptide chain with the help of transfer RNAs (tRNAs). Figure created using Biorender.

Regulation of gene expression may occur at any point in the central dogma, from transcriptional control to post-translational modifications of folded proteins. Gene regulation is critical to organism survival, as it ensures efficient use of resources and appropriate reactions to different stressors and environments. This is especially important for pathogenic bacteria, which must be able to adapt to different environments such as stored food and the human gut. Transcriptional control of gene regulation is the most common, due to being the most efficient use of resources by regulating early on in the process of gene expression. However, post-transcriptional control also occurs, especially in times of stress. It functions primarily by controlling which RNAs are actually translated into proteins.

In order to investigate post-transcriptional control, it is important to establish that not all RNA is translated. Transcription produces many different types of RNA. The three primary kinds of RNA are mRNA, which codes for proteins; transfer RNA (tRNA), which base pairs with mRNA codons to add the correct amino acid to the polypeptide chain; and rRNA, which forms the catalytic core of ribosomes (Nelson & Cox, 2012). An additional class of non-coding RNAs is small RNAs (sRNAs).

I-2: Bacterial sRNAs

sRNAs are non-coding RNA molecules that are involved in gene regulation in bacteria through post-transcriptional control. sRNAs engage in imperfect base pairing with specific messenger RNAs (mRNAs) under different conditions in order to alter which mRNAs are ultimately translated into protein (Wagner & Romby 2015; Figure 2). sRNAs are not expressed constitutively but instead respond to environmental variations in order to appropriately affect gene expression (Wagner & Romby, 2015). The relatively easy synthesis and degradation of

RNA means that sRNAs are valuable for rapid, cost-effective regulation of gene expression in response to environmental cues (Felden & Augagneur, 2021). Regulation of transcription may allow for more efficient use of resources, but it often requires transcription factors. As proteins, transcription factors require translation which takes more time and energy for production than RNA. Due to this difference, sRNAs are able to act more quickly and reversibly. Additionally, the imperfect nature of base pairing means that many sRNAs are able to bind multiple mRNA targets, allowing them to regulate multiple biological functions at once (Felden & Augagneur, 2021; Han et al., 2020; Mai et al., 2019).

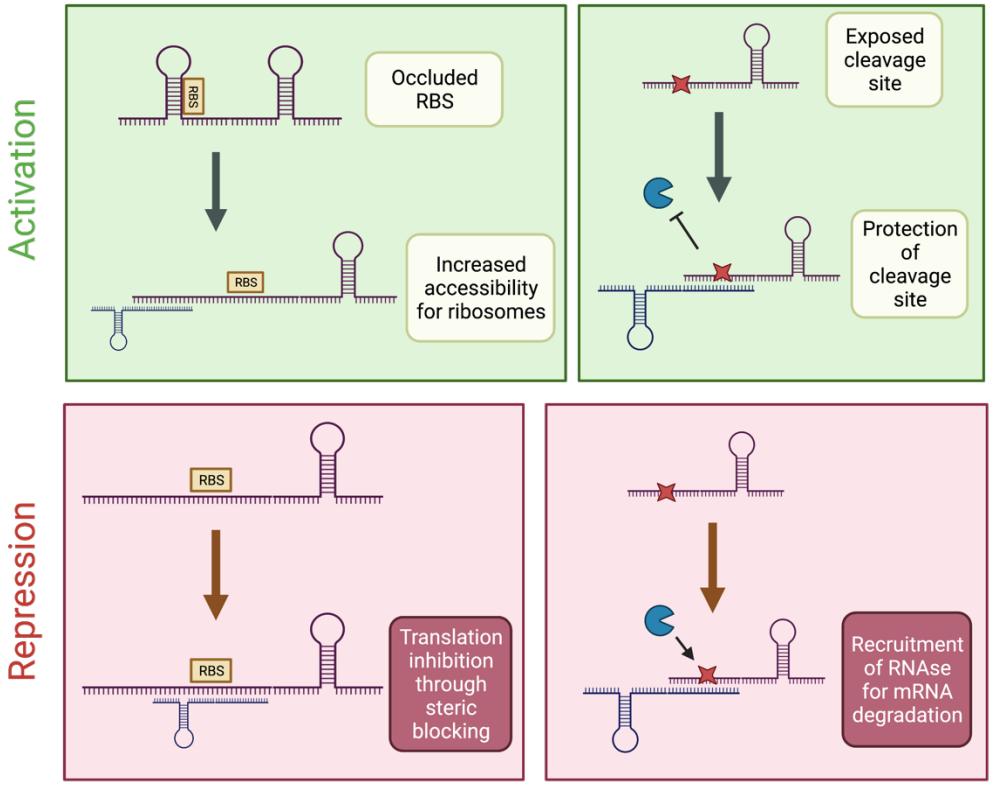


Figure 2: Mechanisms of action used by trans-encoded sRNAs. sRNAs (shown in blue) may lead to the activation or repression of target mRNAs (shown in purple). sRNAs may activate a gene by preventing mRNA from folding the ribosome binding site (RBS) into an inaccessible position or by protecting a cleavage site from degradation by RNAse. sRNAs may repress a gene by sterically blocking a ribosome from a binding site, or by recruiting RNAse to a degradation site. Figure modified from Felden & Augagneur 2021 using Biorender.

I-3: RNA Chaperone Proteins

I-3-i. Hfq

Pairing of sRNAs and mRNAs is sometimes assisted by chaperone proteins, which have the ability to affect either the binding rate or stability of the sRNA-mRNA complex. Perhaps the most well-known of these proteins is the protein hexamer Hfq, which simultaneously binds to sRNA and mRNA to promote base pairing between the two (Wagner & Romby, 2015; Figure 3). Due to this simultaneous binding of both RNAs, this protein is known as an "RNA matchmaker" (Updegrove et al., 2016). Hfq can work with sRNAs in multiple different ways, including protecting sRNAs from degradation, matching sRNAs with target mRNAs, recruiting ribonucleases, and directly interfering with translation (Olejniczak & Storz, 2017). While Hfq is involved in many of these sRNA-mRNA base-pairing events, there are many more sRNAs than those which interact with Hfq and there are many bacterial species that don't have Hfq or an ortholog (Olejniczak & Storz, 2017). This suggests the presence of additional proteins which mediate sRNA-mRNA interactions in bacteria, the search for which has led to the discovery of FinO-domain proteins.

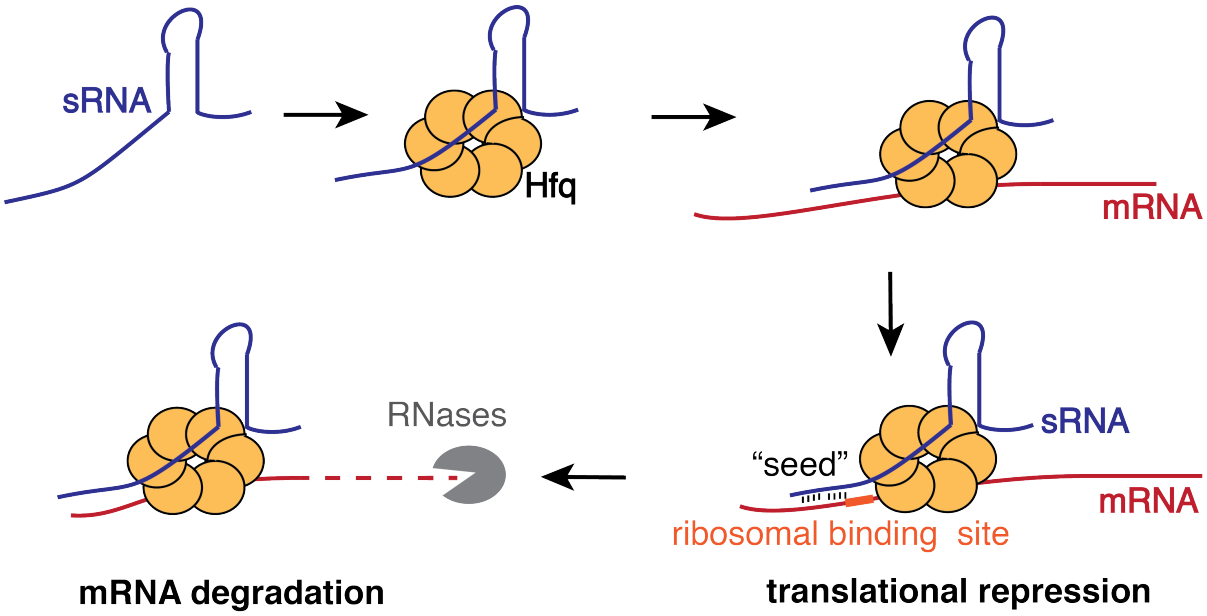


Figure 3: Hfq mediates binding between an sRNA and target mRNA. Hfq simultaneously binds to an sRNA and mRNA, increasing the local concentration of the RNA molecules and facilitating the binding of the "seed" region of the sRNA to the mRNA target. Translation is repressed when Hfq sterically blocks the ribosome from getting onto the mRNA. Hfq is also thought to recruit RNAses, triggering degradation of the mRNA. Figure created by Professor Katie Berry.

I-3-ii. FinO-Domain Proteins

The FinO-domain is named after the central domain of *Escherichia coli* (*Ec*) F-plasmid encoded protein FinO, which consists of 5 α-helical segments and two β strands (Ghetu et al., 2000). FinO binds an sRNA target FinP and an mRNA *traJ*, promoting interactions between FinP and the 5'UTR of *traJ* (Jerome et al., 1999). When FinP interacts with *traJ*, it blocks access to the ribosomal binding site on the mRNA, ultimately inhibiting the translation of *traJ* mRNA. On the eponymous protein, this domain resembles a sort of fist with positively charged residues on the concave surface of the protein (Olejniczak & Storz, 2017; Figure 4). The positively charged residues are of interest as those are most likely to interact with the negatively charged RNA-backbone. The opposite face of the protein has mostly residues of neutral or negative charge, implicating the concave face as the primary site of RNA binding (Olejniczak & Storz, 2017). Interest in positively charged residues was validated by a crosslinking experiment

performed by Ghetu *et al*. (2002), where 12 single-cysteine substituted FinO variants were crosslinked with RNA (Ghetu et al., 2002). The authors were able to show that the cysteine residues positioned near positively charged patches could cross-link with a stem-loop structure on FinP, while cysteine residues near a negatively charged patch did not crosslink to the RNA (Ghetu et al., 2002; Glover et al., 2015; **Figure 4**). This suggests that RNA interacts with FinO in the regions with positively charged amino acids, over those with a majority of negatively charged amino acids.

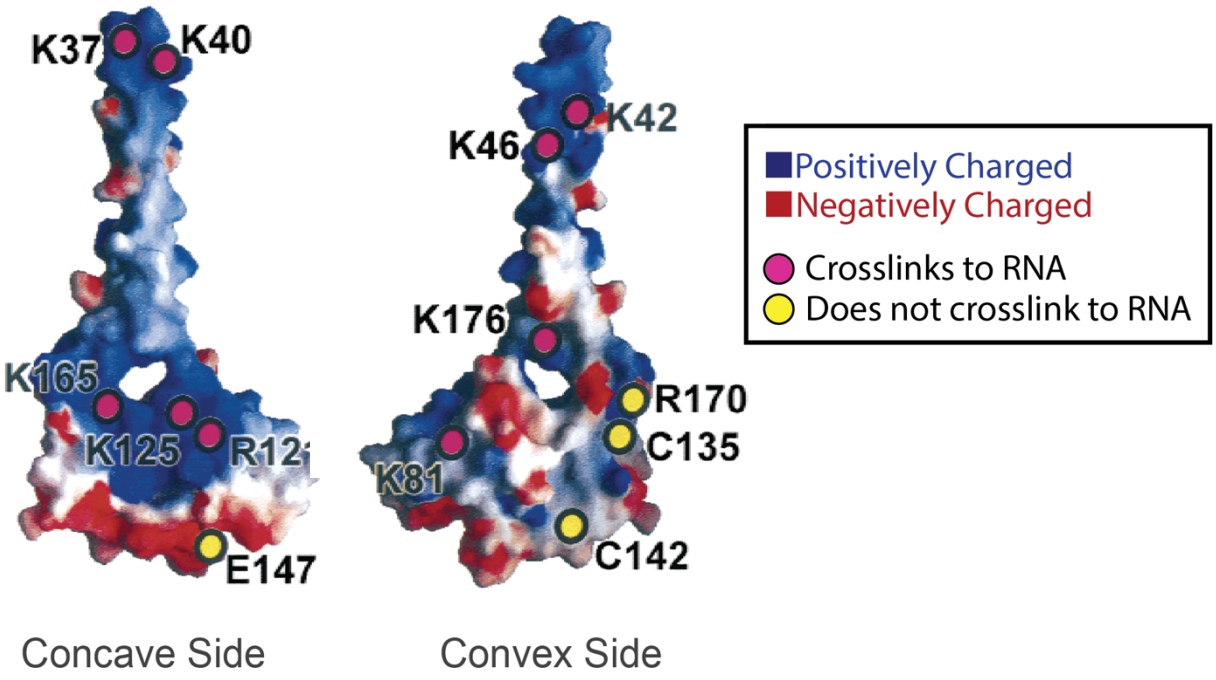


Figure 4: Site-specific crosslinking of FinO and stem-loop structure mapped onto the FinO structure colored by residue charge. Positively charged residues are colored blue and negatively charged residues are colored red. Magenta circles indicate places where an incorporated crosslinker, azidophenacyl bromide (APA-Br) showed significant crosslinking to the RNA target, while yellow circles show sites that do not crosslink to RNA. The concave face of the protein is shown on the left, while the convex face is shown on the right. Figure adapted from Figure 3B in Ghetu *et al*., 2002.

FinO has orthologs in many species of γ-proteobacteria. Orthologs are homologous genes found in different species. Homologous sequences refer to genes with more than thirty percent similarity in amino acid sequence, as this is enough for two proteins to adopt a similar fold (Rost, 1999). Due to the structure/function relationship in proteins, the similar structure of homologous

proteins frequently corresponds to a similar function of the protein *in vivo* (Pearson, 2013). Orthologs of the FinO protein are often referred to as FinO proteins or FinO-domain proteins.

FinO proteins all have a FinO domain which was identified through sequence analysis. Due to the presence of the FinO domain, these proteins are predicted to be RNA-binding proteins. Since there are multiple chromosomally-encoded proteins with FinO domains, this is thought to be a possible new family of bacterial chaperones (Olejniczak & Storz, 2017). The structure of FinO was solved in 2000 by Ghetu *et al.* through X-ray crystallography. Over the past two decades, three additional structures of FinO proteins found to bind RNA have been released on the Worldwide Protein Data Bank (PDB), including the structure of *Neisseria meningitidis* (*Nm*) NMB1681 solved with X-ray diffraction (Chaulk et al., 2010; PDB ID 3mw6), *E. coli* ProQ solved through nuclear magnetic resonance (NMR) (Gonzalez et al., 2017; PDB ID: 5nb9), and *Lpp*1663 in *Legionella pneumophila* (*Lp*) solved with NMR (Immer et al., 2020; PDB ID: 6s10). There has been extensive research implicating the FinO domain of these proteins as the primary site of RNA binding (Arthur et al., 2003; Ghetu et al., 2002; Gonzalez et al., 2017; Immer et al., 2020; Pandey et al., 2020; Sandercock & Frost, 1998; Stein et al., 2020).

I-3-iii. ProQ

ProQ, a chaperone protein found in both *Salmonella enterica* and *E. coli,* impacts the translation of mRNAs through binding to a large suite of sRNAs (Olejniczak & Storz, 2017). ProQ was first discovered in 1983, but its function was unknown and the first identified phenotype was dependent on ProP, a proline transporter (Milner & Wood, 1989; Stalmach et al., 1983). In 2016, the potential importance of ProQ was revealed when genome-wide transcriptomic analyses performed by Smirnov *et al*., showed that more than 400 sRNAs co-immunoprecipitated with ProQ in *S. enterica.* This suggests that ProQ interacts with at least this

many sRNAs in *S. enterica.* Other FinO proteins, such as FinO and RocC, seem to only bind a single sRNA and a few mRNAs (Olejniczak & Storz, 2017). Binding a great number of RNAs targets may allow ProQ to have a larger impact on gene regulation in *Salmonella* and *E. coli*, making it particularly interesting.

The target interactome of ProQ overlaps with that of Hfq (Melamed et al., 2020). This fact, and the knowledge that ProQ binds to sRNAs which act through base-pairing, may suggest that ProQ mediates sRNA-mRNA interactions in a way similar to Hfq (Holmqvist et al., 2020). However, current work has found some traits of RNA targets preferred by ProQ. For one, it has been shown that the presence of double-stranded RNA (dsRNA) in RNA targets is a major determinant of ProQ-RNA interactions (Holmqvist et al., 2020). This suggests that stem-loop structures (as seen in cartoon representation in Figure 2) in the RNA may provide a natural target for ProQ due to the presence of a double-stranded region in this common structure. However, it is unclear why double-stranded regions are a key determinant because the mechanism of RNA binding by ProQ is not yet understood.

ProQ is made up of two domains, a C-terminal domain (CTD) and an N-terminal domain (NTD), joined with an unstructured, short sequence of amino acids (a linker). While the exact mechanism of RNA binding is not yet known, the NTD of ProQ seems to be the primary site of RNA binding. The NTD is the domain with homology to FinO (Olejniczak & Storz, 2017), and the structure of this domain was originally modeled based on FinO due to sequence (Smith et al., 2004, 2007). Additionally, this domain shows similarity in function to FinO. Biochemical *in vitro* studies showed that the isolated NTD of ProQ bound to a particular stem-loop structure in the RNA (a two-stranded stem-loop II RNA) with higher affinity than the full-length protein. In contrast, the isolated CTD had a greatly reduced affinity for the RNA when compared to the full-

length protein (G. Chaulk et al., 2011). This suggests the RNA binding activity of ProQ resides in the NTD. The NTD has also been shown to be able to discriminate between different RNAs (Gonzalez et al., 2017), which suggests that the binding by this domain is substrate-specific. The sequence analysis and experimental data together suggest that the FinO domain is the main RNA binding site for ProQ. Furthermore, past work done by this lab *in vivo* further supports the NTD as the primary site of binding, as removing the C-terminal domain (CTD) from the protein did not significantly alter interaction levels with an sRNA or 3' untranslated region (3'UTR) in a bacterial three-hybrid (B3H) system (Pandey et al., 2020). As the NTD seems to be the primary site of interaction with RNA for this protein, it will be the focus of this study.

Within the NTD of ProQ, specific residues have been highlighted by the field as important to binding of RNA. Through the use of a random mutagenesis screen, Pandey *et al.* (2020) were able to identify several residues on the concave face of ProQ's NTD, and one on the convex face, as necessary for interaction between ProQ and targets sRNA SibB and *cspE* 3' UTR. The single residue on the convex face, R80, naturally drew attention. Prior to this screen, R80 had been identified as a residue likely to mediate interactions between ProQ and RNA due to its nature as a very highly conserved, positively charged residue (Gonzalez et al., 2017). Somewhat surprisingly, residues homologous to R80 are located on the concave face of all other FinO-domain proteins with known structures (Figure 5). The importance of this residue was confirmed when it was pulled out of two separate saturation mutagenesis loss of function screens performed on *Salmonella enterica* ProQ (El Mouali et al., 2021; Rizvanovic et al., 2021). Although these screens were performed in *Salmonella* rather than *E. coli*, the results should be considered applicable to *E. coli* ProQ due to the fact that *Salmonella* and *E. coli* are very closely related and the *E. Coli* ProQ NTD is >99% identical to the *Salmonella* ProQ NTD (Rizvanovic et

al., 2021), suggesting that the structures very likely have an identical fold *in vivo*. Therefore, this study can be taken as further support for the importance of R80 in this RNA binding protein.

A

Ec ProQ	G37	K35	D41	K54	R58	Y70	L71	R80	D82	G85
F' FinO	G99	ala	ala (-1: D102)	his (+1: K118)	R121	Y133	L134	R143	D145	G148
Nm NMB1681	G47	ala	D51	ala	ala (+1: R66)	Y77	L78	R87	D89	asn
Lp RocC	G52	K50	D56	K71	R75	Y87	L88	R97	D99	G102
*Lp*1663	G40	K37	D44	K59	R63	Y75	L76	R86	D87	G90

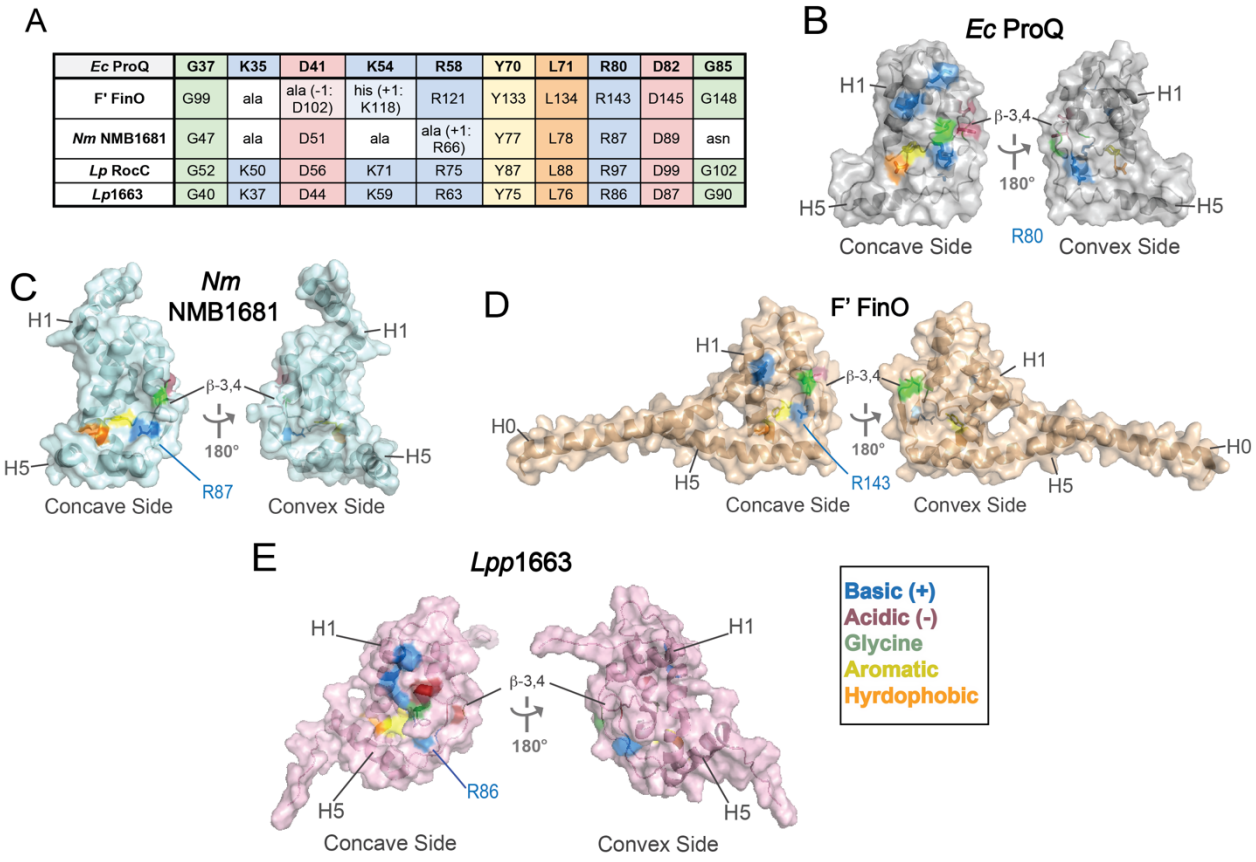


Figure 5: Conservation of *E. coli* ProQ RNA-binding residues across other FinO-domain proteins. A) 10 residues identified by Pandey *et al.* (2020) as important for RNA binding by *E. coli (Ec)* ProQ are listed across the top row, shaded by amino-acid character (blue is basic, orange is hydrophobic, yellow is aromatic, red is acidic, green is glycine). Corresponding residues from *Ec* F' FinO, *N. meningitidis (Nm)* NMB1681, *L. pneumophila (Lp)* RocC, and *Lp Lpp*1663 are listed in rows below. The separate panels depict surface and cartoon representations of **B)** *Ec* ProQ NTD NMR structure (PDB ID: 5nb9), **C)** crystal structure of *Nm* NMB 1681 (PDB ID: 3MW6), **D)** crystal structure of *Ec* F' FinO protein (PDB ID: 1DVO), **E)** NMR structure of *Lp* Lp1663 (PDB ID: 6S10) viewed from the concave face on the left and convex face on the right. Figure adapted from Pandey *et al.* (2020) Supplementary Figure 12, with the addition of *Lpp*1663, the structure of which was only just solved by Immer *et al.* in 2020.

The screen performed by Pandey *et al.* pulled out an additional residue studied in this work, Y70. Y70 was a residue of interest prior to the screen as Y70 was highly conserved, located on the concave face of the NTD (which was proposed to be the primary site of RNA binding), and aromatic (Pandey et al., 2020). Surface-exposed aromatic and hydrophobic residues often mediate intermolecular interactions in the cell, such as interactions between

proteins and nucleic acids. The particularly interesting trait of R80 and Y70 is an incredible sensitivity to mutation, suggesting a very specialized function. When each of the residues was mutated to alanine (a mutation often used to "remove" an amino acid by reducing the length of the side chain and eliminating any reactive groups), binding to RNA by ProQ was greatly diminished. Furthermore, even with a very conservative mutation to either of these residues (such as R80K, Y70S, or Y70F) binding was lost completely (previously reported in Oliver Stockert '21 Thesis; Pandey et al., 2020; Figure 6).

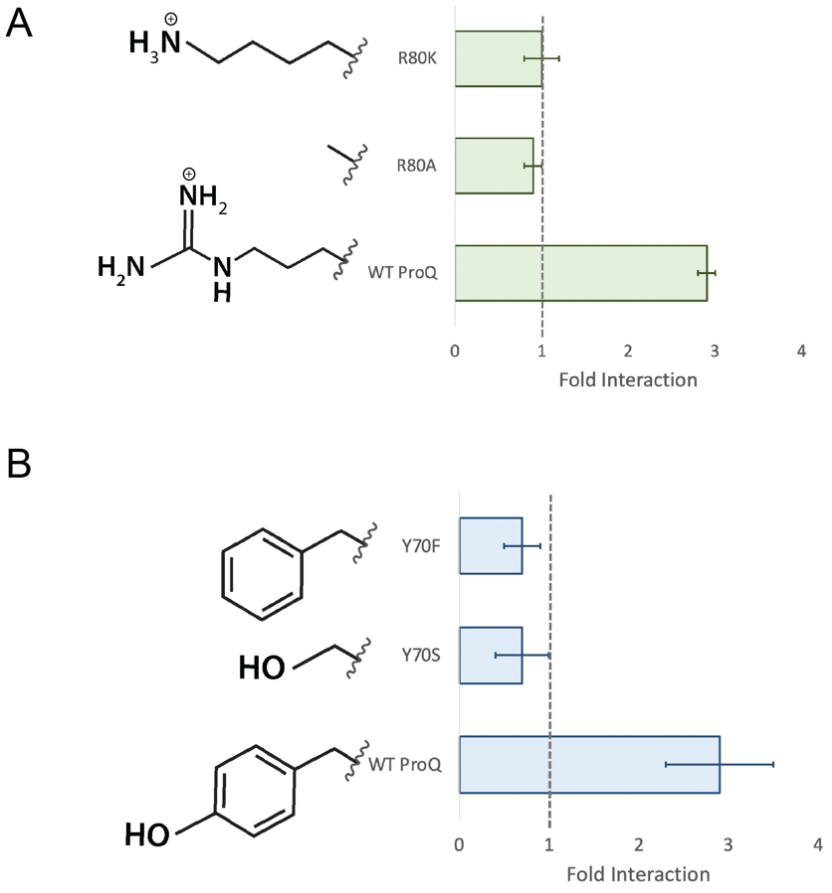


Figure 6: Binding of ProQ NTD mutants to *cspE* 3'UTR. Fold interaction represents relative interaction between the RNA and protein, as measured by a bacterial three-hybrid assay (Berry & Hochschild, 2018). The top panel **A)** depicts data previously presented by Oliver Stockert '21, showing that even a conservative mutation to R80, such as R80K, greatly reduces the ability of ProQ to bind to the RNA target. The same is seen to be true for mutations to Y70 in panel **B)**, where either retaining the hydroxyl group (Y70S) or the aromatic ring (Y70F) still results in a loss of binding ability (Pandey et al., 2020). The amino acid side chains depict the wild type residue, as well as the residues which are mutated to in order to illustrate the conservative nature of the mutations. Amino acid structures from Biorender.

While work performed following the release of the NMR structure for ProQ in 2017 has brought the field closer to a model of RNA binding by ProQ, in the absence of a solved costructure with ProQ and RNA the specific details of the interaction remain uncertain. In attempts to learn more about this mechanism, the NMR structure of ProQ has since been cited in many papers on ProQ in both *Salmonella enterica* and *E. coli* (El Mouali et al., 2021; Leonard et al., 2021; Pandey et al., 2020; Rizvanovic et al., 2021; Stein et al., 2020) although it is possible that this structure is not the most accurate model of ProQ.

I-4: AlphaFold

I-4-i. The protein-folding problem

In 1961, Anfinsen *et al.* introduced the "protein folding problem" when they determined that the structure of a protein was encoded in the amino acid sequence. The initial experiment involved denaturing a bovine pancreatic ribonuclease (RNase) and allowing the protein to refold. After undergoing treatment with urea and a reducing agent, the protein lacked RNase activity and had an altered spectral absorption, indicating a change in structure. When the urea was removed and molecular oxygen was added in order to encourage oxidation, the protein returned to the native state as determined by monitoring both the spectra and RNase activity of the protein (Anfinsen et al., 1961). This study demonstrated that all information needed to fold correctly was contained within the amino acid sequence of the protein. With this knowledge came the belief that the structure of proteins could be predicted from only the amino acid sequence. An effort to computationally predict protein structures was born. A few decades later, the Critical Assessment of protein Structure Prediction (CASP) began. Starting in 1994, once every two years CASP assessed the protein-structure prediction field blindly (AlQuraushi, 2019).

In the first few CASP challenges, there was a lot of progress made as researchers rapidly recognized the value in modeling based on similar proteins and fragment assembly (Bouatta, Sorger, & AlQuraishi 2021). Sequence analysis allowed homologous proteins to be identified and used to assist in modeling (Perrakis & Sixma, 2021). Progress on the protein-folding problem slowed in the late 2000s and early 2010s, accelerating at the end of the 2010s. In December of 2020, the organizers of CASP14 stated that DeepMind had solved the protein folding problem with their AlphaFold2 algorithm (Bouatta et al., 2021). Improvement seen in the last decade is primarily driven by the development of co-evolution methods based in statistical physics and the application of deep-learning techniques to protein structure prediction (Bouatta et al., 2021).

Advancements have also been the result of a shift from researchers attempting to predict protein structure from one sequence alone to using the information found in multiple sequence alignments (MSAs) (Bouatta et al., 2021). The revolutionary AlphaFold2 algorithm operates on MSAs at its core (Jumper et al., 2021; Figure 7). However, utilizing this information was really only possible with recent advancements in computer science. Progress in machine learning and neural networks has allowed the incorporation of MSAs to be more effective. Current neural networks have unparalleled pattern-recognition capabilities (Bouatta et al., 2021). These capabilities are very valuable in protein structure prediction. While there are many theoretically possible conformations of amino acid sequences, in reality, these conformations are limited by biophysics. Within this, evolution has only explored a small subset of all possible protein conformations (Bouatta et al., 2021). In other words, it is very likely that the majority of proteins follow similar folding patterns. As a result, advancements in pattern recognition have allowed machine learning algorithms to more effectively identify patterns in known protein

structures. This may include folding motifs, which are common three-dimensional structures seen in a variety of different proteins. Once known, patterns may then be applied to sequences with currently unknown structures.

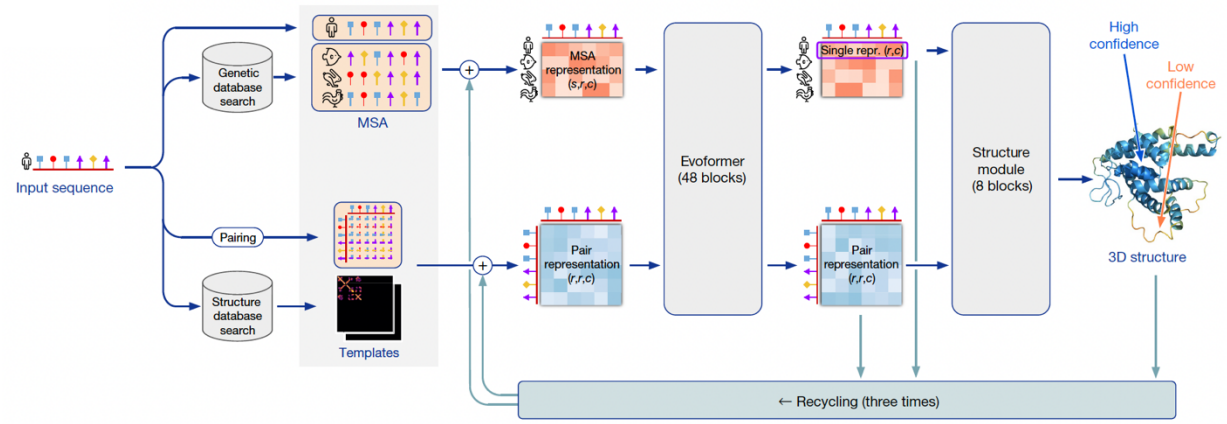


Figure 7: The architecture of the AlphaFold2 algorithm. Arrows show the flow of information among various components. Parentheses on the array structures specify the geometry of the array, with s corresponding to the number of sequences, r corresponding to the number of residues, and c corresponding to the number of channels. From an input sequence, an MSA is generated and evolutionarily coupled residues are recognized as "pairs." There is a flow of information between the MSA and pair representation, which allows both of these features to continually inform each other in the Evoformer module. This mechanism of information exchange allows direct reasoning about spatial and evolutionary relationships. A structural hypothesis arises early in the Evoformer block and is continuously refined. The structure module introduces an explicit 3D structure in the form of a "residue gas," in which the chain is able to break to allow for local refinement of many parts of the structure to occur simultaneously. Again, this structure is iteratively refined which is credited as making a great contribution to the accuracy of the modeling (Jumper et al., 2021). Adapted from Jumper *et al.* 2021 Figure 1e.

I-4-ii. AlphaFold2's solution

When AlphaFold first appeared at CASP13 it represented a major advancement in computational structure prediction. The advancement made by AlphaFold2 was even greater. This algorithm depends heavily on data gathered by experimental scientists over the last 50 years, namely sequences and structures of proteins (Jumper et al., 2021). The input protein sequence is used to create an MSA, which provides AlphaFold2 information both on conserved structures and evolutionarily coupled residues. The benefit of recognizing conserved structures is clear, for if a certain sequence of amino acids is known to fold a certain way then recognizing the sequence in a protein of unknown structure gives an idea of how that particular area might fold. The use of evolutionarily coupled residues is slightly more complex. AlphaFold2 uses MSAs to

look for inter-residue correlations in the MSA in order to model co-evolution (Bouatta et al., 2021). Co-evolution is the idea that if two residues are close in space and interact with each other, even if the residues are far apart in the amino acid sequence, the residues will evolve together in order to preserve the structure and function of the protein. This means that if the properties of one of the residues changes as a result of a mutation, the other residue must change in order to continue to interact and preserve the structure and function of the protein (Perrakis & Sixma, 2021). This concept allows AlphaFold to glean information on spatial contacts in the protein, even for residues that are far away in the primary sequence of the protein (Bouatta et al., 2021). Further along in the algorithm, in the structure module (Figure 7), the AlphaFold2 algorithm then focuses on spatially local interactions in order to perform structure refinement such as fixing steric clashes and improving secondary structure (Bouatta et al., 2021).

AlphaFold2 implements an ensemble-based prediction method, considering groups of homologous proteins over individual sequences (Bouatta et al., 2021). There is some concern about the use of this method, especially when considering solving the structures of proteins that lack many homologous sequences or homologs with solved structures (Bouatta et al., 2021). However, it should be noted that while some of the exact inner workings of AlphaFold2 remain unknown, there is some independence from homologs. This is seen in the fact that AlphaFold2 has been able to accurately predict structures with few available sequences, most notably the SARS-CoV-2 Orf8 protein, which had very few orthologs (Flower & Hurley, 2021; Kovner, 2021). This could be due to the previously mentioned idea that throughout evolution proteins have not explored all available conformations, and therefore basic patterns in protein folding hold true.

While the ideas of co-evolution and ensemble-based prediction are not new to the field of protein prediction, there are a few details that set AlphaFold2 apart from previous work. The inputs and outputs of the algorithm are slightly different from other computational methods. The algorithm starts with raw MSAs as inputs, rather than summarized statistics as some structure prediction methods have shifted to. AlphaFold2 predicts a 3D structure as the output, (rather than an intermediate recording of the location of atoms, as other prediction algorithms have) which is thought to possibly allow for better use of many iterative refinements. Protein structures predicted by AlphaFold2 go through the prediction process 48 times, per the published paper (Jumper et al., 2021). This is a very high number of iterative refinements. The iterative refinements are better taken advantage of with the use of attention, a new concept in artificial intelligence which allows the coders to indicate what it is the program should "pay attention" to (Cho et al., 2014). This method has also been used for other advancements in artificial intelligence, including language translation, image analysis, and modeling brain information processing (Cho et al., 2014; Kriegeskorte, 2015; Xu et al., 2015). A final component important to the success of AlphaFold2 is the ability of the algorithm to recognize the same protein rotated slightly as the same object. Historically this has been challenging for computers, which often have considered the same protein rotated slightly to be an entirely different computational object (Bouatta et al., 2021). It must also be noted that the true differences between machine learning algorithms are challenging to discern, as the connections that the algorithm makes based on training data remain unknown to even those who created the algorithm (Hohman et al., 2018). Therefore, comprehensive knowledge of the differences between AlphaFold2 and previous algorithms cannot be determined.

I-4-iii. Broader impact of AlphaFold2

AlphaFold2 does not increase our understanding of protein folding, as some may wish accurate protein prediction methods would. Some suggest that it should be considered more like a powerful new experimental method for determining structures (Bouatta et al., 2021). Like experimental methods for structure determination, AlphaFold2 is not aware of energy minima or other factors in folding (Perrakis & Sixma, 2021). It does not tell us how amino acid sequence determines the structure, or provide clarity on protein folding pathways in the cell, as proposed by Levinthal (Levinthal, 1969). Rather, this algorithm is simply focused on finding and replicating patterns.

Structural biology remains important in order to increase our understanding of how proteins work and interact with each other (Perrakis & Sixma, 2021). Furthermore, like all artificial intelligence, AlphaFold2 was developed based on what it was fed, which was solved protein structures. For this reason, it is possible that unexpected structures will not be predicted with as consistent accuracy. Regardless, AlphaFold2 may be considered a source of testable hypotheses on protein structure (Perrakis & Sixma, 2021).

I-4-iv. AlphaFold proposed model for ProQ

AlphaFold2 greatly outperformed other methods at CASP14, and was found to have accuracy close to experimental methods in many cases (Jumper et al., 2021). From the approximately 100 protein targets in Casp14, AlphaFold2 performed poorly on five of them. Three of these targets were oligomeric complexes (structures formed when multiple separate polypeptide chains come together) and the other two were structures determined by NMR (Bouatta et al., 2021). Predictions by the original AlphaFold algorithm at CASP13 were poor matches for NMR structures as well. It must be noted that these inconsistencies could be due to

how NMR data has been converted into a protein model rather than the actual protein structure (Callaway, 2020). NMR can provide information about the distances between atoms, but these distance constraints must then be fed into a computer which provides a group of structures that are consistent with the constraints (Nelson & Cox, 2012). It is possible that the software used to develop structures from the data did not account for the constraint data well. In these cases, it would be interesting to compare the structure predicted by AlphaFold2 directly to the NMR constraints. The AlphaFold2 structure may still be in line with the experimental measurements, even if it is not in line with the proposed structure determined from those measurements.

The AlphaFold database of protein structures was launched in July of 2021. The original release included predicted structures for all proteins in twelve model organisms. The predictions were made by AlphaFold2, but will be referred to as AlphaFold structures from this point onward in order to match the database name and broader conversation about this software. The predicted structures released included the *E. coli* genome, and most notably for the purposes of this study, ProQ. The structure predicted for ProQ by AlphaFold2 looks dramatically different from the NMR structure for ProQ, but looks more similar to FinO, with a more distinct concave and convex face (Figure 8). Variation between these structures added to concern about the accuracy of the NMR structure, due to the high level of accuracy seen from AlphaFold2.

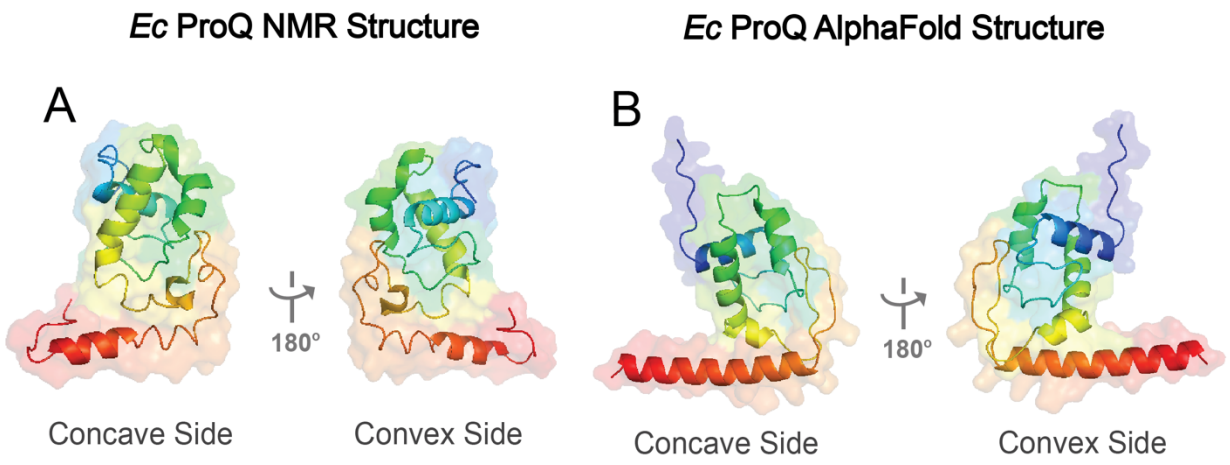


Figure 8: The two proposed structures for *E. coli* ProQ. A) displays the currently accepted structure for ProQ, solved through NMR in 2017 by Gonzalez *et al.* (PDB ID: 5nb9) while **B)** shows the alternative structure solved computational by AlphaFold (Jumper et al., 2021). The AlphaFold structure has more distinct concave and convex faces and a long helix extending from the core of the protein.

It is not unreasonable to consider the AlphaFold model as a legitimate structural model for ProQ. Despite the very recent release of the AlphaFold database, it has already been used to accurately predict the folding of an RNA-binding protein in humans, hnPNK. This protein was predicted to have multiple RNA binding motifs, with three of the domains thought to be the main sites for RNA interaction. Unexpectedly, one of the domains predicted to interact with RNA did not bind to nucleic acids when purified as a single domain but was active in binding when it was folded with another one of the domains predicted for RNA interaction. This was true even when the second domain was mutated to make it not functional for binding of RNA (Yao et al., 2021). The crystal structure of this protein showed that these two domains folded as a single domain (Yao et al., 2021), which had been previously predicted by AlphaFold (Jumper et al., 2021). This prediction of an unlikely folding pattern supports a conclusion that the CASP14 performance hinted at: recent developments in computational methods of protein prediction have led to structures that should be considered seriously in attempts to understand unexpected results.

I-5: Studying RNA-Protein Interactions

I-5-i: Computational insight

RNA binding proteins may be studied computationally in order to gain hypotheses into which residues are most likely involved with RNA binding. Positively charged residues are more likely to interact with the negatively charged sugar-phosphate backbone of RNA. Additionally, conserved residues in any protein are more likely to be important to protein structure and/or function, as these specific amino acids have been held on to by evolution. For this reason, this study will highlight both conserved and positively charged residues on the structure of ProQ and other FinO-domain proteins.

I-5-ii: *in vitro* study of RNA binding

Traditional methods of studying protein binding of RNA include *in vitro* methods such as electromobility shift assays (EMSAs) in which the interactions between purified RNA and purified protein are studied. This and other *in vitro* methods, such as protein-RNA crosslinking and gel-based fluorescence resonance energy transfer (gelFRET), have been helpful for the study of FinO-domain proteins (Ghetu et al., 2002; Stein et al., 2020). However, these methods require careful purification of both protein and nucleic acid – a labor-intensive process. Furthermore, *in vitro* experiments cannot perfectly replicate the environment in the cell, and therefore may be less informative on how the protein and RNA interact in the natural environment than *in vivo* experiments.

I-5-iii: *in vivo* study of RNA binding

Previously mentioned studies on ProQ used loss of function mutagenesis screens to identify residues important to protein function (El Mouali et al., 2021; Rizvanovic et al., 2021). In one of these studies, the authors simply looked for cells that exhibited a phenotype previously

seen in a Δ*proQ* strain, which indicated inactivation of the protein. These cells were then sent for sequencing to check for amino acid substitutions in the ProQ protein (El Mouali et al., 2021). The other study described the development and use of a fluorescence-based reporter which could be used with fluorescence-activated cell sorting (FACS) and high-throughput sequencing. The authors fused a fluorescent protein to a gene for which transcription was indirectly activated by ProQ. This allowed loss-of-function phenotypes to be identified in cells with lower levels of fluorescence (Rizvanovic et al., 2021). While both of these studies offer examples for how phenotypes can be used to probe molecular mechanisms, loss of function studies are dependent on a specific phenotype. This limits the application of the work, as it is not as easy to specifically examine the interactions between a single RNA target and protein. The ability to focus on a specific interaction is especially valuable for global RNA binding proteins such as ProQ, which impact many sRNAs and mRNAs and therefore may be impacting the results of loss of function screens in unknown ways.

In vivo assays can allow for more targeted focus on specific interactions in the cell, by allowing the researchers to clone in the appropriate RNA and protein of their choice. The first such assay created for the study of RNA-protein interactions was in *Saccharomyces cerevisae* (SenGupta et al., 1996). While yeast three hybrid assays are helpful for the study of eukaryotic RNA-protein interactions, it is not possible to study prokaryotic RNA-protein interactions with such an assay.

This study makes use of genetic methods of studying RNA-protein interactions with a bacterial three-hybrid assay developed by Berry & Hochschild (2018). This assay has been shown to detect protein-RNA interactions *in vivo.* The three components consist of a DNA-bound protein fused to an MS2 coat protein ($MS2^{CP}$), a hybrid RNA, and a fusion protein. The

hybrid RNA has a region constant to all bait constructs, which includes a MS2 hairpin ($MS2^{hp}$), a specific stem-loop structure made from RNA (cartoon representation shown in the gray part of the cartoon in Figure 9). This structure is bound by the $MS2^{CP}$, a protein isolated from bacteriophage (Berry & Hochschild, 2018) which holds this specific structure in the RNA and allows us to tether the RNA bait upstream of the *lacZ* reporter gene on the DNA. The hairpin is joined to a varied region in the RNA, where targets or mutant targets of the RNA binding protein can be cloned in. This allows for different RNA constructs to be tested in the assay. The RNA binding protein of interest is fused to the alpha subunit of *E. coli* RNA polymerase (Figure 9). The relative locations of the RNA construct held by the MS2 coat protein and the RNA binding protein fused to RNA polymerase means that interaction between the RNA and protein stabilizes RNA polymerase on a promoter, leading to transcription of *lacZ* (Figure 9). *lacZ* encodes for the enzyme β-galactosidase, which allows transcription of this reporter gene to be monitored by measuring β-galactosidase activity (Berry & Hochschild 2018). Through this assay, β-galactosidase activity by bacteria grown in liquid culture can be measured and used as a proxy for interaction between RNA and protein. Fold interaction between RNA and protein is determined by dividing the β-galactosidase activity in Miller Units by the highest basal level of activity. Basal activity is that seen when any one of the three components of the assay is left out (Berry & Hochschild, 2018).

This assay may also be used to show RNA-protein interactions in bacteria grown on plates, with the addition of a specific indicator X-gal to the agar. When β-galactosidase is produced, it hydrolyzes X-gal which spontaneously dimerizes to create an insoluble blue pigment (Padmanabhan et al., 2011). Therefore, the more β-galactosidase produced, the more blue pigment is produced on the plate. Tracing this back to the B3H assay, the higher the level of

interaction between the RNA and prey protein, the more β-galactosidase produced and the more blue the bacteria appears on the plate. This allows us to observe levels of interaction between RNA and protein by looking at the level of blue/whiteness on the plate. This can be used either as an additional assay for liquid experiments (Stein et al., 2020) or for genetic screens using the bacterial three-hybrid assay (Berry & Hochschild, 2018; Pandey et al., 2020).

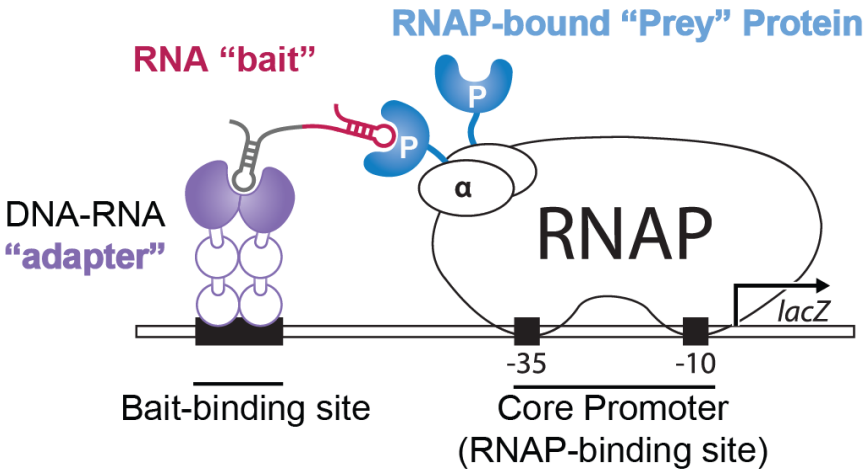


Figure 9: Schematic of the B3H system. Interaction between the "prey" protein and the "bait" RNA stabilizes RNA polymerase on the promoter region, leading to transcription of test promoter *lacZ*. *lacZ* encodes β-galactosidase. β-galactosidase activity in the cell is then measured, acting as an indication of the level of interaction between the "prey" protein and "bait" RNA. Figure adapted from Berry and Hochschild, 2018.

1-6: Statement of Purpose

In this study, I gathered data in order to determine the most accurate structure of *E. coli* ProQ. I computationally examined proposed NMR and AlphaFold structures of ProQ alongside solved structures of other FinO-domain proteins. By examining the structures of proteins in this family solved with different experimental techniques, I developed a reasonable idea of general structure features and relative locations of key residues in this domain. This provided a point of comparison for the two proposed structures for *E. coli* ProQ, and the reliability of each structure was further assessed computationally using basic quality measurements. To complement

computational analysis, important structural features in ProQ were investigated with experimentally, including a screen to search for any amino acids which allowed for RNA binding by ProQ at two highly conserved positions. Finally, interactions present only in the AlphaFold structure were probed with site-directed mutagenesis in order to examine the accuracy of this structure. Together, this work will allow for more effective future study of ProQ, by encouraging development of a model for ProQ binding RNA using the most relevant structure available. Additionally, this study adds to the growing body of work analyzing the utility of DeepMind's AlphaFold2 algorithm.

Chapter II: Materials and Methods

II-1. Bacterial Strains

E. coli strains used in this study are listed in Table 2. NEB5-alpha, purchased from New England Biolabs, was the recipient strain for all cloned plasmids. KB473 cells harbored the B3H assay and acted as the reporter strain for both liquid and plate-based ß-galactosidase assays. Each strain has specific antibiotic resistance, listed with the following abbreviations: AmpR (ampicillin and carbenicillin), CmR (chloramphenicol), KanR (kanamycin), StrR (streptomycin), and TetR (tetracycline). All strains were stored as glycerol stocks at -80ºC.

Table 2: Strains used in this study.

Name:	Description:	Used for:	Antibiotic resistance:	Reference/Source:
KB473	FW102 Δhfq cells with an F' episome which has test promoter placOL2-62 fused to *lacZ*	Harboring the B3H	KanR, StrR	(Berry & Hochschild, 2018)
NEB 5-alpha F'Iq	*E. coli* cloning cells with the F' episome and *lacIq*	Constructing plasmids	TetR	New England Biolabs

II-2. Plasmid Construction

II-2-i. Plasmids

Plasmids used in this study are listed in Table 3, with antibiotic resistance indicated using the same abbreviations outlined above. Plasmids with the prefix pCG, pCH, pCW, pKB, and pSP were cloned by Chandra Gravel (MHC '20), Courtney Hegner (MHC '19), Clara Wang (MHC '19), Professor Katie Berry, and Smriti Pandey (MHC '19) respectively.

Table 3: Plasmids used in this study. Listed in alphabetical order by plasmid name.

Name:	Description:	Details:	Antibiotic resistance:	Reference
pACλCI	Empty vector	Encodes full-length λCI under control of *lacUV5* promoter	CmR	(Dove et al., 1997)
pBrα	Empty vector	pPrey plasmid without a protein linked on, serves as a negative control	AmpR	(Dove et al., 1997)
pCG55	pBrα-ProQ^{FL}-R80X	R80 codon randomized in pKB949	AmpR	Cloned by Chandra Gravel '20
pCH1	pCDF-1xMS2^{hp}	pBait plasmid without an RNA bait attached, serves as a negative control	SpcR	(Pandey et al., 2020)
pCW17	pAC-p$_{constitutive}$-λCIMS2^{CP}	Residues 1-248 of CI fused to an MS2 coat protein; transcription of this protein under the control of a constitutive promoter	CmR	(Pandey et al., 2020)
pKB949	pBrα-ProQ^{FL}	Residues 1-248 of alpha fused via three alanine residues to full-length wild type *E. coli proQ*, serves as positive control	AmpR	(Pandey et al., 2020)
pKB955	pBrα-ProQ$^{\Delta CTD}$	Residues 1-248 of alpha fused via three alanine residues to residues 1-176 of wild type *E. coli proQ*	AmpR	(Pandey et al., 2020)
pKB1210	pCDF-pBAD-1xMS2-malM-3'UTR	3'UTR of *E. coli malM* (final 90 nts) cloned into the pBait construct of the B3H assay	SpecR	(Stein et al., 2020)
pKB1225	pBrα- ProQ^{FL}-C24L	C24L mutation in pKB949	AmpR	Cloned by Prof. Katie Berry
pKB1226	pBrα- ProQ^{FL}-C24V	C24V mutation in pKB949	AmpR	Cloned by Prof. Katie Berry

pKB1227	pBrα- ProQ^{FL}-C24I	C24I mutation in pKB949	AmpR	Cloned by Prof. Katie Berry
pKD01	pBrα-ProQ$^{\Delta CTD}$-Y70X	Y70 codon randomized in pKB955	AmpR	This study
pKD02	pBrα-ProQ^{FL}-R80X+Y70X	Both R80 and Y70 codons randomized in pKB949	AmpR	This study
pKD03	pBrα- ProQ$^{\Delta CTD}$ -D41K	*proQ* D41K mutation introduced to pKB955	AmpR	This study
pKD04	pBrα- ProQ$^{\Delta CTD}$ -K35D	*proQ* K35D mutation introduced to pKB955	AmpR	This study
pKD05	pBrα- ProQ$^{\Delta CTD}$ -D41K+K35D	*proQ* K35D+D41K mutations simultaneously introduced to pKB955	AmpR	This study
pKD06	pBrα- ProQ$^{\Delta CTD}$ -Y70E	*proQ* Y70E mutation introduced to pKB955	AmpR	This study
pKD07	pBrα- ProQ$^{\Delta CTD}$ -Y70W	*proQ* Y70W mutation introduced to pKB955	AmpR	This study
pKD08	pBrα- ProQ$^{\Delta CTD}$ -H95R	*proQ* H95R mutation introduced to pKB955	AmpR	This study
pKD09	pBrα- ProQ$^{\Delta CTD}$ -R80H	*proQ* R80H mutation introduced to pKB955	AmpR	This study
pKD10	pBrα- ProQ$^{\Delta CTD}$ -R80H+H95R	*proQ* R80H+H95R mutations both introduced to pKB955	AmpR	This study
pKD11	pBrα- ProQ$^{\Delta CTD}$ -H95F	*proQ* H95F mutation introduced to pKB955	AmpR	This study
pKD12	pBrα- ProQ$^{\Delta CTD}$ -H95Q	*proQ* H95Q mutation introduced to pKB955	AmpR	This study
pKD13	pBrα- ProQ$^{\Delta CTD}$ -H95A	*proQ* H95A mutation introduced to pKB955	AmpR	This study
pRM16	pFW11tet-O_L2-85	pACλCI with the promoter from pCW17	CmR	Cloned by Rachel Mansky ‘20

pSP10	pCDF-1xMS2^{hp}-*cspE*-3'UTR	*E. coli* 3'UTR of *cspE* cloned into the pBait construct of the B3H assay	SpecR	(Pandey et al., 2020)
pSP14	pCDF-1xMS2^{hp}-SibB	*E. coli SibB* sRNA cloned into the pBait construct of the B3H assay	SpecR	(Pandey et al., 2020)
pSP117	pBrα- ProQ$^{\Delta CTD}$ -R80A	*proQ* R80A mutation introduced to pKB955	AmpR	(Pandey et al., 2020)

II-2-ii. Q5 Site-directed mutagenesis

Single site mutagenesis was performed using Q5 cloning to create pKD03, pKD04, pKD06, pKD08, pKD09, pKD10, pKD11, pKD12, and pKD13 (Table 3) for this study. Single-strand forward and reverse primers (Table 4) were designed using NEBaseChanger. The primers were at least partially complementary to the backbone sequence. These primers were diluted appropriately and used in a polymerase chain reaction (PCR) with 2X Phusion Master Mix (New England Biolabs). Cycling conditions for 30 µL PCR reactions are listed in Table 5. Following the PCR reaction, the product was run on 1% agarose gel with GelRed nucleic acid stain (Biotium) with New England Biolabs 6x loading dye for approximately 28 minutes at 100 watts. It was run next to the template plasmid and a 1kb ladder in order to check for a product of the correct length. The gel was visualized with UV light. If a band was observed at the correct length, the PCR product underwent KLD treatment. KLD treatment phosphorylates (K=kinase) and ligates (L) the PCR product and digests (D) the original template DNA so that the genetic material will be more easily taken up by cells and the plasmid will not be contaminated by the template. In this study, a homemade KLD treatment was used by combining 1 µL PCR Product, 1 µL T4 DNA Ligase (New England Biolabs), 1 µL T4 DNA Ligase buffer (New England Biolabs), 1 µL T4 polynucleotide kinase (PNK; New England Biolabs), 1 µL DpnI (New England Biolabs), and 5 µL MilliQ water. The KLD product was transformed into NEB 5alpha

cloning cells (Table 4). Cells were plated with glass beads and single colonies were picked to be sent for sequencing in order to find a colony with the intended mutation and only that mutation. The library plasmids used in this study (pCG55 and pKD01 (Table 3)), were also created using Q5 cloning, with slight differences in the primers and method for plasmid extraction. Partially complementary single-strand forward and reverse primers (Table 4) were designed using NEBaseChanger. For each of these plasmids, a single codon of the primers was replaced by a blend of 25% of each nucleotide in each of the three sites, written as "nnn" in the primer sequence (Table 4). The primers were appropriately diluted and both PCR and KLD treatment were performed as outlined above.

Table 4: Oligonucleotides used in this study. Listed alphabetically by name. Here n = A/U/G/C**.**

Oligo Name:	Description:	Used for:	Sequence:
oCG64	F ProQ R80X library	Q5 PCR	CGGCGCAACGnnnGTCGATCTTG
oKB1077s	Sequencing oligo for pBrα plasmids	Sequencing	GAACAGCGTACCGACCTGG
oKD01	F ProQ Y70X library	Q5 PCR	GAGCTGGCGTnnnCTTTACGGTG
oKD02	R ProQ Y70X library	Q5 PCR	GAAGTGTAGAGACGTAAAGC
oKD03	F ProQ D41K	Q5 PCR	GGTATTTTTCAGaagTTGGTCGATCGTGTTG
oKD04	R ProQ D41K	Q5 PCR	GATTTTCAGCGGACGCGC
oKD05	F ProQ K35D	Q5 PCR	GGTATTTTTCAGGACTTGGTCGATCG
oKD06	R ProQ K35D	Q5 PCR	GATgtcCAGCGGACGCGCTTC
oKD07	F ProQ H95R	Q5 PCR	GGACGAGCAAcgtGTAGAGCATG
oKD08	R ProQ H95 mutations	Q5 PCR	AGCTCACCGCATGGGTTG

oKD09	F ProQ R80H	Q5 PCR	CGGCGCAACGcatGTCGATCTTG
oKD10	R ProQ R80H	Q5 PCR	GGTTTAACACCGTAAAGATAACGCC
oKD11	F ProQ H95F	Q5 PCR	GGACGAGCAAtttGTAGAGCATGCTCGC
oKD12	F ProQ H95Q	Q5 PCR	GGACGAGCAAcaaGTAGAGCATG
oKD13	F ProQ H95A	Q5 PCR	GGACGAGCAAgctGTAGAGCATGCTCGC
oSP92	R ProQ R80 mutation	Q5 PCR	GGTTTAACACCGTAAAGATAAC

Table 5: Q5 cloning PCR cycling conditions. T_a varies based on the primers used. The T_a used with these primers was the lower of the two (one for each primer) recommended by NEBaseChanger.

Step		Temperature (ºC)	Time
Initial melt		98	5 sec
25x	Denature	98	5 sec
	Anneal	T_a + 3	5 sec
	Extension	72	2 min 45 sec
Final extension		72	10 min
Hold		10	∞

II-2-iii. Collection of library plasmid

The KLD products for pCG55 and pKD01 were transformed into NEB 5-alpha cells. Cells were spread on an LB-carbenicillin plate using the glass bead method and incubated at 37ºC overnight. The following morning, the plates were checked for an appropriate amount of colony growth. For a 64 codon library, such as pCG55 and pKD01, the goal was ~500-600 colonies in order to have sufficient coverage of the library. Once there were enough total colonies for each library, 5 mL of LB broth was added directly to each plate using a pipette aid.

Colonies were gently loosened from the plate using a sterile glass spreader, which had been soaked in ethanol overnight and was flamed immediately before each use. The resulting "cell slurry" was transferred to 1.5 mL Eppendorf tubes. If the library required multiple plates for appropriate coverage, some cell slurry was added from each plate to the Eppendorf in order to maximize diversity. Once the Eppendorf was full, the tubes were spun down at 5000 rpm for 2 minutes. The supernatant was vacuumed off. More cell slurry was added, the tubes were spun down again, and the supernatant was vacuumed off again until there was a sufficient amount of cells in the Eppendorf (approximately 100 µL of cells). Cells were resuspended in 600 µL MilliQ and miniprepped following Zymo Zyppy miniprep protocol. Any remaining cell slurry was added to Eppendorf tubes, labeled, and stored at -80ºC for future use.

II-3. Bacterial three-hyrid assays

II-3-i. Liquid assays

Reporter cells (KB473) were co-transformed with pAC-, pBR-, and pCDF- derived plasmids. pAC constructs express the CI-MS2^{CP} fusion protein, while pCDF-pBAD constructs express the MS2^{hp} fusion RNA and pBR-α expresses α-ProQ protein. For each transformation there were three controls, one where each of these core plasmids was replaced with an "empty" construct. Single colonies from each transformation were inoculated into 1 mL of LB broth supplemented with 0.2% arabinose and antibiotics: carbenicillin (100 µg/mL), chloramphenicol (25 µg/mL), kanamycin (50 µg/mL), and spectinomycin (100 µg/mL) in a 2 mL 96 well deep well block (VWR) sealed with breathable film (VWR) and shaken at 900 rpm and 37ºC overnight. Overnight cultures were back-diluted (1:40) into 200 µL with the same antibiotics and arabinose as outlined above, as well as 0 µM, 5 µM, or 50 µM IPTG (isopropyl-ß-D-thiogalactoside, exact amount used specified in figure caption for corresponding data) into

optically clear 200 µL flat bottom 96-well plates covered with plastic lids (Olympus). The back dilution was grown to mid-log at 37ºC, shaking at 900 rpm. Mid-log was determined by measuring the OD_{600} of each well with a microplate spectrophotometer (Molecular Devices SpectraMax), and was defined as roughly OD_{600} 0.3-0.9. Once grown, mid-log cells were transferred into a new 96 well plate with rLysozyme and PopCulture reagent (EMD Millipore). The cells were left to lyse for 0.5-4 hours. Lysate was transferred into a fresh optically clear 96 well plate (Olympus) with Z-buffer, ONPG (O-nitrophenyl-ß-D-galactopyranoside), and ß-mercaptoethanol. ß-galactosidase activity was measured by taking OD_{420} values every minute at 28ºC for 1 hour using a microplate spectrophotometer (Molecular Devices SpectraMax). OD_{420} readings were normalized using the OD_{600} values from directly before lysis in order to give ß-galactosidase activity in Miller units (Stockert et al., 2022; Thibodeau et al., 2004). ß-galactosidase activity was averaged over three replicates for each experimental condition, and then divided by the highest relevant negative control in order to give the fold interaction. Negative controls refer to transformations in which one component of the assay (pPrey, pBait, or pAdapter) has been replaced with an empty construct. Error for fold interaction was propagated from the standard deviations of experimental and negative control averages. Assays were conducted in biological triplicate on at least three separate days, with the exception of one set of experiments (specified in corresponding figure caption, see Results).

II-3-ii. Plate-based bacterial three hybrid assay

A sterile, optically clear 200 µL flat bottom 96-well plate (Olympus) was filled with 200 µL in each well of liquid broth supplemented with 0.2% arabinose and antibiotics: carbenicillin (100 µg/mL), chloramphenicol (25 µg/mL), kanamycin (50 µg/mL), and spectinomycin (100 µg/mL. Mid-log growth from the OD_{600} plate was back-diluted into this fresh sterile plate at a

concentration of 1:100. The plate was left to sit for 5-10 minutes. 3.2 µL from each well of the plate was pipetted onto a large LB agar plate supplemented with inducers (0.2% arabinose and 1.5 µM IPTG), antibiotics (carbenicillin (100 µg/ml), chloramphenicol (25 µg/ml), kanamycin (50 µg/ml) and spectinomycin (100 µg/ml)) and indicators (X-gal (40 µg/ml) and TPEG (200 µM)). While pipetting, a 96 block grid was placed below the plate for guidance. Plates were incubated at 37ºC overnight, and then moved to a 4ºC fridge for at least one day before pictures were taken of the plate.

II-3-iii. Forward genetic screen using the B3H

The library plasmids were transformed along with pKB1210 into eKB473 cells which had pCW17 pre-transformed. pKB1210 is the bait construct for the 3'UTR of mRNA *malM* (Table 1), which we have determined to have a high level of interaction with ProQ in our assay. After screening of the R80X library (pCG55) was well underway and partially into screening of the Y70X library (pKD01), we shifted to using eKB473 cells with both pCW17 and pKB1210 pre-transformed in order to maximize transformation efficiency (Figure 10). Cells were heat shocked and plated on LB agar supplemented with inducers (0.2% arabinose and 1.5 µM IPTG), antibiotics (carbenicillin (100 µg/ml), chloramphenicol (25 µg/ml), kanamycin (50 µg/ml) and spectinomycin (100 µg/ml)) and indicators (Xgal (40 µg/ml) and TPEG (200 µM)). Plates were incubated at 37ºC overnight. Each time that a transformation was performed, positive and negative controls were transformed alongside the experimental conditions in order to allow comparison of blue/white levels. The positive control for this screen was wild type, full length *E. coli* ProQ (pKB949, Table 3) and the negative control was an empty alpha plasmid (pBrα; Table 3). In order to ensure that cells with medium levels of interaction were not missed, ProQ mutants which were previously established to have a medium level of interaction in the B3H (pKB1225,

pKB1226, and pKB1227; β-galactosidase assays performed by Amy Wang ‘22) were transformed as an additional control (Table 3). The plates were transferred from the 37ºC incubator to a 4ºC refrigerator once colonies had grown to sufficient size, ~18 hours. Plates were in the fridge for a minimum of ~5 hours and a maximum of ~3 days before being examined for the presence of blue colonies. The total number of colonies and number of blue colonies was recorded for each plate. Blue colonies were picked and restruck on plates identical in composition to the original plates to ensure clear, single blue colonies. Single blue colonies were miniprepped and sent for sequencing. Sequences were aligned to the sequence of wild-type *E. coli ProQ* in order to check for mutations at the codon of interest.

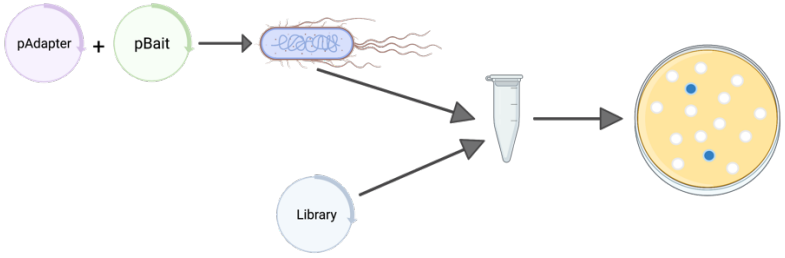


Figure 10: *E. coli* cells were pre-transformed with pAdapter and pBait constructs in order to allow for more efficient library screens. Library screening with pre-transformed cells allowed for greater transformation efficiency, and therefore more effective screening. Here cells were pre-transformed with the adapter and bait components of the B3H so that many different prey plasmids (containing different codons at either position 80 or position 70) could be screened.

Following the discovery of non-tyrosine or non-arginine amino acids at positions 70 or 80, respectively, in blue colonies, the prey plasmid was isolated for further testing. The miniprep used for sequencing was diluted ~1:100 and transformed into NEB 5-alpha cells, plated onto LB-carbenicillin (100 µg/ml). Single colonies from this transformation were patched onto 3 separate plates, one with carbenicillin (100 µg/ml), one with spectinomycin (100 µg/ml), and one with chloramphenicol (25 µg/ml). We selected a colony which was resistant to carbenicillin but sensitive to spectinomycin and chloramphenicol, as this indicates a colony which contains only the pPrey plasmid and not the “bait” or “adapter” components of the assay. A colony which was

resistant to only carbenicillin was restruck for single colonies on a carb plate. A single colony from this plate was used to inoculate an overnight for miniprep, which was then sent for sequencing a second time in order to make sure that we got back a clean read of the correct sequence. This sequence was then given a plasmid name and used in a β-galactosidase assay.

II-4. Dot Blot

2.5 μL of cell lysates from a liquid ß-galactosidase experiment were transferred to nitrocellulose Protran membranes (Amersham) with a multichannel pipette. Separate membranes were used for each IPTG concentration, which was 0 μM and 50 μM in the case of this study. See the corresponding figure legend for an indication of the IPTG concentration used in the specific data shown. Membranes were allowed to dry, and then blocked with a solution of 10% Tween, 10X TBS, and powdered milk before being probed with a primary antibody (1:10,000 anti-ProQ; kindly provided by G. Storz) overnight at 4ºC. The membrane was then washed three times with a 1X TBS/Tween wash buffer, and probed with a horseradish peroxidase conjugated secondary antibody (1:10,000 anti-rabbit IgG; Cell Signaling Technology). The membrane was then rinsed three more times, and treated with ECL Plus western blot detection reagents (BioRad). Chemiluminescent signal was captured with a c600 imaging system (Azure), operated according to manufacturer's instructions. Densitometry analysis was performed using ImageJ to measure the intensity of the dots on the resultant images. In order to normalize the data, four measurements of intensity were only of the membrane, with no lysate spotted on. These four measurements were averaged to get a background value, which was then subtracted from all other intensities. The intensities were normalized by the OD_{600} values corresponding to the individual transformations. The normalized intensities were calculated in duplicate for each

transformation and averaged between the two. The densitometry analysis was performed separately for each IPTG concentration.

II-5. Computational work

II-5-i. Study of orthologs

Conservation in orthologs were examined with the help of ConSurf (Ashkenazy et al., 2016). The prompts of "The ConSurf Server" were followed to create a figure for a protein with known structure for *N. meningitidis* NMB1681 (PDB ID 3mw6), *E. coli* ProQ (PDB ID: 5nb9), and *L. pneumophila Lpp*1663 (PDB ID: 6s10). ConSurf generated the multiple sequence alignment (MSA) for each protein. PyMol was used to visualize the resulting data (Schrödinger, LLC, 2015). For NMB1681, all but one chain (chain A) were deleted to focus on a single binding domain for more appropriate comparison to other structures. A conservation figure was generated for the AlphaFold structure by opening the .pdb file generated by AlphaFold in a text editor and pasting the conservation scores generated by the ConSurf Server for the NMR structure of ProQ into the b-factor column. The proteins were also colored by residue charge using code detailed on PyMol Wiki (*Show Charged - PyMOLWiki*, 2009). Finally, the AlphaFold structure for ProQ was directly compared to the other available structures in PyMol. Structures for the FinO domain in orthologs were aligned to the NTD (residues 1-119) of the AlphaFold structure of ProQ. For the NMR structure for ProQ, the 14aa tag used to solve the structure was removed. For NMB1681, only a single chain was shown for more appropriate comparison between structures. For each alignment, the root mean square deviation (RMSD) reported by PyMol was recorded.

II-5-ii. Search for interactions in the AlphaFold structure

The AlphaFold database for predicted protein structures includes predicted interactions between residues in each structure. I went through each residue in the structure predicted by AlphaFold for *Ec* ProQ and recorded any interactions between sidechains. This did not include interactions between different areas of the protein backbone or interactions between sidechains and the backbone, as such interactions would be more challenging to probe through mutagenesis in our assay. After a list of potential sidechain interactions had been collected, the corresponding residues were examined in the NMR structure. In order to assess the viability of a certain reaction in the NMR structure, the residues involved were rotated around single bonds into the most favorable position and the distance between the ends of the residues was measured using the measurement wizard in PyMol (Schrödinger, LLC, 2015). The released structure file for PDB entry 5nb9 includes many different potential structures based on the NMR data. For these reason, the distances between residues were checked and recorded for every structure in the file.

After locating interactions between side chains unique to the AlphaFold structure, mutations which could potentially "rescue" the interaction were proposed. This was done by considering the chemical and geometric traits potentially involved in each interaction and selecting amino acids which maintained at least some of the key chemical traits. Mutagenesis was performed on the specific residues on the AlphaFold structure in PyMol (Schrödinger, LLC, 2015). This illustrated what the mutation would look like in the AlphaFold structure if the protein were to fold the same way as the wild type protein, with the understanding that the protein may fold differently following mutation.

II-5-iii. Validation work using Coot

Coot (Emsley et al., 2010) was used to further investigate the findings of the PDB validation report for 5nb9, which showed that this entry was of lower quality in clash between residues, locations of residues on a Ramachandran plot, and sidechain outliers (PDB, 2020). In this work, I focused on analysis of clash and Ramachandran plots. Ramachandran plots were created for PDB entries 1dvo, 6s10, 3mw6, and 5nb9 as well as the AlphaFold entry for ProQ. For NMB1681 (3mw6), the atoms for all but one chain (chain A) were deleted to focus on a single binding domain for more appropriate comparison to other structures. For both NMR structures, *Lpp*1663 and ProQ, atoms for all but the first model were deleted by opening the PDB file in TextEdit and deleting atoms. This was done because Coot reported clashes and outliers for all models in the NMR structure file at once. Removing all but one model allowed for more fair comparison to the other structures, for which the clashes and outliers were only reported for a single model. For the NMR ProQ structure (5nb9) the 14-aa tag included with the structure (Schrödinger, LLC, 2015) was removed for this analysis. For the AlphaFold ProQ structure, the CTD was removed (leaving residues 1-133). Following appropriate modifications, each structure was then loaded into Coot. A Ramachandran plot was created for each structure. Clash determination was run for each structure as well, by using the "Probe clashes" feature in Coot and displaying only "bad overlap" (Emsley et al., 2010).

Chapter III: Results

III-1. Single-Codon Genetic Screens

Previous work has shown that mutations at positions 70 and 80, even conservative ones, abolish binding of ProQ to RNA (Pandey et al., 2020; Stockert, 2021). In order to determine if any amino acids besides those found in wild-type ProQ at position 80 or position 70 would facilitate ProQ binding of RNA, we mutated those codons to random nucleotides to create two libraries. This was done through PCR, with the nucleotides at the respective codon on the primer replaced with a blend containing 25% of each nucleotide (Table 4). The PCR product was transformed into cells and grown to a near lawn. Plasmid was collected from a "cell slurry" made from the near lawn to optimize diversity. The result was one library with random nucleotides at position 70, the Y70X library, and one with random nucleotides at position 80, the R80X library. These libraries were then screened on plates for colonies that maintained interaction between ProQ and RNA using the B3H assay and cells that had pre-transformed adapter and bait components.

The total number of colonies screened, as well as the number of blue colonies found from the R80X library can be seen in Table 6. After sending many colonies for sequencing, each of the six Arg codons were returned (

Table 7). Since no codons for other amino acids were returned and the library was covered 29 times, we feel confident saying that no other amino acid at this position allows ProQ binding to *malM*.

Table 6: Summary of colonies screened in R80X library screen. Times the library was covered is an estimate, determined by calculating (colonies screened)/(size of the library).

Entries in R80X Library	Total Colonies Screened	Times Library Covered (Approx)	Blue Colonies Found	% Colonies Blue
64	1895	29	66	3.48

Table 7: Summary of results from R80X library screen. 19 clean reads were obtained from the screen: 4 CGT, 1 CGC, 1 CGA, 1 CGG, 8 AGA, and 4 AGG. All codons returned code for arginine.

Codons	Codes for	Times Returned
CGT	Arginine	4
CGC	Arginine	1
CGA	Arginine	1
CGG	Arginine	1
AGA	Arginine	8
AGG	Arginine	4

The total colonies screened, as well as the blue colonies found for the Y70X library can be seen in Table 8. Both tyrosine codons were returned in this screen, as were GAG (glutamic acid) and TGG (tryptophan). The pPrey plasmids for the two colonies with non-tyrosine codons were isolated (see Methods) so that each could be tested again in the liquid β-galactosidase assay, which allows for more easily quantifiable data. While these two amino acids did appear in the screen, each only appeared once while one tyrosine codon was returned sixteen times (TAT) and the other three times (TAC) (Table 9). The uneven distribution indicates that tyrosine is greatly preferred at this position.

Table 8: Summary of colonies screened in Y70X library screen. Times the library was covered is an estimate, determined by calculating (colonies screened)/(size of the library).

Entries in Y70X Library	Total Colonies Screened	Times Library Covered (Approx)	Blue Colonies Found	% Colonies Blue
64	1064	16	36	3.38

Table 9: Summary of results from Y70X library screen. 21 clean reads were obtained from the screen: 16 TAT (tyrosine), 3 TAC (tyrosine), 1 GAG (glutamic acid), and 1 TGG (tryptophan).

Codons	Codes for	Times Returned
TAT	Tyrosine	16
TAC	Tyrosine	3
GAG	Glutamic Acid	1
TGG	Tryptophan	1

ProQ sequences containing the codon for tryptophan and glutamic acid were isolated from a blue colony, which indicated that those amino acids could be at position 70 in a protein which maintained the ability to bind to RNA. However, since each of these amino acids was only returned once in the primary screen results, I wanted to verify the ability of these amino acids to maintain ProQ's interaction with RNA. The pPrey plasmid was isolated from these colonies through miniprep, and the sequences of these colonies were checked again. A sequence with only a clean Y70W mutation and another sequence with only a clean Y70E mutation were each transformed back into eKB473 cells. These new pPrey plasmids were transformed along with the pBait and pAdapter components (pKB1210 and pCW17, respectively) in order to run a liquid β-galactosidase assay and gather more quantitative data on the differences in RNA binding between these variants. Additional pPrey plasmids with mutations at Y70 were also included to allow for comparison to past results (Pandey et al., 2020). All variant proteins tested had loss of interaction when compared to wild type ProQ (Figure 11).

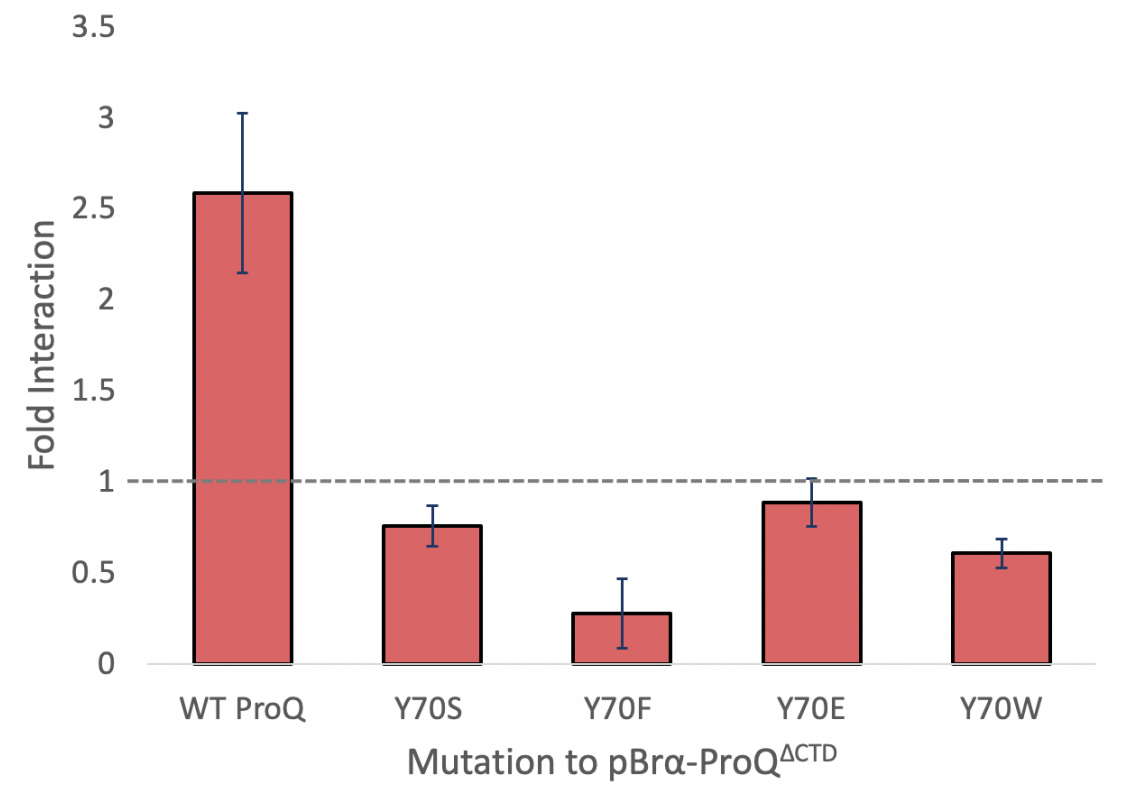


Figure 11: Liquid B3H assay data for Y70 mutants identified in the screen. The variants were tested against a *malM* pBait construct. All variants showed a complete loss of binding to *malM*. The plate assay was also run for this experiment and was in agreement with the results found in liquid. Results shown here are an average of three independent experimental conditions with standard deviation. This experiment was performed on three separate days, and this data is representative of that found each time.

III-2. Ortholog Study

III-2-i. Charge and conservation

In order to investigate residues potentially involved in RNA binding on ProQ orthologs, we generated images of FinO-domain proteins colored by conservation (created with ConSurf (Ashkenazy et al., 2016)) and images of the proteins colored by residue charge. On all proteins except for *Ec* ProQ, a large patch of highly conserved residues can be seen on the concave face of the protein (Figure 12). The three non-ProQ proteins have more distinct concave and convex faces, an observation best seen in three-dimensional (3D) renderings of these proteins. All of the proteins have a central area of the concave face which is made of positive and neutral charged residues, with negatively charged residues focused around the rim of this face. With the

exception of ProQ, the convex face of the proteins include negatively charged residues across the center of the domain. FinO and Lpp1663 are relatively lacking in positively charged amino acids on the convex face of the domain, while both ProQ and NMB1681 have greater concentrations of positively charged amino acids on the convex face (Figure 12).

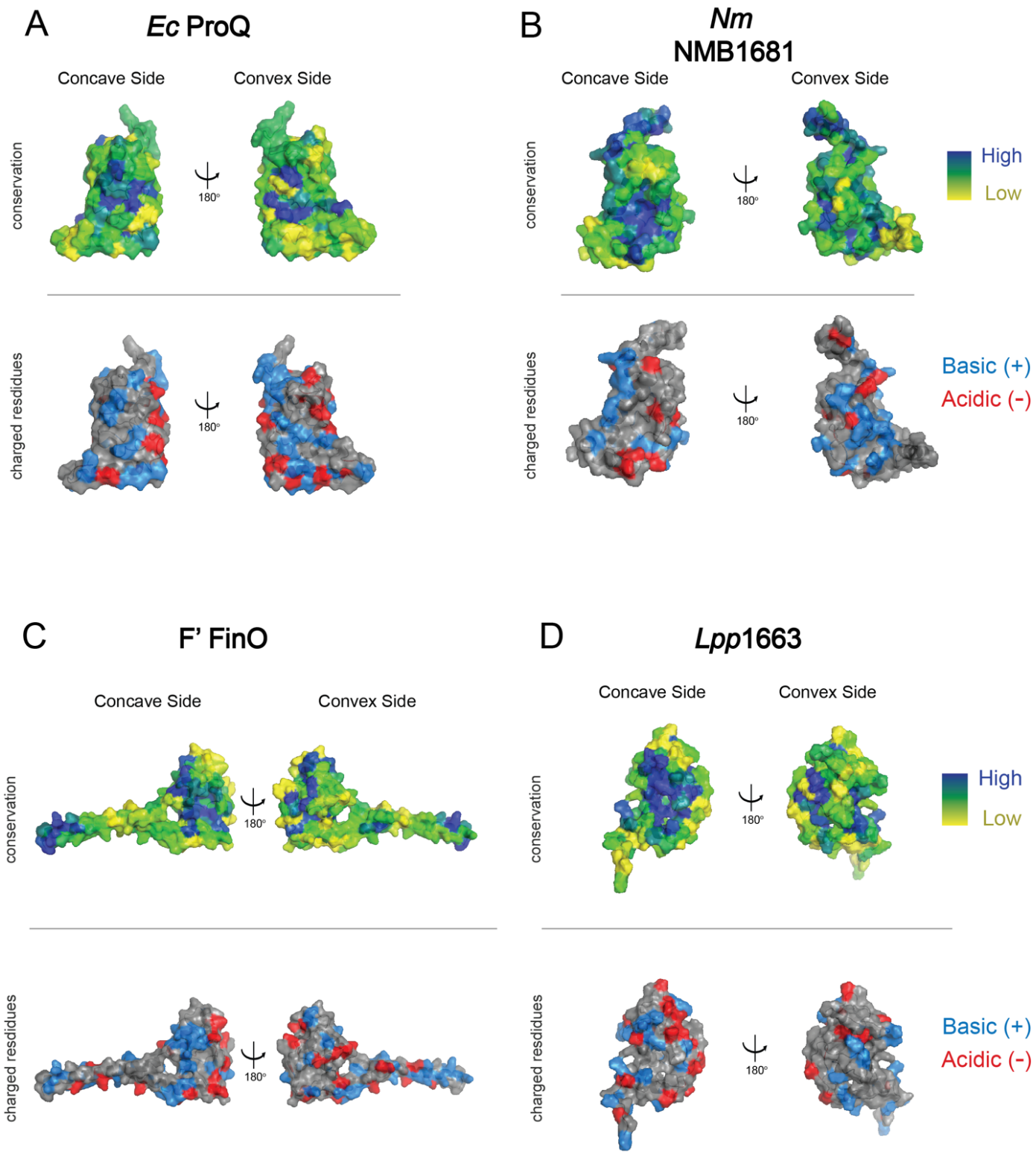


Figure 12: Structures of ProQ and orthologs, with residues colored by conservation and charge. For **A)** *Ec* ProQ NTD NMR structure (Gonzalez et al., 2017; PDB ID: 5nb9), **B)** crystal structure of *Nm* NMB1681 (Chaulk et al., 2010; PDB ID: 3mw6), **C)** crystal structure of *Ec* F' FinO protein (Ghetu et al., 2000; PDB ID: 1dvo), and **D)** NMR structure of *Lp* Lpp1663 (Immer et al., 2020; PDB ID: 6s10). Each structure is colored using ConSurf (Ashkenazy et al., 2016) and residue charge coloring in PyMol (*Show Charged - PyMOLWiki*, 2009). High levels of

conservation are shown in dark blue, with moderate levels of conservation in varied shades of green and minimally conserved residues in yellow. Residues predicted to be positively charged at cellular pH (arginine, lysine, and histidine) are colored blue, neutral residues are gray, and residues predicted to be negatively charged at cellular pH (aspartate and glutamate) are colored red.

Following the release of the AlphaFold database, the same set of figures were made for the ProQ structure predicted by AlphaFold. This structure has distinct concave and convex faces, with a highly conserved patch at the center of the concave face. There is a limited amount of conservation visible on the convex face. The center of the concave face has mostly positively and neutrally charged residues, with some negatively charged residues around the rim of the face. There are negatively charged residues across the convex face of the protein (Figure 13).

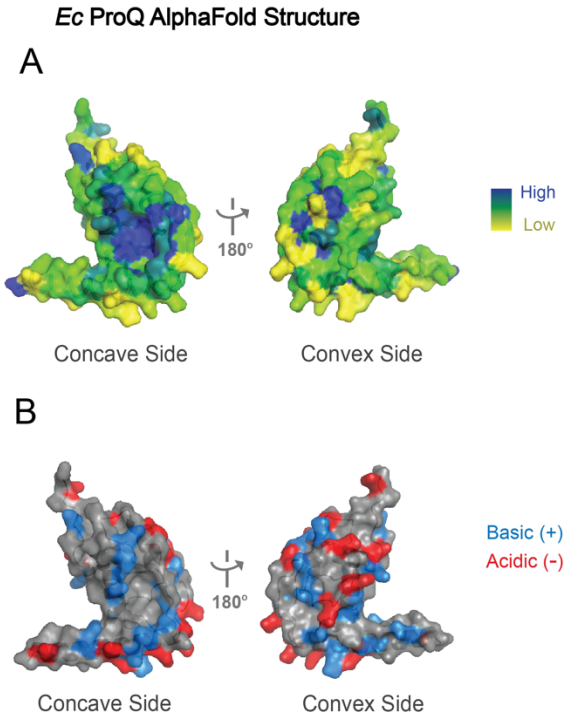


Figure 13: AlphaFold model for ProQ with residues colored by conservation and charge. AlphaFold model of ProQ predicted by Jumper *et al*., 2021, colored using A) ConSurf (Ashkenazy et al., 2016) and B) residue charge coloring in PyMol (*Show Charged - PyMOLWiki*, 2009). As the above figure, high levels of conservation are shown in dark blue, with moderate levels of conservation in varied shades of green and minimally conserved residues in yellow. Residues predicted to be positively charged at cellular pH (arginine, lysine, and histidine) are colored blue, neutral residues are gray, and residues predicted to be negatively charged at cellular pH (aspartate and glutamate) are colored red.

To allow for a more direct comparison between the AlphaFold structure and currently available experimental structures, I created alignments between the AlphaFold structure and all

other available structures using PyMol. For each alignment, the root mean square deviation (RMSD) was recorded. RMSD is a commonly used quantitative measure of similarity between two or more protein structures, which represents the average distance between individual atoms in two different protein structures (Kufareva & Abagyan, 2012). The more similar the structures, the better the two will align and the lower the RMSD values will be. The AlphaFold structure was closest to the FinO structure, with a root mean square deviation (RMSD) of 1.1 Å (83 atoms). It was second closest to Lpp1663, with an RMSD of 1.6 Å (88 atoms). The RMSD between the AlphaFold structure and NMB1681 is 3.5 Å (84 atoms), though this may be skewed higher due to large deviations in the structure towards the top of the protein, while the center of the domains line up closely. Despite being the same amino acid sequence, the ProQ NMR structure has the largest RMSD when aligned with the AlphaFold structure, at 4.2 Å (106 atoms).

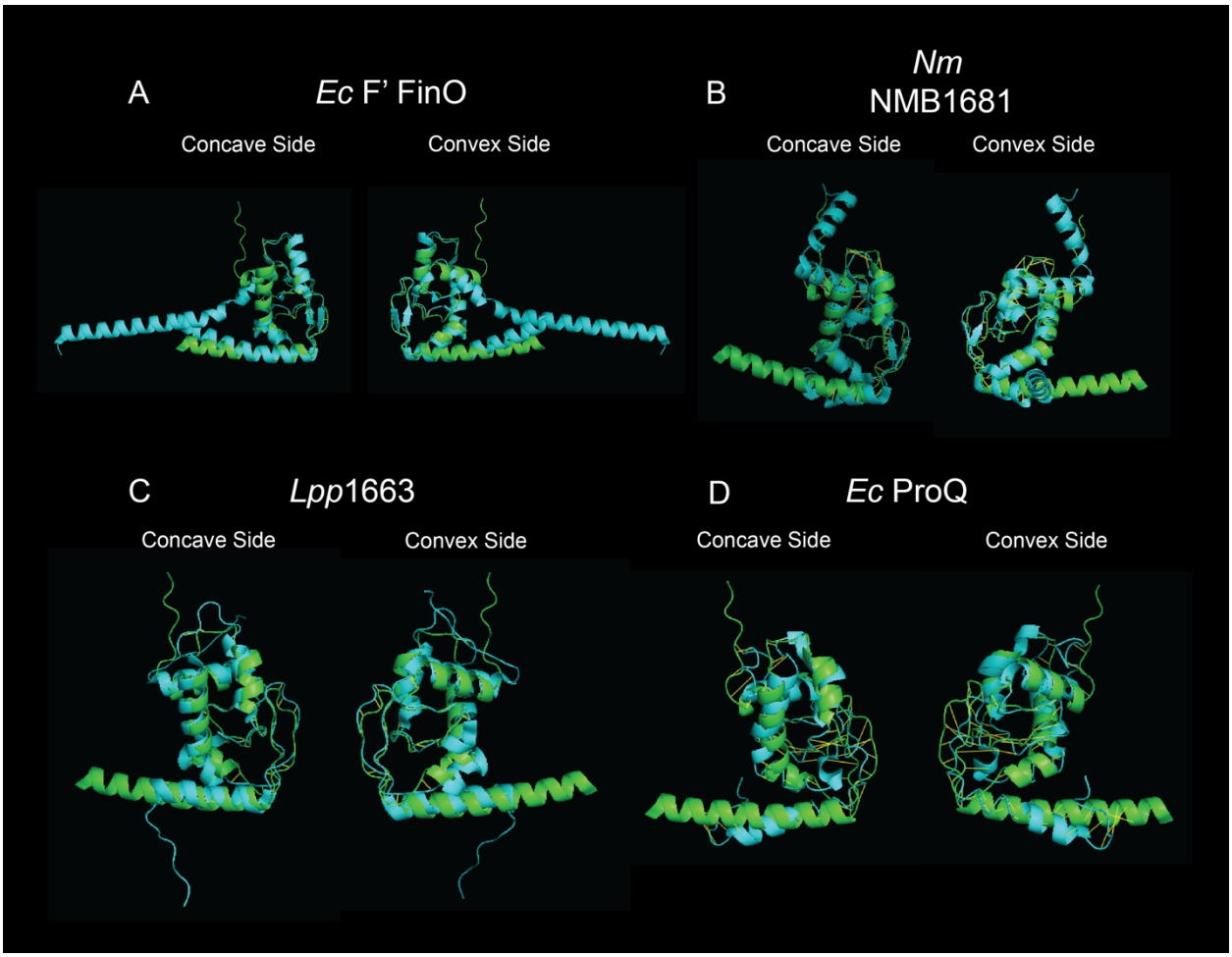


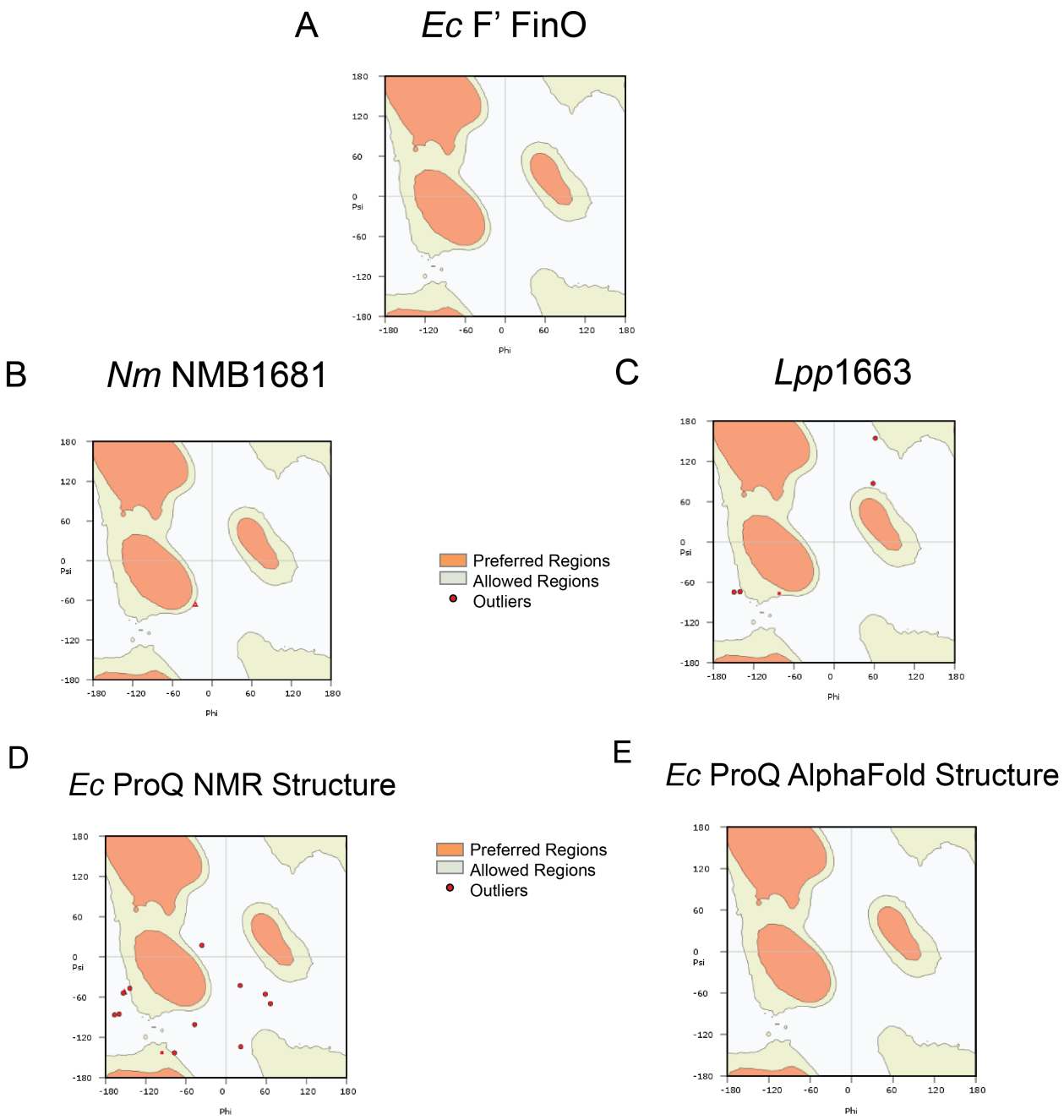
Figure 14: Alignments between the AlphaFold prediction for ProQ NTD and the FinO domain of available structural models. The AlphaFold structure is shown in green, while the other proteins A) *Ec* FinO (Ghetu et al., 2000; PDB ID: 1dvo) B) *Nm* NMB1681 (Chaulk et al., 2010; PDB ID: 3mw6) C) *Lp* Lpp1663 (Immer et al., 2020; PDB ID: 6s10) and D) ProQ (Gonzalez et al., 2017; PDB ID: 5nb9) are shown in a light blue. The distances between amino acid chains are seen in bright yellow. Alignment performed using PyMol.

III-2-ii. Validation

After analysis of general structure, conservation, and charge distribution, differences between the available FinO domain structures were clear. For this reason, we examined and compared the quality of the structures to give an idea of the reliability of each structure. Coot was used to follow up on weaknesses highlighted in the PDB validation report for the NMR structure of ProQ, including clashscore and Ramachandran outliers (*Full WwPDB NMR Structure Validation Report for PDB ID 5nb9*, 2020). Validation measures were performed for

the same structures as above, including *Ec* FinO, *Nm* NMB1681, *Lp* Lpp1663, the NMR structure for *Ec* ProQ, and the *Ec* ProQ structure predicted by AlphaFold.

Ramachandran plots were created for all five structures. The plots shown in Figure 15 display only the outlying residues to give a more readable figure. The percentage of outliers in the two crystal structures, FinO and NMB1681, is incredibly low, with only one outlier in NMB1681 and none in the FinO structure (Figure 15). Furthermore, the great majority of amino acids in these structures fall into the "preferred regions" for angles, with 99% of residues in preferred regions for FinO and 98% of residues within preferred regions for NMB1681. The percentage of amino acids in preferred regions falls substantially for both NMR structures, with *Lpp*1663 at 90% of residues in preferred regions and ProQ at 74%. However, the ProQ structure seems to be markedly weaker than even the other NMR structure in this family of proteins. Not only are 15% fewer of the residues in preferred regions, but the ProQ structure has nearly three times the percentage of outlying residues, at 11% compared to 4.2% for *Lpp*1663. The AlphaFold structure for ProQ performs similarly to the crystal structures for FinO and NMB1681, with 97% of residues in preferred regions and 0 outliers (Figure 15).



F

	F' FinO	*Nm* NMB1681	*Lpp*1663	*Ec* ProQ by NMR	*Ec* ProQ by AlphaFold2
In Preferred Regions (%)	99.33	98.35	90.00	74.36	96.58
In Allowed Regions (%)	0.67	1.50	5.83	14.53	3.42
Outliers (%)	0.00	0.15	4.17	11.11	0.00

Figure 15: Ramachandran plots generated by Coot for FinO family proteins including A) crystal structure of *E. coli* F' FinO protein (PDB ID: 1DVO; Ghetu et al., 2000), B) crystal structure of *N. meningitidis* NMB 1681 (PDB ID: 3MW6; Chaulk et al., 2010), C) NMR structure of *L. pneumophila* Lpp1663 (PDB ID: 6S10; Immer et al., 2020) D) *E. coli* ProQ NTD NMR structure (PDB ID: 5nb9; Gonzalez et al., 2017), and E) the NTD of the *E. coli*

ProQ structure predicted by AlphaFold2 (Jumper et al., 2021). Panel F) Depicts the summary statistics for the Ramachandran plots of each protein, also generated by Coot (Emsley et al., 2010).

In the PDB validation report for the NMR structure for ProQ, clash between amino acids was also identified as a point of weakness. For this reason, I chose to examine clash in all experimentally determined structures for FinO proteins and the AlphaFold structure of ProQ. Coot was used to determine clash through MolProbity, specifically by looking at the output of MolProbity's Reduce and Probe functions (Emsley et al., 2010). The reduce function adds hydrogen atoms to the model in order to more accurately represent the space taken up by each residue. Probe examines all areas within 0.5 Å of the Van der Waals surfaces of atoms to identify overlap between pairs of nonbonded atoms (Chen et al., 2010). When non-donor-acceptor atoms overlap by more than 0.4 Å, Probe denotes this as a "serious clash," (Chen et al., 2010) represented by hot-pink spikes in Coot (Figure 16). Overlaps this significant indicate that two atoms have been modeled in the same place at the same time. This cannot happen in reality, and therefore the presence of such overlap in a structure indicates that at least one of the atoms has been modeled incorrectly.

There is minimal clash in the structures for FinO, NMB1681, and *Lpp*1663, with only a few pink spikes visible. There are many more pink spikes visible on the NMR structure of ProQ, indicating more clash in this structure. It is notable that this structure has more clash than even the other structure which was solved with NMR, the Lpp1663 structure. The AlphaFold structure appears more similar in clash level to the other protein structures, with only one or two small patches of pink spikes visible (Figure 16).

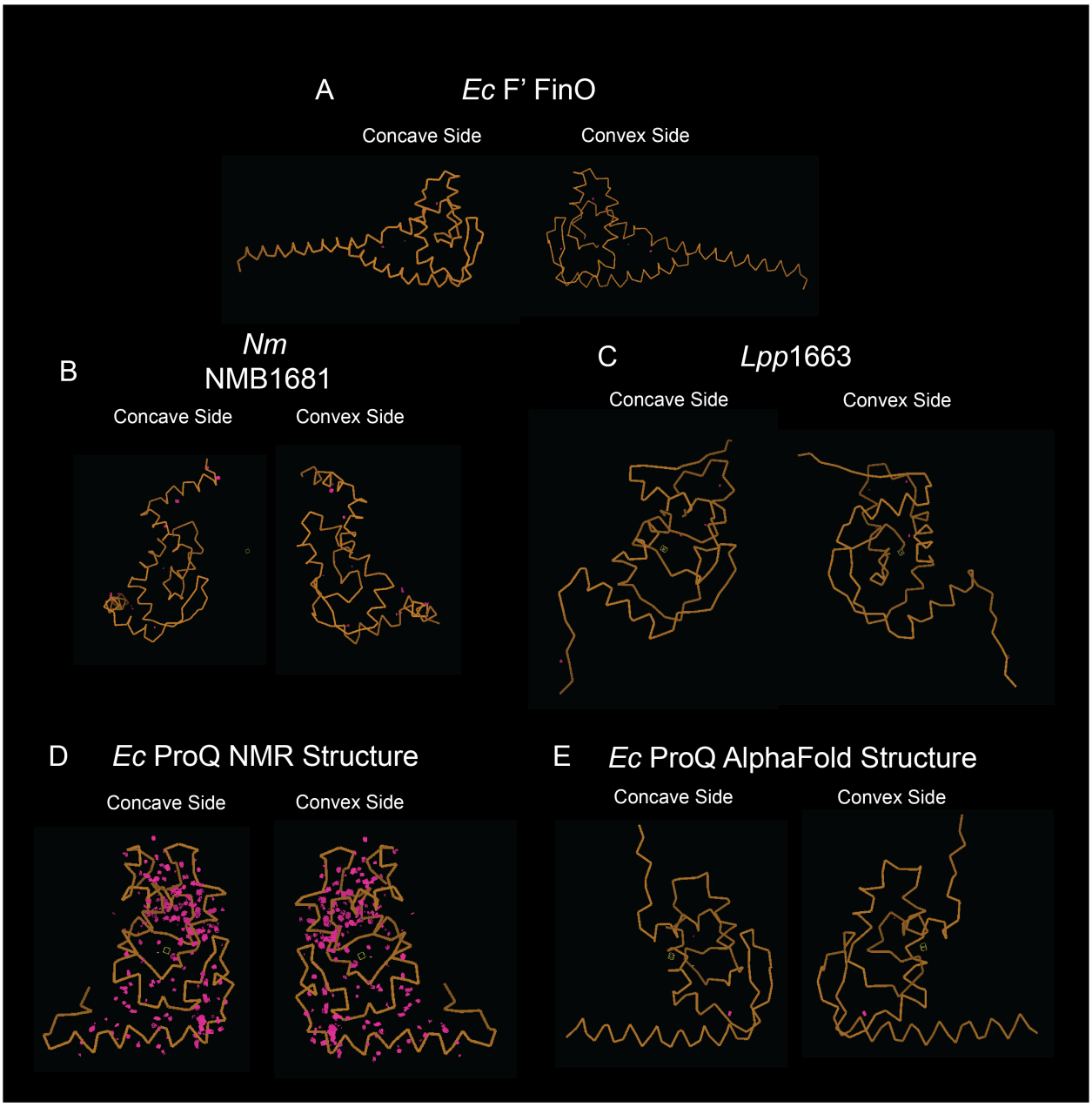


Figure 16: Clash as determined by Molprobity for FinO family proteins including **A)** crystal structure of *E. coli* F' FinO protein (PDB ID: 1DVO; Ghetu et al., 2000), **B)** crystal structure of *N. meningitidis* NMB 1681 (PDB ID: 3MW6; Chaulk et al., 2010), **C)** NMR structure of *L. pneumophila* Lpp1663 (PDB ID: 6S10; Immer et al., 2020), **D)** *E. coli* ProQ NTD NMR structure (PDB ID: 5nb9; Gonzalez et al., 2017) and **E)** the NTD of the *E. coli* ProQ structure predicted by AlphaFold2 (Jumper et al., 2021). The backbones of the protein are shown in orange, while the bright pink hash marks represent "serious clashes," which is when non-donor-acceptor atoms overlap by more than 0.4 Å, indicating that at least one of the two atoms has been modeled incorrectly (Chen et al., 2010; Emsley et al., 2010).

III-3. Site-directed Mutagenesis

III-3-i. Salt bridge between K35 and D41

Once the AlphaFold structure became available, we were able to design experiments targeted at distinguishing between the two available structures. To do this, we identified interactions unique to just one of the structures. I located interactions unique to the AlphaFold structure for ProQ by searching for predicted interactions between side chains in the AlphaFold database and checking for the possibility of the interaction in the NMR structure. This resulted in the discovery of a putative salt bridge between K35 and D41 and a cation-pi interaction between H95 and R80. In order to test for the presence of a salt bridge, I performed a swap mutation by creating a K35D variant of ProQ, a D41K variant of ProQ, and a K35D+D41K variant of ProQ. These variants were compared to each other and variants in which either K35 or D41 had been mutated to alanine. We hypothesized that if the salt bridge was present in the protein, it would be possible to see a loss of binding with the single mutants and the binding would be restored by the double mutant. Prior to running the experiment, there was an understanding that swapping the charges may cause a significant enough disruption in the protein folding process that the double mutant would not be functional.

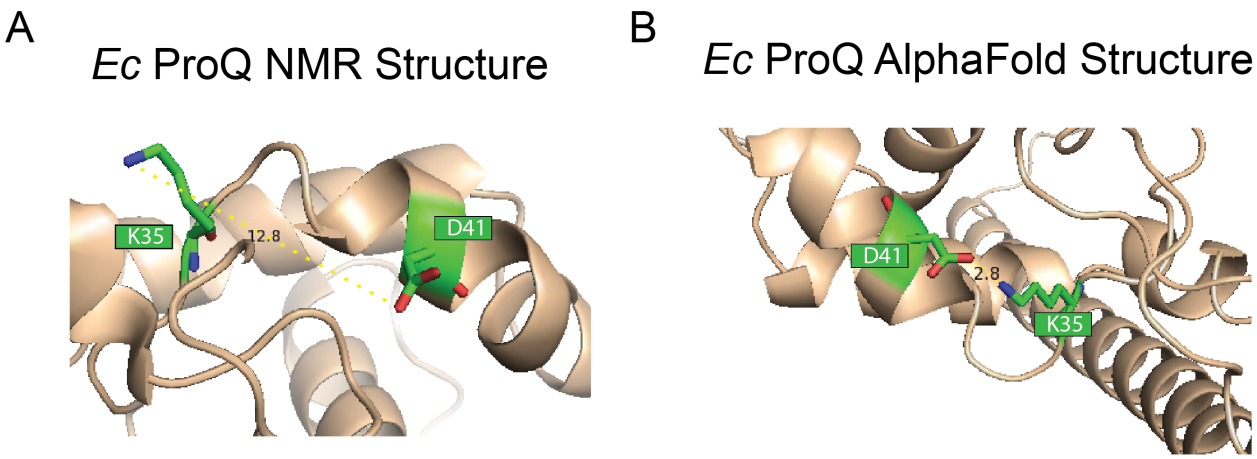


Figure 17: Potential salt bridge between K35 and D41 shown in the AlphaFold ProQ structure. Depicted above are the distances between atoms at the end of K35 and D41 side chains in **A)** the NMR structure for ProQ solved by Gonzalez *et al*., 2017 and **B)** the structure predicted by AlphaFold2 (Jumper et al., 2021). The difference in distances suggests that while these residues could interact in the AlphaFold2 structure, they would not in the NMR structure. It should be noted that the PDB entry for the NMR structure includes 17 possible structures for ProQ, across which the distances between the end of these side chains ranges from 2.6 Å to 13.4 Å, with an average distance of 9.2 Å between K35 and D41 (for the full list of distances between the end of the side chains by model, see Supplementary Table 1 in Appendix) . In other words, the NMR file includes some models in which this interaction could happen, but it is not the case for the majority of the models.

Mutations cloned into the NTD of ProQ were tested in both liquid and plate-based bacterial three-hybrid assays. Both K35A and D41A mutations led to a loss of binding to all three RNA substrates (Figure 18) which is in line with previous results for *cspE* and SibB (Pandey et al., 2020). Single mutations for the swap led to an even greater loss of binding than the mutations to alanine, which is especially clear in the tests against *malM* (Figure 18). In the liquid assay, the double mutant did not restore ProQ binding to any of the RNA substrates (Figure 18). In order to examine the potential impact of the stability of ProQ variants on observed differences in interaction levels, we performed an immunoblot experiment. A preliminary immunoblot (not yet repeated) showed that all ProQ variants were stably expressed, which can be observed both visually and in the densitometry data (Figure 18D, Supplementary Figure 1).

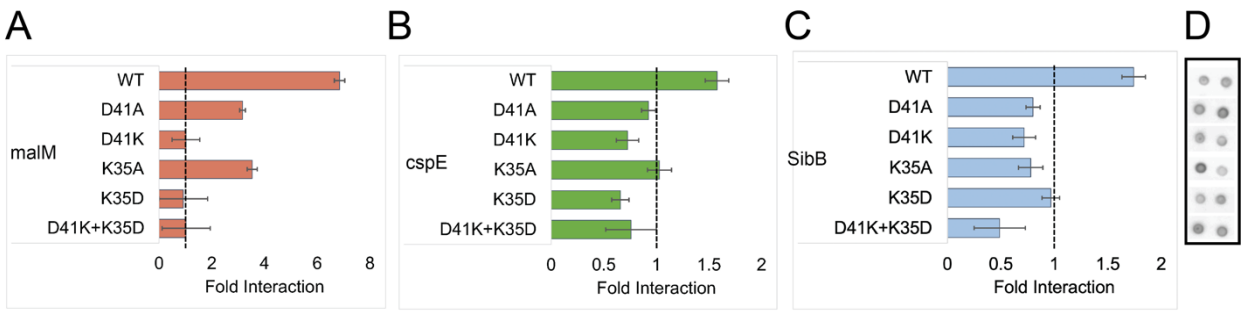


Figure 18: Impact of mutation to residues K35 and D41 on ProQ binding of RNA targets *malM*, *cspE*, and SibB. In this figure, the red bars in A) represent the fold interaction observed with *malM*, green bars in B) represent fold interaction observed with *cspE*, and blue bars in C) represent fold interaction observed with SibB in the bacterial three-hybrid system. The gray dashed line shows one fold interaction on the graph. This data shows that mutation to residues D41 and K35 significantly impacts binding, even for a mutation to alanine. The impact on binding is more dramatic when the residues are mutated to a residue of the opposite charge, as for the D41K and K35D variants. The D41K+K35D variant does not show an increased ability to bind to the RNA targets when compared to the single mutants at those residues. Results shown here are an average of three independent experimental conditions with standard deviation. This experiment was performed on three separate days, and this data is representative of that found each time. Panel D) shows preliminary results from an immunoblot experiment which indicates that the mutant proteins were expressed stably in the cell. The spots in panel D are ordered from top to bottom to match the corresponding B3H data for that particular variant. Complete immunoblot data, including densitometry analysis of the membrane, can be seen in Supplementary Figure 1 in the Appendix. Supplementary Figure 1 shows the membrane spots for alpha-empty, which are blank.

The plate-based assay for this experiment offers a slight hint of an increase in RNA binding by the K35D+D41K variant ProQ protein, relative to the K35D variant protein. In Figure 19, the patches for the protein with a double mutation appear to be more blue than the patches for K35D alone. This suggests that the K35D+D41K variant of ProQ may be able to bind to RNA better than the K35D variant. One possible explanation for this impact seen on the plate but not in liquid is a difference in the β-galactosidase activity levels in the negative controls corresponding to different mutant proteins. Calculations of fold interaction involve dividing the experimental condition by the highest negative control for that condition, where the negative controls are conditions in which one component of the assay has been replaced by an empty vector.

To explore this possibility, I plotted raw β-galactosidase data for the liquid experiment. In this case, the β-galactosidase activity for the CI-empty condition for the K35D+D41K variant of ProQ is higher than for other prey protein constructs, particularly when compared to the negative

controls for the single mutants, K35A and D41A (Figure 20). This could lead to a lower calculated fold interaction when in reality the absolute β-galactosidase activity is comparatively the same or even better. For further analysis of differences between plate and liquid data, see Discussion.

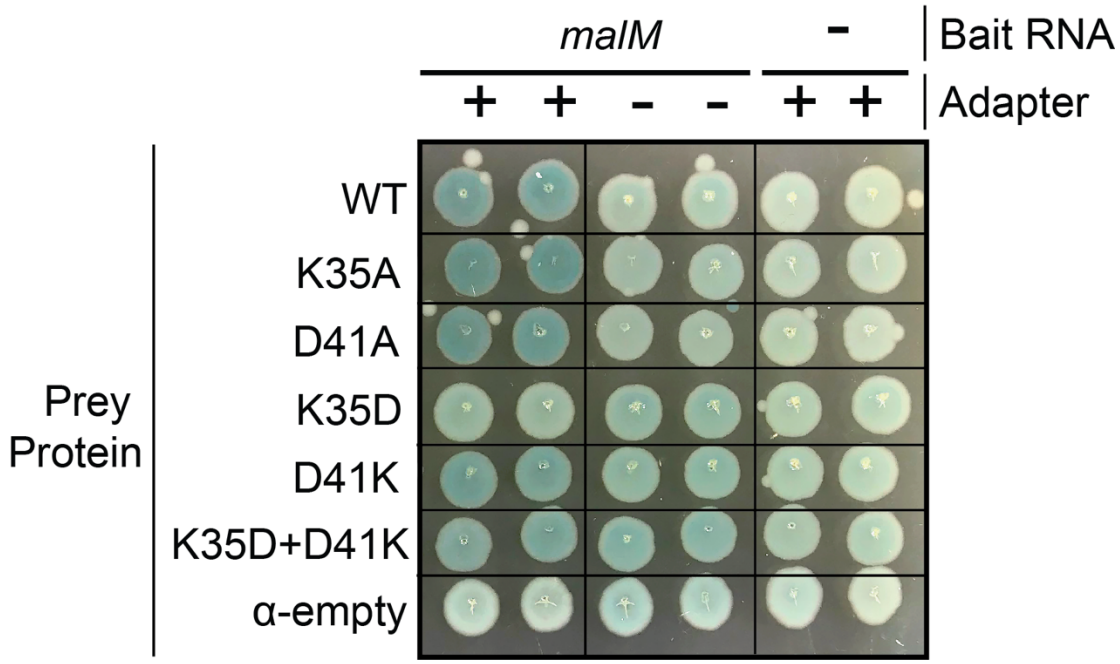


Figure 19: Plate-based assay for K35 and D41 variants with *malM*. Transformations which include all components of the assay are shown in the leftmost column, followed by transformations with an empty adapter construct in the center column and transformations with an empty bait construct in the rightmost column. Transformations with an empty prey construct are shown in the bottom row. The top row shows the resulting growth on plates for WT ProQ NTD (pKB955, see Methods), while the subsequent rows show growth resulting from a mutated ProQ NTD, labeled accordingly. This experiment was performed on two separate days, and this data is representative of that found each time.

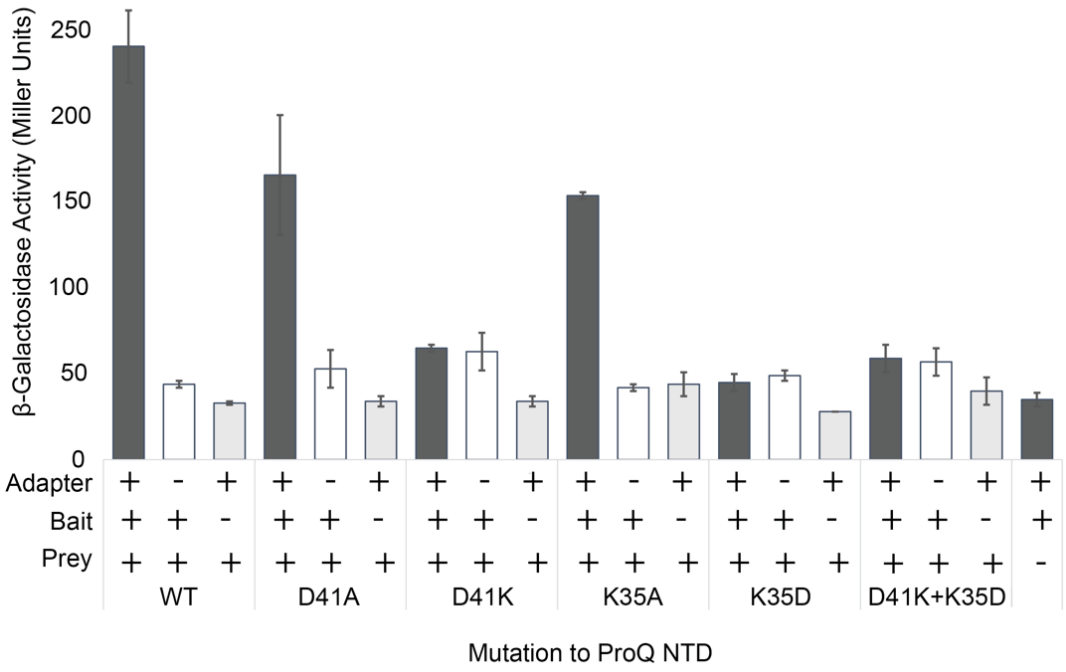


Figure 20: Raw β-galactosidase activity from liquid B3H with K35 and D41 variants. In this figure, the darkest bars represent β-galactosidase activity observed in the experimental conditions, while white bars represent β-galactosidase activity seen when the adapter component of the assay is replaced with an empty construct, and light gray bars show β-galactosidase activity observed when the bait RNA component is replaced with an empty construct. For calculations of fold interaction, the β-galactosidase activity seen in the experimental condition is divided by the highest β-galactosidase activity seen in a negative control (either CI empty, MS2 empty, or α-empty). For this reason, higher negative controls can lead to lower fold interactions for the same absolute β-galactosidase activity production. Results shown here are an average of three independent experimental conditions with standard deviation. This experiment was performed on three separate days, and this data is representative of that found each time. The data shown here corresponds to the data used to calculate fold interaction above. This particular set of data is focused on the findings observed with *malM* for clarity. *malM* was selected due to the high level of interaction with ProQ in our assay, which allows for the detection of more subtle differences in binding.

III-3-ii. Cation-pi interaction between H95 and R80

In order to test for the presence of a cation-pi interaction between H95 and R80 (shown in Figure 21) , I again performed a swap mutation as well as additional mutations to *ProQ*. With the knowledge that ProQ does not tolerate mutations to R80 established in the library screen, these mutations focused on H95. The histidine side chain has multiple chemical traits which could be leading to interaction with R80: the presence of an aromatic ring and the ability to form hydrogen bonds. For this reason, I elected to separate some chemical traits to be probed individually. To do so, I created both an H95F and H95Q ProQ variant. The mutation to

phenylalanine maintained the aromatic ring seen in histidine. Glutamine was selected over other amino acids which would maintain hydrogen bonding ability, such as glutamate or aspartate, due to the length of the side chain and the ability to form multiple hydrogen bonds, which we hoped would increase the chances of creating a variant protein which could maintain the predicted interaction with arginine. Prospective mutations to test these features were first made in the AlphaFold structure in PyMol (as shown in Figure 22) to check for appropriate geometry.

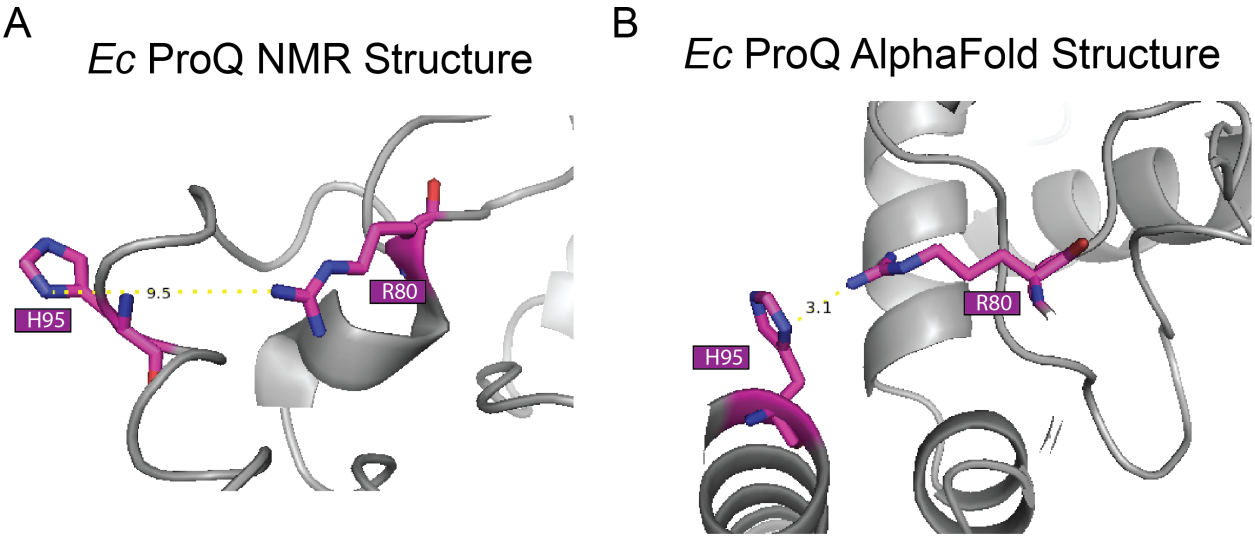


Figure 21: Potential cation-pi interaction between R80 and H95 shown in AlphaFold ProQ structure. In A) the NMR structure (Gonzalez et al., 2017), side chains are not close enough to interact while they are in B) the structure predicted by AlphaFold2 (Jumper et al., 2017). For the 17 proposed structures in the NMR file, these two side chains range from 9.5 Å apart to 10.9 Å apart, with an average distance of 10.0 Å between the two (for the full list of differences between the ends of the side chains in each model, see Supplementary Table 1 in Appendix). In every structure, these residues are too far to interact.

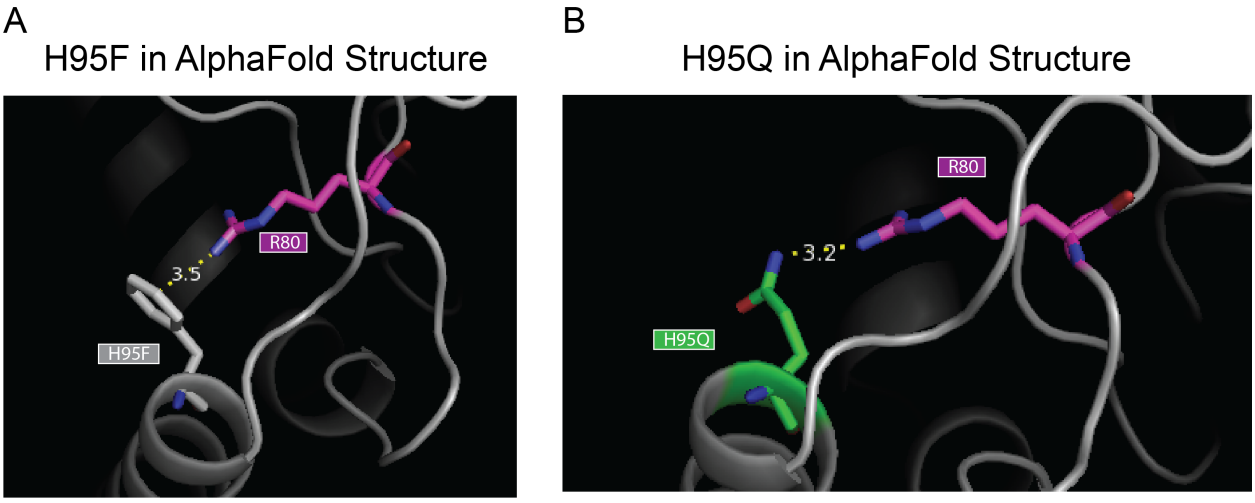


Figure 22: Proposed mutations to H95 in ProQ. In order to probe the potential cation-pi interaction between R80 and H95, I proposed variants that would maintain either A) the aromatic ring found in a histidine side chain, by mutating it to phenylalanine, or B) the ability of a histidine side chain to form hydrogen bonds, by mutating it to glutamine. Both mutations are only shown in the AlphaFold structure, as the previous figure for this interaction demonstrated that the interaction between H95 and R80 is not proposed to take place in the NMR structure.

The effects of histidine substitutions were evaluated in both liquid and with the plate-based assay. Surprisingly, mutating the histidine residue to alanine only had a limited impact on ProQ's binding of RNA targets (Figure 23, Figure 24). Since a mutation to alanine can be thought of as effectively removing a side chain by minimizing length and chemical properties, this data indicates that it is unlikely that any interaction involving H95 has an impact on ProQ's ability to bind RNA. An H95R mutation led to a slightly greater loss of binding with *malM* and SibB, but not *cspE* (Figure 23). Any mutation which included mutating R80 had a complete loss of binding (Figure 23). A swap mutation was not able to rescue the RNA-binding ability of ProQ (Figure 23). This is supported clearly in the plate-based assay as well, where the patches for variants with an R80 mutation are much more white than patches grown with other prey proteins (Figure 25). The H95F variant demonstrated a greater loss in binding when compared to other H95 variants (Figure 24). This data is supported in the plate-based assay as well, where the patch grown with the H95F variant seems to be a slightly lighter blue than the other H95 variants (Figure 25). The H95Q variant showed a level of interaction similar to the H95A variant (Figure 24).

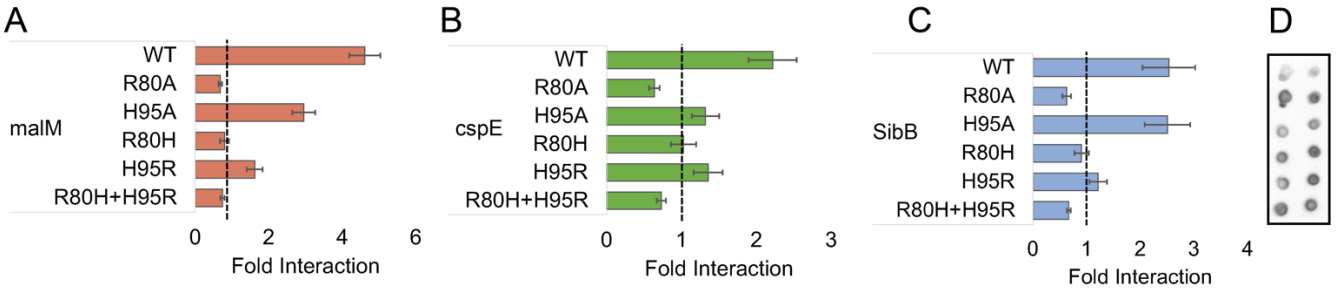


Figure 23: Impact of mutation probing a potential H95 and R80 interaction on ProQ binding of RNA targets *malM*, *cspE*, and SibB. In this figure, the red bars in A) represent the fold interaction observed with *malM*, green bars in B) represent fold interaction observed with *cspE*, and blue bars in C) represent fold interaction observed with

SibB in the bacterial three-hybrid system. The gray dashed line indicates one fold interaction on the graph, which can be taken as a basal level of interaction. Surprisingly, mutations to H95 and namely H95A did not impact binding in a major way. As a mutation to alanine can be thought of as removing a side chain, this indicates that any interaction involving H95 in ProQ does not have a significant impact on ProQ's ability to bind to RNA. Results shown here are an average of three independent experimental conditions with standard deviation. This experiment was performed on three separate days, and this data is representative of that found each time. Panel D) shows preliminary results from an immunoblot experiment which indicates that the mutant proteins were expressed stably in the cell. The spots in panel D are ordered from top to bottom to line up with the corresponding B3H data for that particular variant. Complete immunoblot data, including densitometry analysis of the membrane, can be seen in Supplementary Figure 2 in Appendix. Supplementary Figure 1 shows the membrane spots for alpha-empty, which are blank.

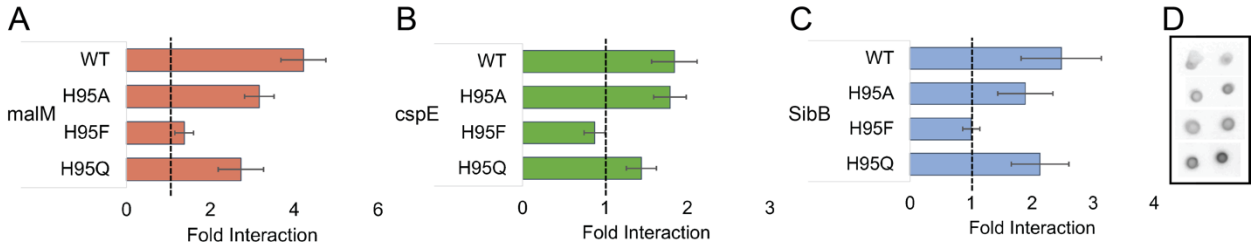


Figure 24: Impact of mutations to H95 on ProQ binding of RNA targets *malM*, *cspE*, and SibB. In this figure, the red bars in A) represent the fold interaction observed with *malM*, green bars in B) represent fold interaction observed with *cspE*, and blue bars in C) represent fold interaction observed with SibB in the bacterial three-hybrid system. The gray dashed line indicates one fold interaction on the graph, which can be taken as a basal level of interaction. An H95F mutation impairs binding when compared to H95A or H95Q. The H95A variant is able to bind to RNA at comparable levels to the H95Q variant. Results shown here are an average of three independent experimental conditions with standard deviation. This experiment was performed on two separate days, and this data is representative of that found each time. This particular experiment was only repeated twice, rather than three times, due to an error in data collection in the final step of the attempted third repeat. Panel D) shows preliminary results from an immunoblot experiment which indicates that the mutant proteins were expressed stably in the cell. The spots in panel D are ordered from top to bottom to line up with the corresponding B3H data for that particular variant. Complete immunoblot data, including densitometry analysis of the membrane, can be seen in Supplementary Figure 2 in Appendix. Supplementary Figure 1 shows the membrane spots for alpha-empty, which are blank.

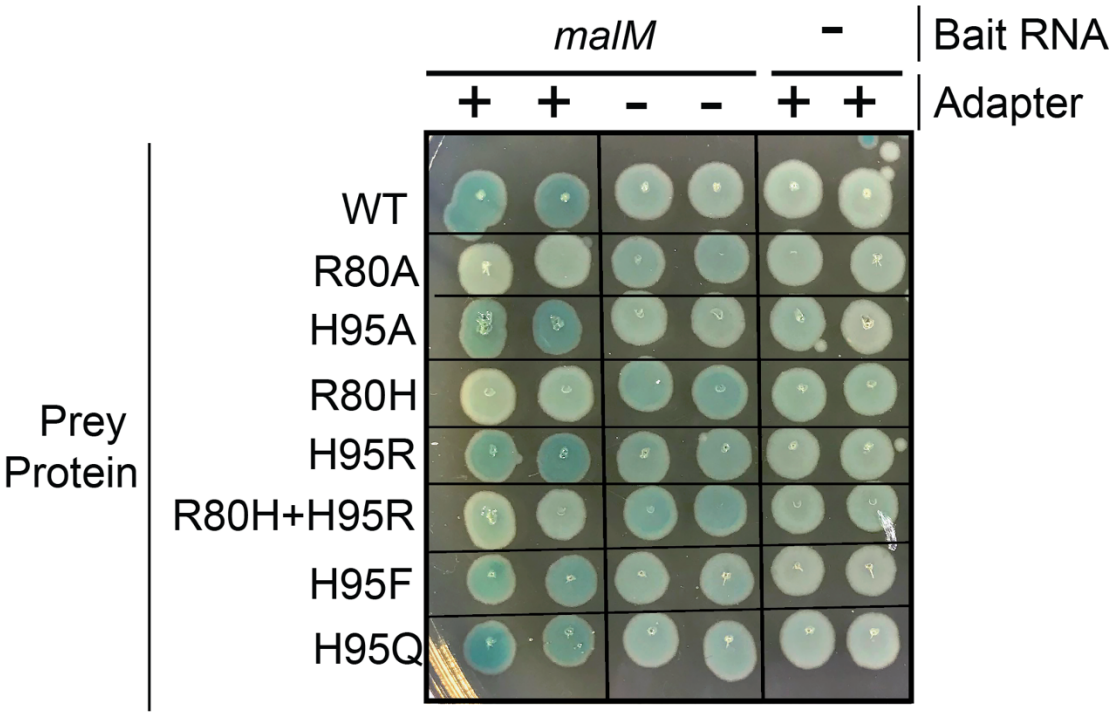


Figure 25: Plate-based assay for H95 and R80 variants with *malM*. Transformations that include all components of the assay are shown in the leftmost column, followed by transformations with an empty adapter construct in the center column and transformations with an empty bait construct in the rightmost column. The top row shows the resulting growth on plates for WT ProQ NTD (pKB955, see Methods), while the following rows show growth resulting from a mutated ProQ NTD, labeled accordingly. Both R80 variants, R80A and R80H, result in lighter colonies. The patch with the H95F variant appears to be slightly lighter than the other H95 variants, H95A and H95R. This experiment was repeated twice, and this data is representative of that found each time.

Like for the K35 and D41 variants, the stability of these variants *in vivo* was examined a single time with the use of an immunoblot. Again, based on this preliminary data all variants were stably expressed in cells. This is seen in images of the membrane processed with a chemiluminescent imager (Figure 18) and corresponding densitometry data (Appendix: Supplementary Figure 2).

Chapter IV: Discussion

In this work, I have presented an analysis of two different structural models of the bacterial RNA binding protein ProQ. By running a saturation mutagenesis screen to search for any other amino acids which supported RNA binding at position 70 or 80 in ProQ, I determined that no amino acids aside from tyrosine and arginine, respectively, worked at these positions. ProQ was found to vary in chemical qualities, including distribution of charge and conservation, when compared to the structural models of other FinO-domain proteins. Finally, using site-directed mutagenesis I highlighted the importance of residues predicted to be involved in an interaction in the AlphaFold model of ProQ, but not predicted to form such an interaction in the NMR structure. Overall, my analysis highlights inconsistencies between the currently available data on ProQ and the structural model used by the field. The hope of this work is to support the future investigation of the RNA binding mechanism of ProQ by both calling attention to potential issues with the NMR structure for ProQ and bringing forth a structure that appears to be more in line with currently available data on ProQ and other FinO-domain proteins.

IV-1.Tyrosine and arginine are strictly required at positions 70 and 80 in ProQ

This work began with the R80X and Y70X library screens. After covering both the Y70X and R80X library more than fifteen times, no other amino acid was found to function at either position. This screen was performed only with *malM*, but it is likely that these results would apply to other sRNA targets of ProQ based on previous work. In studies performed with the same B3H assay, both Y70 and R80 appeared as a residue necessary for binding in screens performed with the *cspE* 3'UTR and SibB (Pandey et al., 2020) and even conservative mutation to these residues led to fold interaction below one with SibB and *cspE* 3' UTR (Pandey et al., 2020; Stockert, 2021).

An important question regarding the validity of our saturation mutagenesis screen is the representation of all amino acids in the library. All codons for the wild type amino acids were returned, however, there was an over-representation of the codons found in the original (template) plasmids seen in the sequencing results (see Methods). It is possible that some of this is due to bias introduced into the library either in PCR or in the growth of colonies during the library preparation. In the PCR tube, the primer which perfectly complemented the wild-type template could have annealed slightly better than the other primers, leading to more replication of that specific sequence. This trend would explain why there was bias to one tyrosine codon and one arginine codon over the others (Table 7, Table 9). It is also possible that cells that were transformed with wild type ProQ following KLD treatment had a slight growth advantage, leading to overrepresentation in the library preparation.

While it is clear that the library did contain some bias, the fact that each codon for the wild-type amino acid was returned suggests that the screen was saturated. This demonstrates that the methodology used was able to identify mutants of ProQ which could interact with mRNA target *malM*, as it is unlikely that sequences with alternate codons for the wild-type amino acids had the same advantages the original sequence did in the PCR reaction.

This screen demonstrated that functional ProQ always had tyrosine at position 70 and arginine at position 80, suggesting that the unique chemical properties of Y70 and R80 are important. In order for mutations to these residues to impact levels of interaction between RNA and protein observed in the B3H, the mutations must disrupt ProQ's interaction with RNA either directly or indirectly through disruption of the overall structure and function of the protein. ProQ has been shown to be stably expressed in the cell with a mutation at each of these positions (Pandey et al., 2020). If mutant proteins were severely misfolded in the cell, the proteins would

be degraded by the cell machinery. It is therefore unlikely that the loss in binding observed is due to misfolding alone. Additionally, Y70 and R80 are located on the surface of the protein in both models (Gonzalez et al., 2017; Jumper et al., 2021). In contrast, key structural residues are often found deep in the core of the protein. The loss of binding seen in proteins with a mutation at either Y70 or R80 is therefore more likely because these residues are directly involved in interaction with RNA.

An Argonaute (Ago) protein from *Aquifex aeolicus* (*Aa*) provides an example of tyrosine and arginine side chains interacting directly with RNA (Yuan et al., 2006). Ago proteins are found in processes of gene regulation mediated by small RNA molecules in both eukaryotes and prokaryotes. In this process, the Ago protein binds to a small "guide" RNA in order to identify specific RNA targets which have sequences complementary to the guide (Lisitskaya et al., 2018). The role of the protein in encouraging complementary base pairing is functionally analogous to Hfq, FinO, and perhaps ProQ. The *Aa* Ago protein has a solved crystal costructure with a regulatory small interfering RNA (siRNA). The structure shows a tyrosine, Y119, and an arginine, R123, pi-pi stacking and forming a cation-pi interaction with the final base pairs of the double-stranded RNA (Figure 26; Yuan et al., 2006). It is possible that ProQ would be able to form a similar interaction if it folds as the AlphaFold structure suggests. In the AlphaFold structure for ProQ, residues Y70 and R80 are a similar distance apart in 3D space as residues Y119 and R123, at 5.5 Å and 4.0 Å, respectively.

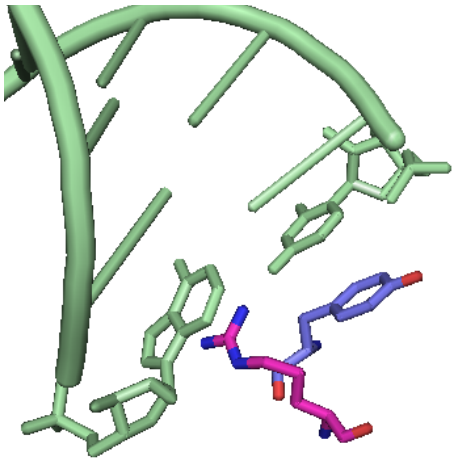


Figure 26: Costructure of Argonaute (Ago) protein with an siRNA in *Aquifex aeolicus.* The structure shows the terminal RNA bases pi-pi stacking with a tyrosine, Y119, and forming a cation-pi interaction with an arginine, R123 (Yuan et al., 2006). This costructure offers an example of how tyrosine and arginine residues may be involved in binding double-stranded RNA.

Furthermore, the position of residues homologous to Y70 and R80 are conserved across the FinO domains with solved structures. In all structures with the exception of the NMR structure for ProQ, these two residues are close to each other in the center of a highly conserved patch (Figure 27, conservation can be seen in Figure 12). In the non-ProQ structures, this is also a distinct concave "pocket" on the protein. The location of these residues on *Lpp*1663 was not missed by Immer *et al*., the authors who solved the NMR structure for that protein. Citing the establishment of Y70 and R80 as critical to ProQ's binding of RNA in Pandey *et al*. (2020), the authors made single point mutations to alanine at homologous residues Y76 and R86 in *Lpp*1663. These variants were confirmed to fold correctly with NMR spectroscopy. However, when either residue was mutated to alanine, RNA binding was lost completely. This suggests that these two residues have a role in RNA binding by Lpp1663, which is a key part of the data which leads the authors to conclude that the concave face of Lpp1663 is the primary site of RNA binding (Immer et al., 2020).

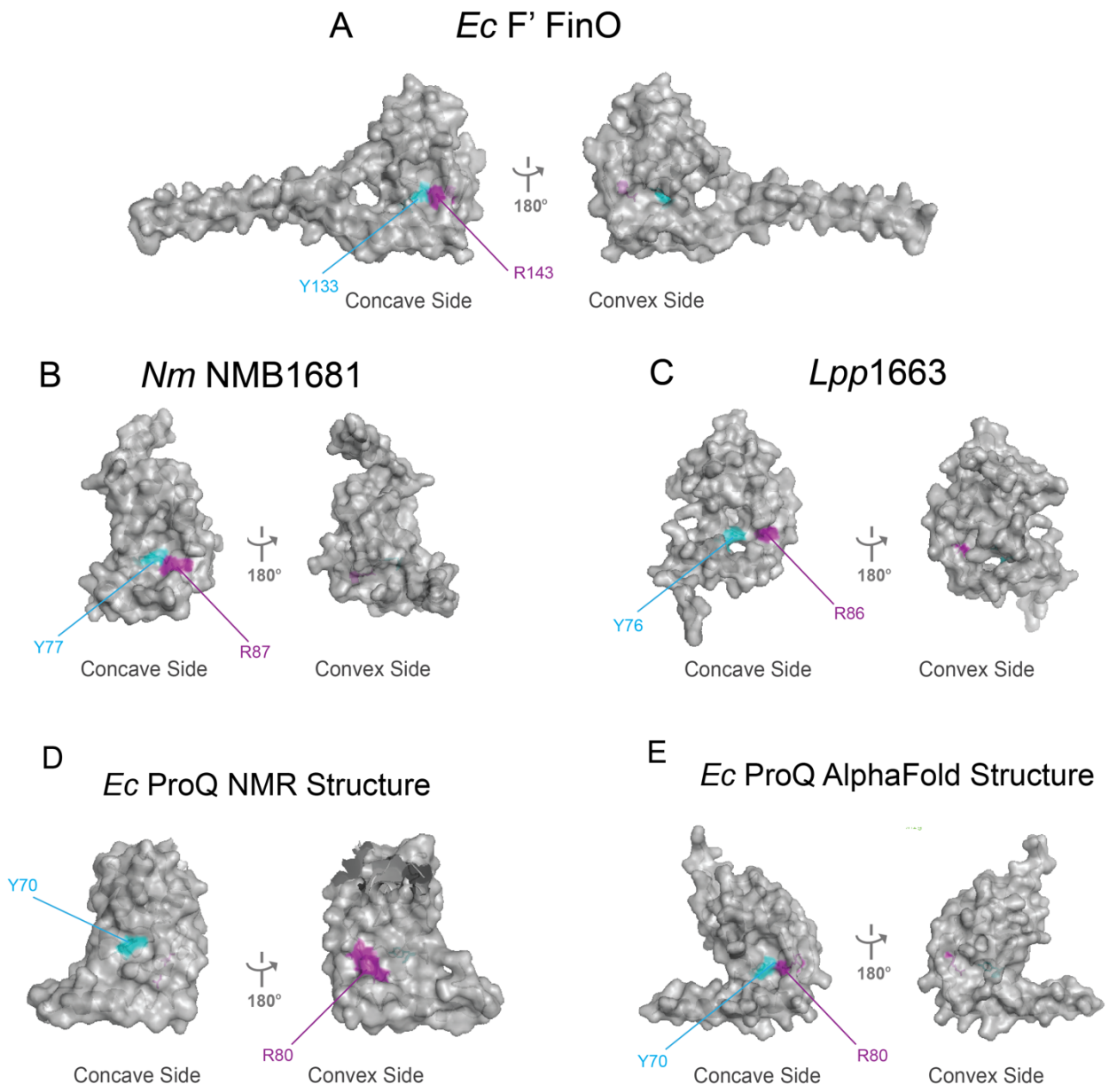


Figure 27: The location of Y70/R80 residues and homologous residues in available structures for FinO family proteins. Y70 and homologous residues have been colored bright blue. R80 and homologous residues have been colored purple. The remaining residues have been colored grey. Depicted are A) FinO (PDB ID: 1DVO; Ghetu et al., 2000), B) NMB1681 (PDB ID: 3MW6; Chaulk et al., 2010), C) Lpp1663 (PDB ID: 6S10; Immer et al., 2020), D) the NMR structure for ProQ's NTD PDB ID: 5nb9; Gonzalez et al., 2017) and E) the AlphaFold structure for ProQ's NTD. On the NMR structure for ProQ, Y70 is located in the center of the concave face of the protein, and R80 is located slightly to the left of the convex side of the protein. For all other structures, Y70 and R80 are located directly next to each other at the center of a pocket on the concave face of the protein.

IV-2. Ortholog's FinO domain differ from ProQ's in both quality and chemical characteristics

The available structures of FinO domains (in FinO, NMB1681, and Lpp1663) vary from the NMR structure for ProQ in several ways. Firstly, these structures all have more distinct concave and convex faces, with patches of high levels of conservation in the center of the concave face (Figure 12). This supports the idea of RNA binding in the center of the concave face of these proteins, as amino acids needed for the proper protein function (such as binding of RNA in this case) are more likely to be held onto by evolution. The non-ProQ proteins also all have residues of positive or neutral charge across the concave face of the protein, with negatively charged residues around the rim of the concave face of the protein and across the convex face. It is possible that negatively charged residues repel the negatively charged phosphate backbone of the RNA. The ring of negatively charged residues on the concave face may guide the RNA to the center of that face for binding. Neither the pattern of conservation nor the pattern of charge distribution seen in other structures is present on the NMR structure for ProQ. However, as these proteins are all orthologs with a shared domain (Olejniczak & Storz, 2017) and orthologs are expected to adopt similar folds in the cell (Rost, 1999), these structures of other proteins should be taken seriously in the search for an accurate model of RNA binding by ProQ. Furthermore, the majority of these structures are comparable to each other well in conservation and charge distribution while the ProQ structure is an outlier.

These differences have been previously highlighted by Immer *et al*. in the paper revealing the structure for Lpp1663. The structure of this FinO-domain protein was determined with NMR in 2020. The authors aligned the lowest energy structure from the NMR file, often selected as the most likely structure due to the role of energy minimization in protein folding in the cell (Nelson

& Cox, 2012), to the structures of FinO domains available at the time. Immer *et al*. compared the structures using through alignments with recorded RMSD values. For Lpp1663, the RMSDs were 1.8 Å when aligned to the FinO protein, 1.9 Å when aligned to the NMB1681 protein, and 4.2 Å when aligned to the ProQ NMR structure (Figure 28). As the authors highlight, this result is surprising as ProQ has the highest sequence identity to Lpp1663, at 34%, compared to 28% for FinO and 25% for NMB1681. This means that Lpp1663 shares the most similarity in sequence to ProQ, which would suggest the fold of these two proteins would be the closest. The RMSD does not align with this, by showing that the available NMR structure for ProQ is the most different from Lpp1663 among all the FinO protein structures.

The structure of ProQ is not only inconsistent with that of Lpp1663, but also with the other available FinO domain structures. The NTD of ProQ (PDB ID: 5nb9), which is the FinO domain, has high RMSD values when aligned with either of the available X-ray structures, at 4.5 Å for FinO and 5.6 Å for NMB1681 (Immer et al., 2020). These findings complement the comparison between structures discussed here which show that the NMR structure of ProQ is quite different from the available structures of other FinO family proteins in shape, conservation, and charge distribution. The AlphaFold structure is more similar to the available structures, not only in the overall structure, location of key residues, charge distribution, and conservation distribution, but as determined with the same RMSD measurement. Our analysis showed that the AlphaFold structure for ProQ's NTD is very close to the FinO and Lpp1663 structures, and in fact has the highest RMSD when compared to the NMR structure for ProQ's NTD (PDB ID: 5nb9). Despite having the same chain of amino acids, the AlphaFold structure adopts a fold that is more similar to that seen in other FinO-domain structures.

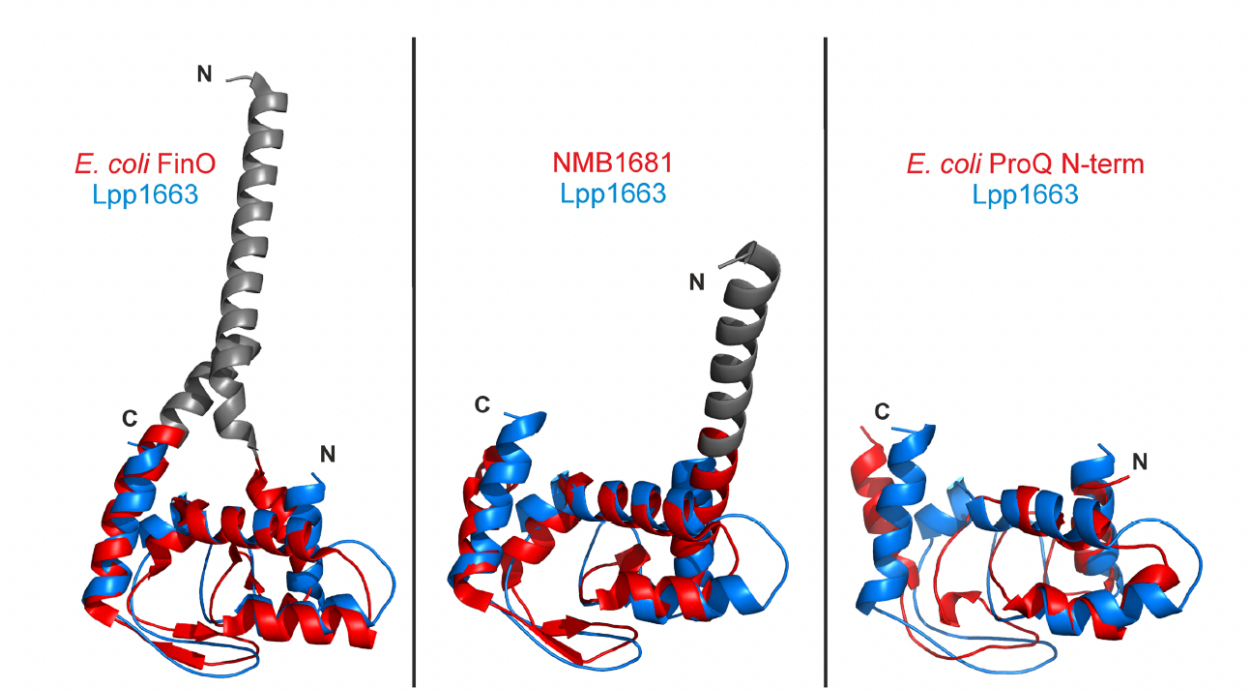


Figure 28: The NMR structure of Lpp1663 aligned with other solved structures for FinO-domain proteins. The lowest energy Lpp1663 structure from the NMR structure file is shown in blue, superposed with the structures of other FinO/ProQ domains colored in red. The N terminal domains of FinO and NMB1681 are colored in gray. For ProQ, only the FinO domain (which is the N-terminal domain) is shown. Lpp1663 closely matches the crystal structures for FinO and NMB1681, with RMSD values of 1.8 Å and 1.9 Å respectively. The structures of Lpp1663 and ProQ do not align as closely, with an RMSD value of 4.2 Å. Figure from Immer *et al*., 2020.

The differences in structure prompted me to look at the quality of the structures. The structure validation work done here indicates that the quality of the non-ProQ structures are greater than the quality of the NMR structure for ProQ. The non-ProQ structures have a lower percentage of Ramachandran outliers and less steric clash. Ramachandran outliers represent bond angles seen in the protein backbone which are geometrically unfavorable. If the backbone were to adopt these angles, the sidechains would be closer than is desired. These angles would take a large amount of energy to maintain and are therefore very unlikely (Nelson & Cox, 2012). While it is possible for the occasional outlier to be present in a protein, it is unlikely for many outliers to be present. Clash represents places where the structure has two atoms modeled in the same place at the same time, which is physically impossible. It should be highlighted that the NMR structure for ProQ has more Ramachandran outliers and clash even when compared to the other

NMR structure of the FinO domain, Lpp1663 (Immer et al., 2020). Furthermore, the AlphaFold structure for ProQ has fewer Ramachandran outliers and less clash. This indicates that the AlphaFold structure forms more favorable bond angles in the cell, and accounts adequately for the presence of all atoms.

IV-3. Mutation of proposed interactions in the AlphaFold structure

In an effort to experimentally validate the AlphaFold structure, we located interactions between amino acid side chains that were predicted to form in the AlphaFold structure, but not the NMR model. After examining both structures, we were able to identify two interactions between side chains seen only in the AlphaFold structure: a salt bridge between K35 and D41 and a cation-pi interaction between R80 and H95.

The hope was to find an interaction that could be both disrupted and restored with careful substitution. The ability to disrupt an interaction between sidechains with a site-specific mutation would support the interaction contributing to the structure and/or RNA binding of ProQ. The ability to restore the interaction with additional mutation would demonstrate knowledge of the chemical traits needed for the interaction. For example, if there was a loss of binding observed when either K35 or D41 was mutated to alanine, effectively removing all functional groups of the side chain, that would support the involvement of each side chain in an important interaction. If swapping the charges of the residues was able to lead to a greater level of binding than either alanine mutation, this would show that the charge of the residues were important. If single mutations to the opposite amino acid (such as K35D alone) did not restore binding, that would suggest the two residues need to be opposite charges. The functionality of the protein depended on both residues complementing each other in chemical characteristics rather than just the presence of either residue would support an interaction between K35 and D41.

It must be noted that our assay measures the binding of ProQ to RNA, therefore such an interaction would have to have a clear impact on either protein stability or the ability of ProQ to bind to RNA substrates for disruption and restoration of the interaction to be observed. If we were able to identify an interaction that could be disrupted and restored, this would support the presence of the interaction and therefore provide data in support of one model over the other.

IV-2-i. Modest indication of interaction seen in K35/D41 variants

The AlphaFold entry for ProQ predicted a salt bridge between K35 and D41. A mutant protein in which these amino acids were swapped did not show interaction with *malM,* SibB, or *cspE* in a liquid β-galactosidase assay (Figure 18). This indicates that these variants all were unable to bind to the RNA bait well enough to stimulate transcription of *lacZ* over the background level. However, this doesn't line up with the results seen in the plate data, which shows a modest increase in interaction between RNA and protein in the swap variant patch when compared to the single K35D variant. A high level of β-galactosidase activity in the CI empty control was observed for the double mutant in liquid, which would lead to a lower calculated fold interaction. In our assay, controls are used to demonstrate that increase in β-galactosidase activity is due to increased interaction between the RNA and protein and not other factors in the cell. Each component is transformed into the cell and has been fused to other macromolecules, such as the alpha subunit. Negative controls have the power to highlight increases in β-galactosidase activity not due to the specific RNA and protein. For example, ProQ was found to interact with a previously used negative RNA control which had two hairpins rather than one (Pandey et al., 2020; Stockert et al., 2022). This finding showed that interactions between ProQ and RNA detected in the assay may have been due to the constant portion of the hybrid RNA rather than the unique traits of specific sRNA or mRNA targets. While this can provide some

information about how ProQ interacts with RNA broadly, it does not provide information on specific interactions with target RNAs in the cell and should not be interpreted as such.

In the case of a high CI empty control, however, it is unlikely that the higher level of β-galactosidase activity reflects the presence of a false interaction in the assay. In this particular control, the "adapter" component of the assay (Figure 9) is replaced by an empty construct. In order for there to be a stimulation of β-galactosidase activity in this condition, the protein would have to be doing a two-hybrid interaction with the MS2 coat protein which typically holds the bait construct. This is unlikely to happen, as it would necessitate the MS2 coat protein binding the DNA sequence in order to stabilize RNAP on the promoter. Therefore, the high β-galactosidase activity observed in the CI empty control is probably due to a cause outside of the bait RNA and prey protein constructs, such as random binding of RNAP to the promoter region for *lacZ*. Unfortunately, this high negative control means that there was a lower calculated interaction for the K35D+D41K variant even as the absolute β-galactosidase activity was higher than for the K35D condition. As previously mentioned, increased β-galactosidase activity indicates increased interaction between the RNA and protein. The method of calculating fold interaction is used to allow for even comparison between experimental conditions, but it is not without limitations.

There are additional reasons why the patterns seen in the liquid and plate-based assays may not match. Differences between these two assays have been observed in our lab previously, such as plate-based screens identifying variant ProQ proteins that maintain interaction with RNA, only for such variants to fail to bind to RNA in the liquid β-galactosidase assay (this work, unpublished data collected by Linda Wang '23). However, the reasons behind these differences are not fully understood. Even though both assays utilize the B3H machinery, the two assays are

slightly different. For one, the bacteria are in different phases of growth: stationary phase on plates and log phase in liquid. These two phases of growth differ in physiology and gene expression (Caglar et al., 2017; Ishihama, 1997; Jaishankar & Srivastava, 2017). This could impact the results seen in our assay. Additionally, the plate data is observed as an absolute β-galactosidase activity while liquid data typically shows relative β-galactosidase activity in the form of fold interaction to allow for a more appropriate comparison between conditions. There have been attempts in the lab to develop an image analysis pipeline in order to quantify differences between blue and white observed on plates, which may allow for a calculation of something more similar to fold interaction. However, as we have yet to find an entirely satisfactory method for such calculations, there is still a gap in our ability to accurately compare blue and white colors. It is possible that while the patches for an experimental condition appear to be more blue than the patches for another experimental condition, the control patches are also different shades. This would be an example of a situation in which the difference in blue observed in experimental conditions is a result of a difference in background levels of *lacZ* transcription rather than increased interaction between protein and RNA.

While the assays performed with K35/D41 variant proteins did not generate data in agreement with each other, there is some indication of a salt bridge between the two residues in ProQ. K35 was previously identified as important to ProQ binding of RNA (Pandey et al., 2020). There were similar levels of interaction seen for K35A and D41A variants, which both effectively removed the respective side chains. This supports the possibility of these residues serving similar roles in the protein, such as involvement in this salt bridge. There is some indication of an increase in interaction between ProQ and RNA when the residues are swapped, which could indicate at least partial restoration of this interaction. Unfortunately, when used with

ProQ, this assay does not have refined enough sensitivity to detect subtle changes in binding. Additionally, our current methods of calculation do not allow for the interpretation of fold interaction below one. If either of these challenges are overcome in the future, it may become easier to interpret data in experiments such as this one.

IV-2-ii. H95 is not important to ProQ binding of RNA

As previously mentioned, in order for disruptions of interactions to be seen in this assay, the interaction must have an impact on RNA binding by ProQ. Mutations to H95 did not have distinct impacts on RNA binding by ProQ in our assay, even when histidine was mutated to alanine, effectively removing all chemical properties of the side chain. If this residue was important to RNA binding by ProQ, we would have expected to see a more distinct loss of binding when the H95A mutation was made. There was a more substantial loss of binding observed for the H95F variant protein, both in liquid and on plates. This is likely because the phenylalanine residue is large and bulky, which could lead to more disruption in the folding of the protein than smaller residues. These findings ultimately support the hypothesis that mutations to R80 have dramatic impacts on ProQ binding of RNA due to direct interactions between R80 and the RNA target, rather than interactions between R80 and other residues.

IV-4. The AlphaFold structure for ProQ is more consistent with experimental data for ProQ and other FinO proteins

The AlphaFold structure for ProQ presents a compelling alternative structure of the protein. As previously highlighted, this structure is more in line with the currently available structures of FinO family proteins in the distribution of conservation and charge as well as the location of key residues in ProQ and orthologs. Most notably, the two key residues of Y70 and R80 are located next to each other in the center of a highly conserved pocket on the AlphaFold

structure, while on the NMR structure these residues are on entirely opposite faces of the protein (Figure 29). The location in the center of the potential binding pocket offers a better explanation for the importance of the two residues, by providing a clear model for how both of these residues may be simultaneously involved in RNA binding.

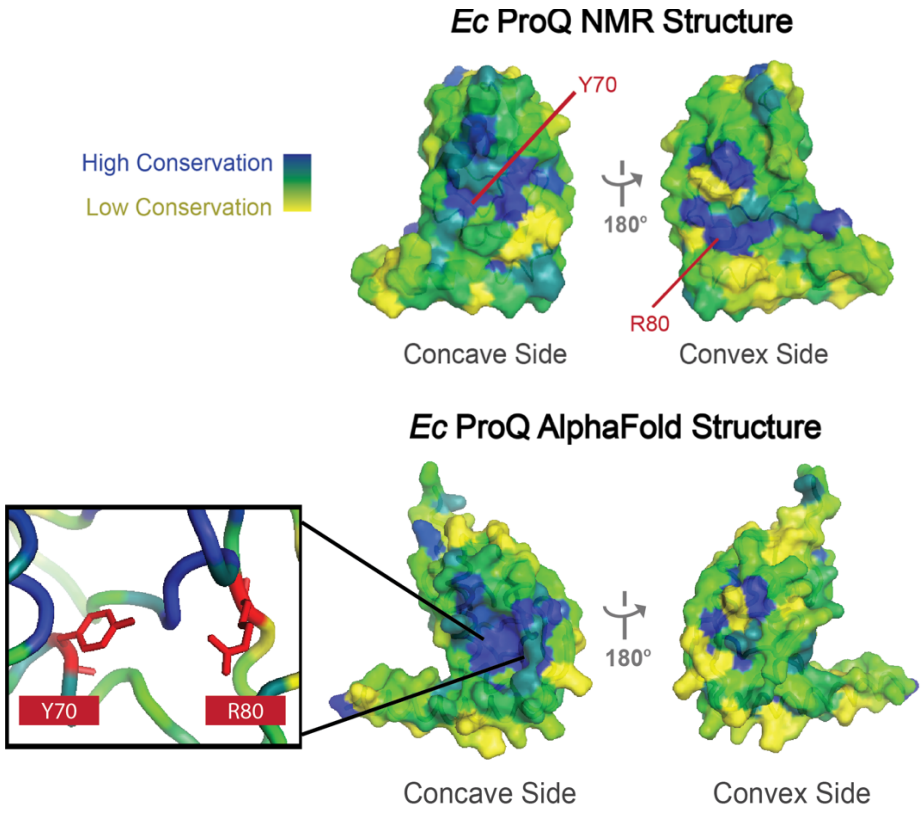


Figure 29: The location of key residues R80 and Y70 on the AlphaFold and NMR Structures for ProQ. On the NMR structure (top), the residues are on opposite faces of the protein. On the AlphaFold structure (bottom), the two residues are located in the center of a highly conserved pocket on the concave face of the protein. As in previous conservation figures, more highly conserved residues are colored in dark blue ranging to the least conserved residues are colored yellow. Conservation scores for ProQ were generated with the use of ConSurf (Ashkenazy et al., 2016).

It is possible that these residues are involved in binding the end of a double-stranded region of RNA, as the previously mentioned Argonaute protein does (Yuan et al., 2006). Studies on the RNA targets of ProQ have highlighted the presence of double-stranded RNA as a major determinant of ProQ-RNA interactions (Holmqvist et al., 2020; Stein et al., 2020). An RNA base pair, as is found at the end of a double-stranded region, fits into the conserved binding pocket of

ProQ (Figure 30). It is possible that these residues are functioning to encourage interactions between sRNAs and mRNAs regulated by ProQ.

The primary role of sRNA binding proteins is believed to be facilitating the sRNA-mediated regulation of mRNAs, which is possible when the sRNA imperfectly base pairs with the mRNA. FinO has been shown to encourage duplex formation between an sRNA and mRNA, in part by encouraging the RNA targets to not take on their native fold but instead fold in a way that promotes imperfect base pairings between the two molecules (Arthur et al., 2003). It is possible that Y70 and R80, as part of the potential binding pocket, encourage the RNA molecule to shift from its highly structured shape into something which better allows base pairing between the sRNA and mRNA. Pi-pi stacking interactions are energetically favorable, at 2-6 kcal/mol (Corley et al., 2020), while cation-pi interactions provide 2-5 kcal/mol (Dougherty, 2013). Comparatively, hydrogen bonds are only 1-2 kcal/mol each. It is possible that the interactions between the RNA bases and the side chains on the protein are more favorable than the hydrogen bonds holding the bases together, helping to encourage the unfolding of the double-stranded region so that it may better pair with the other RNA.

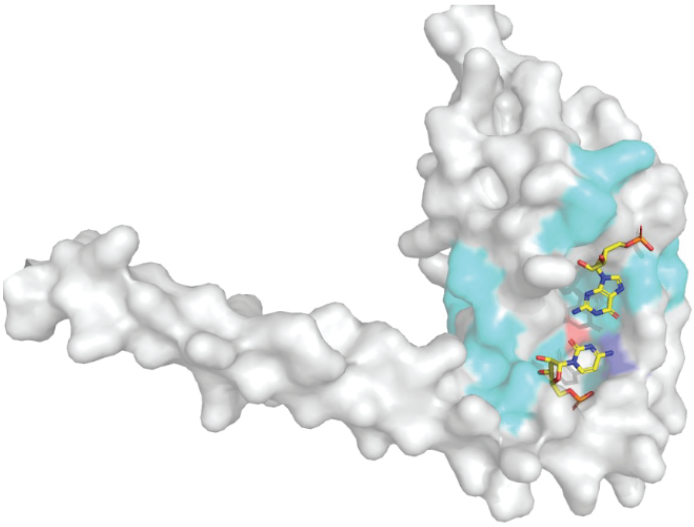


Figure 30: An RNA base pair placed in the concave potential binding pocket of the AlphaFold structure for ProQ for context and relative size. The base pair fits into the pocket of the protein. Figure created by Professor Katie Berry.

Furthermore, it is important to note that the AlphaFold structure for ProQ performs better than the NMR structure on measures of basic validation, namely examination of Ramachandran outliers and steric clash. These validation measures are intended to determine the favorability of the protein fold in cells. Better performance on the validation measures suggests that the primary sequence of ProQ is more likely to fold into the AlphaFold structure in cells, as this structure is more energetically and physically favorable. This could be because the AlphaFold algorithm has a strong awareness of the biophysically limited protein folding space, developed in the training on other structures. It is unlikely that the software used by Gonzalez et al. to model the NMR data for ProQ is able to account for general protein folding patterns to the same degree.

IV-5. Applications

The analysis of current available ProQ structures presented here hopes to help ensure that future research on ProQ, most notably the development of a model for RNA binding, is based on the most accurate available structure for the protein. This has been done by both highlighting weaknesses in the current structure and offering an alternate structure for ProQ. These

weaknesses include a lack of clear explanation for key residues Y70 and R80, which are directly next to each other in the alternate model and orthologous structures. Due to an extreme sensitivity to mutation, these residues very likely contribute directly due to RNA binding. Without the correct structure for ProQ, interactions between the protein and RNA substrates will not be accurately determined. The list of known FinO family proteins is growing, and it is possible that developing a model for RNA binding by ProQ could assist in the understanding of more general FinO domain binding of RNA. The broader utility of this study to the field is underscored by the role of research on other FinO structures in both this work and the work presenting the newest FinO-domain protein structure (Immer et al., 2020).

This work also highlights a value to be found in computational methods of protein prediction. While these methods cannot, and should not, replace empirical benchwork, these structures can provide a source of hypotheses. In some cases, such as that of ProQ, the available structure may fall short in some areas. Computationally predicted structures can be used as an alternate framework to understand confusing results. We believe it is possible to use these newly available computational structures to generate benchwork experiments that may support or refute the accuracy of the computational structures.

IV-6. Limitations

While this work has the potential to be valuable to the field, there are limitations to multiple techniques used in the work which must be considered. The B3H itself, on which a lot of this work is based, is relatively new technology and is still undergoing optimization (Berry & Hochschild, 2018; Wang et al., 2021). Currently, there is a low level of signal observed by our lab for some RNAs which are known to interact with ProQ. This makes the detection of subtle differences in binding as a result of mutations more challenging to discern. Additionally, with the

current method of calculating fold interaction, the meaning of differences in results that are below one fold is unclear. For RNA targets with an already low signal, this makes determining differences in levels of interaction all the more difficult. As highlighted by the discussion of the K35/D41 salt bridge in this work, this can make it challenging to see the true effects of mutations. Further optimization of the assay for use with ProQ may allow for better detection of interactions, and therefore more nuanced discussion of the effect of mutations to either the RNA or protein.

There are additional challenges specific to the use of an *in vivo* system. While this system allows more accurate replication of the conditions found in cells, that strength is also a weakness as many factors in the cell are not within our control. The presence of degradation machinery in particular is a challenge for this assay. It is possible for RNA and protein mutants to fold incorrectly and therefore be degraded by the intracellular machinery. This can lead to low levels of interaction, not due to lower interaction between RNA and protein but due to less protein or RNA present in the cell than in the wild type condition. Such results can mislead researchers to attempt to find structural reasons for these effects caused by degradation. This challenge can be partially remedied with the use of Northern and Western blots to examine the stability of RNA and protein in the cell, respectively. However, there is still an unknown element of folding in the cell. The RNA has the potential to fold incorrectly when fused to the MS2 moiety, which could lead to an interaction in the assay not observed when the RNA adopts its native fold in cells. Mutations in prey proteins lead to unknown consequences, as computational models of mutation can only offer a prediction. Even with the emergence of groundbreaking discoveries such as AlphaFold, we don't have a comprehensive understanding of all of the factors involved in protein folding. While Western blots can allow us to measure the stability of proteins in the cell,

we cannot be sure how specific portions of the protein are folding. It is likely that mutations (especially those that are not conservative) lead to protein misfolding and a greater alteration to the structure than intended.

In addition to the use of the B3H, this work depended on the examination of the structure of other FinO-domain proteins. While these may provide a helpful point of comparison for a protein with a structure in question, it must be remembered that these are ultimately different proteins with slightly different functions. There are differences in the fold of each structure, which can be seen even in the differences between the three other available structures (PDB ID: 1DVO, PDB ID: 3MW6, PDB ID: 6S10; Chaulk et al., 2010; Ghetu et al., 2000; Immer et al., 2020). There are differences between the other FinO-domain proteins and ProQ. This work focuses on general trends gleaned from multiple proteins, but it must be recognized that the specifics will vary from structure to structure. Additionally, none of the proteins mentioned in this paper have a crystal costructure with an RNA target. This would provide a level of knowledge on the molecular mechanism of interaction that is currently not available. This gap means that there is ultimately less information on FinO domain binding of RNA which can be applied to ProQ. There are varying levels of information on the different orthologs, from models of binding informed by crosslinking and gelFRET experiments (Ghetu et al., 2002) to no identified RNA targets (Immer et al., 2020). As a result, in this work, the predictions of RNA binding regions in the proteins were often based on chemical traits rather than experimental data in order to best utilize all available structures.

There are also limitations to the AlphaFold structure prediction software. The structures created by AlphaFold are not equivalent to experimentally determined structures and should be considered more as a source of hypothesis generation rather than fact. The algorithm does not

factor in cellular conditions, and cannot predict happenings in the cell such as post-translational modifications of protein or cofactors (Bagdonas et al., 2021). As this technology is still relatively new, the use of AlphaFold structures in the study of proteins requires experimental data supporting the AlphaFold structure for that specific protein.

IV-7. Future directions

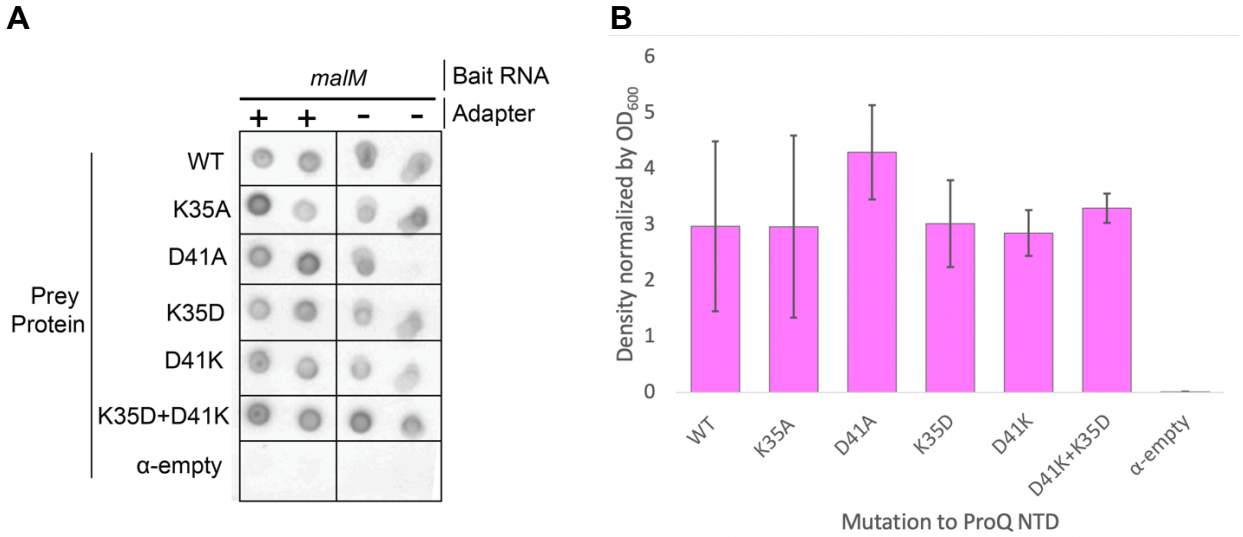
An ideal next step for this area of study would be to determine a crystal costructure of ProQ with RNA. This would show the molecular mechanism for how ProQ binds to its RNA targets in detail. Short of a costructure, a solved crystal structure for this particular protein would still be valuable. Given the conflict between the currently available structure and the experimental data available for this protein, as presented in this work, there would be value in an additional experimentally determined structure for this protein. Crystal structures bypass the need for developing a model from limited data, unlike NMR structures which are based on a model which fits the provided constraints (Vranken, 2014). While it is true that it can be challenging to get crystals of some proteins (Carpenter et al., 2008; Grey & Thompson, 2010), the existence of crystal structures for both NMB1681 and FinO (Chaulk et al., 2010; Ghetu et al., 2000) indicate that it should be possible to create a crystal of at least the FinO domain of this protein.

Even without an additional structure, it is possible that the NMR constraints could be modeled in an alternate way to produce a more accurate structure. Analysis of the constraints done by our lab showed that some atoms appeared to be out of range of the constraints in the final structure of the protein (Amy Wang ’22, unpublished data). It would be interesting to examine how these constraints fall on the AlphaFold structure. It is possible that the AlphaFold

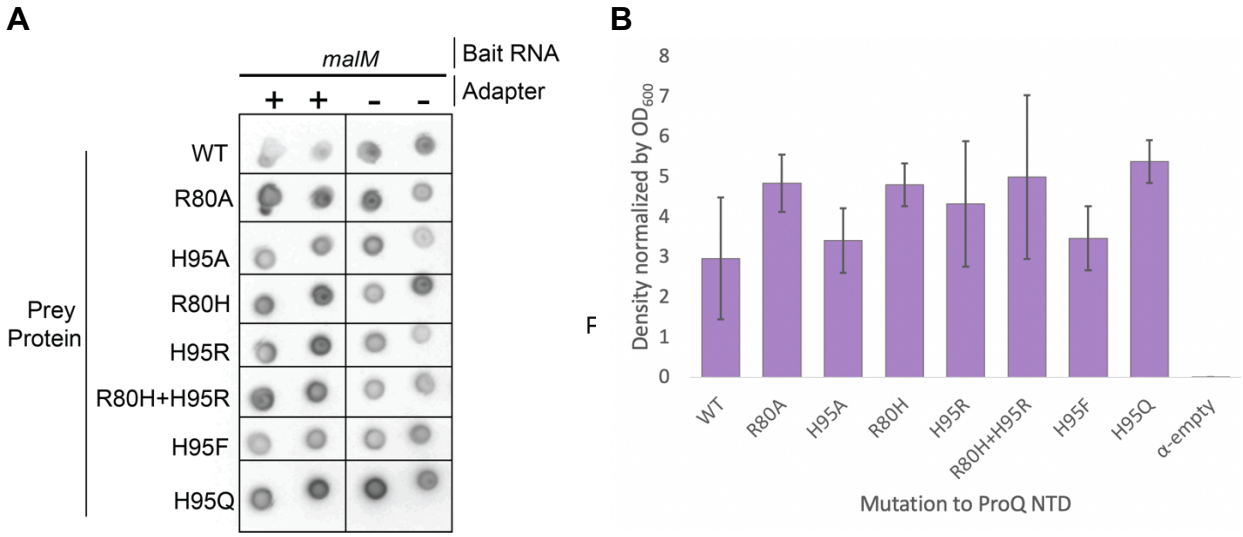
structure is ultimately in line with the NMR data, and the error in structure development occurred in the software used to develop a structure from the constraints.

Furthermore, we can continue our ultimate goal of developing an accurate model for RNA binding by ProQ. One way this can be done is through site-specific RNA footprinting or crosslinking experiments. I do hope that such experiments consider the AlphaFold structure for ProQ in attempts to comprehend the results. With the consideration of this model, the potential RNA binding pocket may be probed through additional mutagenesis experiments. Additionally, scientists with the requisite software experience may try to dock a known RNA target of ProQ on the AlphaFold structure. This can be done with consideration for currently available data on ProQ binding of RNA, such as the residues critical to binding previously identified by our lab (Pandey et al., 2020).

Appendix



Supplementary Figure 1: Immunoblot data for K35 and D41 variants. Panel **A)** depicts the membrane that the cell lysates for cells containing all components of the assay and cells missing an adapter component were spotted onto, while panel **B)** shows the data resulting from densitometry analysis of the spots. The lysates were taken from cells used in a liquid B3H experiment. These cells were grown with 0 IPTG overnight and 0 IPTG during the day. None of the variants were notably destabilized. The absence of a second spot at D41A variant without an adapter is very likely the result of a pipetting issue, since both spots appeared on the membrane run with lysates grown with 0 IPTG overnight and 50 IPTG during the day. This experiment was run only once.



Supplementary Figure 2: Immunoblot data for H95 and R80 variants. Panel A) depicts the membrane that the cell lysates for cells containing all components of the assay and cells missing an adapter component were spotted onto, imaged with a Azure c600 chemiluminescent imaging system, while panel B) shows the data resulting from densitometry analysis of the spots. The lysates were taken from cells used in a liquid B3H. These cells were grown with 0 IPTG overnight and 0 IPTG during the day. None of the variants were notably destabilized. This experiment was run only once.

Supplementary Table 1: Distances between residues in the NMR structure of ProQ for interactions predicted by AlphaFold. These distances were measured between the atoms at the very ends of the side chains. The average distance between the end of lysine and aspartate in the K35/D41 salt bridge was 9.23 Å, with a minimum distance of

2.6 Å and a maximum distance of 13.4 Å. The equivalent distance in the AlphaFold structure is 2.8 Å. The average distance between the end of the arginine and the ring on histidine in the H95/R80 cation-pi interaction was 10.02 Å, with a minimum distance of 9.5 Å and a maximum distance of 10.9 Å. The equivalent distance in the AlphaFold structure is 3.9 Å.

	Distances Between Residues (Å)	
Structure from NMR File	K35/D41 Salt Bridge	H95/R80 Cation-Pi Interaction
1	12.8	9.5
2	12.1	9.7
3	13.4	9.6
4	11.9	10.6
5	13.1	10.6
6	13.4	9.6
7	6.7	9.5
8	4.7	9.8
9	11.7	9.6
10	4.7	10.9
11	13.4	10.9
12	4.7	9.6
13	6.4	10
14	13.1	9.6
15	4.8	10.9
16	7.4	9.8
17	2.6	10.2

References

Anfinsen, C. B., Haber, E., Sela, M., & White, F. H. (1961). The Kinetics of Formation of Native Ribonuclease During Oxidation of the Reduced Polypeptide Chain. *Proceedings of the National Academy of Sciences*, *47*(9), 1309–1314. https://doi.org/10.1073/pnas.47.9.1309

Arthur, D. C., Ghetu, A. F., Gubbins, M. J., Edwards, R. A., Frost, L. S., & Glover, J. N. M. (2003). FinO is an RNA chaperone that facilitates sense-antisense RNA interactions. *The EMBO Journal*, *22*(23), 6346–6355. https://doi.org/10.1093/emboj/cdg607

Ashkenazy, H., Abadi, S., Martz, E., Chay, O., Mayrose, I., Pupko, T., & Ben-Tal, N. (2016). ConSurf 2016: An improved methodology to estimate and visualize evolutionary conservation in macromolecules. *Nucleic Acids Research*, *44*(W1), W344–W350. https://doi.org/10.1093/nar/gkw408

Bagdonas, H., Fogarty, C. A., Fadda, E., & Agirre, J. (2021). The case for post-predictional modifications in the AlphaFold Protein Structure Database. *Nature Structural & Molecular Biology*, *28*(11), 869–870. https://doi.org/10.1038/s41594-021-00680-9

Berry, K. E., & Hochschild, A. (2018). A bacterial three-hybrid assay detects Escherichia coli Hfq–sRNA interactions in vivo. *Nucleic Acids Research*, *46*(2), e12–e12. https://doi.org/10.1093/nar/gkx1086

Bouatta, N., Sorger, P., & AlQuraishi, M. (2021). Protein structure prediction by AlphaFold2: Are attention and symmetries all you need? *Acta Crystallographica. Section D, Structural Biology*, *77*(Pt 8), 982–991. https://doi.org/10.1107/S2059798321007531

Caglar, M. U., Houser, J. R., Barnhart, C. S., Boutz, D. R., Carroll, S. M., Dasgupta, A., Lenoir, W. F., Smith, B. L., Sridhara, V., Sydykova, D. K., Vander Wood, D., Marx, C. J.,

Marcotte, E. M., Barrick, J. E., & Wilke, C. O. (2017). The E. coli molecular phenotype under different growth conditions. *Scientific Reports*, *7*(1), 45303. https://doi.org/10.1038/srep45303

Callaway, E. (2020). 'It will change everything': DeepMind's AI makes gigantic leap in solving protein structures. *Nature*, *588*(7837), 203–204. https://doi.org/10.1038/d41586-020-03348-4

Carpenter, E. P., Beis, K., Cameron, A. D., & Iwata, S. (2008). Overcoming the challenges of membrane protein crystallography. *Current Opinion in Structural Biology*, *18*(5), 581–586. https://doi.org/10.1016/j.sbi.2008.07.001

Chaulk, S. G., Lu, J., Tan, K., Arthur, D. C., Edwards, R. A., Frost, L. S., Joachimiak, A., & Glover, J. N. M. (2010). N. meningitidis 1681 is a member of the FinO family of RNA chaperones. *RNA Biology*, *7*(6), 812–819. https://doi.org/10.4161/rna.7.6.13688

Chen, V. B., Arendall, W. B., Headd, J. J., Keedy, D. A., Immormino, R. M., Kapral, G. J., Murray, L. W., Richardson, J. S., & Richardson, D. C. (2010). *MolProbity*: All-atom structure validation for macromolecular crystallography. *Acta Crystallographica Section D Biological Crystallography*, *66*(1), 12–21. https://doi.org/10.1107/S0907444909042073

Cho, K., van Merrienboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., & Bengio, Y. (2014). Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation. *ArXiv:1406.1078 [Cs, Stat]*. http://arxiv.org/abs/1406.1078

Corley, M., Burns, M. C., & Yeo, G. W. (2020). How RNA-Binding Proteins Interact with RNA: Molecules and Mechanisms. *Molecular Cell*, *78*(1), 9–29. https://doi.org/10.1016/j.molcel.2020.03.011

Dougherty, D. A. (2013). The Cation−π Interaction. *Accounts of Chemical Research*, *46*(4), 885–893. https://doi.org/10.1021/ar300265y

Dove, S. L., Joung, J. K., & Hochschild, A. (1997). Activation of prokaryotic transcription through arbitrary protein-protein contacts. *Nature*, *386*(6625), 627–630. https://doi.org/10.1038/386627a0

El Mouali, Y., Ponath, F., Scharrer, V., Wenner, N., Hinton, J. C. D., & Vogel, J. (2021). Scanning mutagenesis of RNA-binding protein ProQ reveals a quality control role for the Lon protease. *RNA*, *27*(12), 1512–1527. https://doi.org/10.1261/rna.078954.121

Emsley, P., Lohkamp, B., Scott, W. G., & Cowtan, K. (2010). Features and development of *Coot*. *Acta Crystallographica Section D Biological Crystallography*, *66*(4), 486–501. https://doi.org/10.1107/S0907444910007493

Felden, B., & Augagneur, Y. (2021). Diversity and Versatility in Small RNA-Mediated Regulation in Bacterial Pathogens. *Frontiers in Microbiology*, *12*, 719977. https://doi.org/10.3389/fmicb.2021.719977

Flower, T. G., & Hurley, J. H. (2021). Crystallographic molecular replacement using an in silico-generated search model of SARS-CoV-2 ORF8. *Protein Science : A Publication of the Protein Society*, *30*(4), 728–734. https://doi.org/10.1002/pro.4050

Full wwPDB NMR Structure Validation Report for PDB ID 5nb9. (2020). Worldwide Protein Data Bank. https://files.rcsb.org/pub/pdb/validation_reports/nb/5nb9/5nb9_full_validation.pdf

G. Chaulk, S., Smith−Frieday, M. N., Arthur, D. C., Culham, D. E., Edwards, R. A., Soo, P., Frost, L. S., Keates, R. A. B., Glover, J. N. M., & Wood, J. M. (2011). ProQ Is an RNA Chaperone that Controls ProP Levels in *Escherichia coli*. *Biochemistry*, *50*(15), 3095–3106. https://doi.org/10.1021/bi101683a

Ghetu, A. F., Arthur, D. C., Kerppola, T. K., & Glover, J. N. M. (2002). Probing FinO-FinP RNA interactions by site-directed protein-RNA crosslinking and gelFRET. *RNA (New York, N.Y.)*, *8*(6), 816–823. https://doi.org/10.1017/s1355838202026730

Ghetu, A. F., Gubbins, M. J., Frost, L. S., & Glover, J. N. M. (2000). Crystal structure of the bacterial conjugation repressor FinO. *Nature Structural Biology*, *7*(7), 565–569. https://doi.org/10.1038/76790

Glover, M. J. N., Chaulk, S. G., Edwards, R. A., Arthur, D., Lu, J., & Frost, L. S. (2015). The FinO family of bacterial RNA chaperones. *Plasmid*, *78*, 79–87. https://doi.org/10.1016/j.plasmid.2014.07.003

Gonzalez, G. M., Hardwick, S. W., Maslen, S. L., Skehel, J. M., Holmqvist, E., Vogel, J., Bateman, A., Luisi, B. F., & Broadhurst, R. W. (2017). Structure of the *Escherichia coli* ProQ RNA-binding protein. *RNA*, *23*(5), 696–711. https://doi.org/10.1261/rna.060343.116

Grey, J., & Thompson, D. (2010). Challenges and Opportunities for New Protein Crystallization Strategies in Structure-Based Drug Design. *Expert Opinion on Drug Discovery*, *5*(11), 1039–1045. https://doi.org/10.1517/17460441.2010.515583

Han, R., Haning, K., Gonzalez-Rivera, J. C., Yang, Y., Li, R., Cho, S. H., Huang, J., Simonsen, B. A., Yang, S., & Contreras, L. M. (2020). Multiple Small RNAs Interact to Co-regulate

Ethanol Tolerance in Zymomonas mobilis. *Frontiers in Bioengineering and Biotechnology*, *8*. https://www.frontiersin.org/article/10.3389/fbioe.2020.00155

Hohman, F., Kahng, M., Pienta, R., & Chau, D. H. (2018). Visual Analytics in Deep Learning: An Interrogative Survey for the Next Frontiers. *ArXiv:1801.06889 [Cs, Stat]*. http://arxiv.org/abs/1801.06889

Holmqvist, E., Berggren, S., & Rizvanovic, A. (2020). RNA-binding activity and regulatory functions of the emerging sRNA-binding protein ProQ. *Biochimica Et Biophysica Acta. Gene Regulatory Mechanisms*, *1863*(9), 194596. https://doi.org/10.1016/j.bbagrm.2020.194596

Immer, C., Hacker, C., & Wöhnert, J. (2020). Solution structure and RNA-binding of a minimal ProQ-homolog from *Legionella pneumophila* (Lpp1663). *RNA*, *26*(12), 2031–2043. https://doi.org/10.1261/rna.077354.120

Ishihama, A. (1997). Adaptation of gene expression in stationary phase bacteria. *Current Opinion in Genetics & Development*, *7*(5), 582–588. https://doi.org/10.1016/S0959-437X(97)80003-2

Jaishankar, J., & Srivastava, P. (2017). Molecular Basis of Stationary Phase Survival and Applications. *Frontiers in Microbiology*, *8*, 2000. https://doi.org/10.3389/fmicb.2017.02000

Jerome, L. J., van Biesen, T., & Frost, L. S. (1999). Degradation of FinP antisense RNA from F-like plasmids: The RNA-binding protein, FinO, protects FinP from ribonuclease E. *Journal of Molecular Biology*, *285*(4), 1457–1473. https://doi.org/10.1006/jmbi.1998.2404

Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A., Bridgland, A., Meyer, C., Kohl, S. A. A., Ballard, A. J., Cowie, A., Romera-Paredes, B., Nikolov, S., Jain, R., Adler, J., … Hassabis, D. (2021). Highly accurate protein structure prediction with AlphaFold. *Nature*, *596*(7873), 583–589. https://doi.org/10.1038/s41586-021-03819-2

Kovner, A. (2021, January 12). *The Odd Structure of ORF8: Scientists Map the Coronavirus Protein Linked to Immune Evasion and Disease Severity*. News Center. https://newscenter.lbl.gov/2021/01/12/odd-structure-of-orf8/

Kriegeskorte, N. (2015). Deep Neural Networks: A New Framework for Modeling Biological Vision and Brain Information Processing. *Annual Review of Vision Science*, *1*(1), 417–446. https://doi.org/10.1146/annurev-vision-082114-035447

Kufareva, I., & Abagyan, R. (2012). Methods of protein structure comparison. *Methods in Molecular Biology (Clifton, N.J.)*, *857*, 231–257. https://doi.org/10.1007/978-1-61779-588-6_10

Leonard, S., Villard, C., Nasser, W., Reverchon, S., & Hommais, F. (2021). RNA Chaperones Hfq and ProQ Play a Key Role in the Virulence of the Plant Pathogenic Bacterium Dickeya dadantii. *Frontiers in Microbiology*, *12*, 687484. https://doi.org/10.3389/fmicb.2021.687484

Levinthal, C. (1969). Mossbauer spectroscopy in biological systems: Proceedings of a meeting held at allerton house. *University of Illinois*, 22–24.

Lisitskaya, L., Aravin, A. A., & Kulbachinskiy, A. (2018). DNA interference and beyond: Structure and functions of prokaryotic Argonaute proteins. *Nature Communications*, *9*(1), 5165. https://doi.org/10.1038/s41467-018-07449-7

Mai, J., Rao, C., Watt, J., Sun, X., Lin, C., Zhang, L., & Liu, J. (2019). Mycobacterium tuberculosis 6C sRNA binds multiple mRNA targets via C-rich loops independent of RNA chaperones. *Nucleic Acids Research*, *47*(8), 4292–4307. https://doi.org/10.1093/nar/gkz149

Melamed, S., Adams, P. P., Zhang, A., Zhang, H., & Storz, G. (2020). RNA-RNA Interactomes of ProQ and Hfq Reveal Overlapping and Competing Roles. *Molecular Cell*, *77*(2), 411-425.e7. https://doi.org/10.1016/j.molcel.2019.10.022

Milner, J. L., & Wood, J. M. (1989). Insertion proQ220::Tn5 alters regulation of proline porter II, a transporter of proline and glycine betaine in Escherichia coli. *Journal of Bacteriology*, *171*(2), 947–951. https://doi.org/10.1128/jb.171.2.947-951.1989

Nelson, D. L., & Cox, M. M. (2012). *Lehninger Principles of Biochemistry* (6th edition). W.H. Freeman.

Olejniczak, M., & Storz, G. (2017). ProQ/FinO-domain proteins: Another ubiquitous family of RNA matchmakers? *Molecular Microbiology*, *104*(6), 905–915. https://doi.org/10.1111/mmi.13679

Padmanabhan, S., Banerjee, S., & Mandi, N. (2011). Screening of Bacterial Recombinants: Strategies and Preventing False Positives. In *Molecular Cloning—Selected Applications in Medicine and Biology*. IntechOpen. https://doi.org/10.5772/22140

Pandey, S., Gravel, C. M., Stockert, O. M., Wang, C. D., Hegner, C. L., LeBlanc, H., & Berry, K. E. (2020). Genetic identification of the functional surface for RNA binding by Escherichia coli ProQ. *Nucleic Acids Research*, *48*(8), 4507–4520. https://doi.org/10.1093/nar/gkaa144

Pearson, W. R. (2013). An introduction to sequence similarity ("homology") searching. *Current Protocols in Bioinformatics*, *Chapter 3*, Unit3.1. https://doi.org/10.1002/0471250953.bi0301s42

Perrakis, A., & Sixma, T. K. (2021). AI revolutions in biology. *EMBO Reports*, *22*(11), e54046. https://doi.org/10.15252/embr.202154046

Rizvanovic, A., Kjellin, J., Söderbom, F., & Holmqvist, E. (2021). Saturation mutagenesis charts the functional landscape of *Salmonella* ProQ and reveals a gene regulatory function of its C-terminal domain. *Nucleic Acids Research*, *49*(17), 9992–10006. https://doi.org/10.1093/nar/gkab721

Rost, B. (1999). Twilight zone of protein sequence alignments. *Protein Engineering*, *12*(2), 85–94. https://doi.org/10.1093/protein/12.2.85

Sandercock, J. R., & Frost, L. S. (1998). Analysis of the major domains of the F fertility inhibition protein, FinO. *Molecular & General Genetics: MGG*, *259*(6), 622–629. https://doi.org/10.1007/s004380050856

Schrödinger, LLC. (2015). *The PyMOL Molecular Graphics System, Version 1.8*.

SenGupta, D. J., Zhang, B., Kraemer, B., Pochart, P., Fields, S., & Wickens, M. (1996). A three-hybrid system to detect RNA-protein interactions in vivo. *Proceedings of the National Academy of Sciences of the United States of America*, *93*(16), 8496–8501. https://doi.org/10.1073/pnas.93.16.8496

Show charged—PyMOLWiki. (2009, April 30). https://pymolwiki.org/index.php/Show_charged

Smith, M. N., Crane, R. A., Keates, R. A. B., & Wood, J. M. (2004). Overexpression, purification, and characterization of ProQ, a posttranslational regulator for

osmoregulatory transporter ProP of Escherichia coli. *Biochemistry*, *43*(41), 12979–12989. https://doi.org/10.1021/bi048561g

Smith, M. N., Kwok, S. C., Hodges, R. S., & Wood, J. M. (2007). Structural and functional analysis of ProQ: An osmoregulatory protein of Escherichia coli. *Biochemistry*, *46*(11), 3084–3095. https://doi.org/10.1021/bi6023786

Stalmach, M. E., Grothe, S., & Wood, J. M. (1983). Two proline porters in Escherichia coli K-12. *Journal of Bacteriology*, *156*(2), 481–486. https://doi.org/10.1128/jb.156.2.481-486.1983

Stein, E. M., Kwiatkowska, J., Basczok, M. M., Gravel, C. M., Berry, K. E., & Olejniczak, M. (2020). Determinants of RNA recognition by the FinO domain of the Escherichia coli ProQ protein. *Nucleic Acids Research*, gkaa497. https://doi.org/10.1093/nar/gkaa497

Stockert, O. M. (2021). *Elucidation of the mechanisms of RNA binding by the Escherichia coli ProQ N-terminal domain: A molecular genetic approach* [Thesis]. https://ida.mtholyoke.edu/handle/10166/6303

Stockert, O. M., Gravel, C. M., & Berry, K. E. (2022). A bacterial three-hybrid assay for forward and reverse genetic analysis of RNA–protein interactions. *Nature Protocols*, *17*(4), 941–961. https://doi.org/10.1038/s41596-021-00657-4

Thibodeau, S. A., Fang, R., & Joung, J. K. (2004). High-throughput beta-galactosidase assay for bacterial cell-based reporter systems. *BioTechniques*, *36*(3), 410–415. https://doi.org/10.2144/04363BM07

Updegrove, T. B., Zhang, A., & Storz, G. (2016). Hfq: The flexible RNA matchmaker. *Current Opinion in Microbiology*, *30*, 133–138. https://doi.org/10.1016/j.mib.2016.02.003

Vranken, W. F. (2014). NMR structure validation in relation to dynamics and structure determination. *Progress in Nuclear Magnetic Resonance Spectroscopy*, *82*, 27–38. https://doi.org/10.1016/j.pnmrs.2014.08.001

Wagner, E. G. H., & Romby, P. (2015). Small RNAs in Bacteria and Archaea: Who They Are, What They Do, and How They Do It. In *Advances in Genetics* (Vol. 90, pp. 133–208). Elsevier. https://doi.org/10.1016/bs.adgen.2015.05.001

Wang, C. D., Mansky, R., LeBlanc, H., Gravel, C. M., & Berry, K. E. (2021). Optimization of a bacterial three-hybrid assay through in vivo titration of an RNA–DNA adapter protein. *RNA*, *27*(4), 513–526. https://doi.org/10.1261/rna.077404.120

Xu, K., Ba, J., Kiros, R., Cho, K., Courville, A., Salakhudinov, R., Zemel, R., & Bengio, Y. (2015). Show, Attend and Tell: Neural Image Caption Generation with Visual Attention. *Proceedings of the 32nd International Conference on Machine Learning*, 2048–2057. https://proceedings.mlr.press/v37/xuc15.html

Yuan, Y.-R., Pei, Y., Chen, H.-Y., Tuschl, T., & Patel, D. J. (2006). A Potential Protein-RNA Recognition Event along the RISC-Loading Pathway from the Structure of A. aeolicus Argonaute with Externally Bound siRNA. *Structure (London, England : 1993)*, *14*(10), 1557–1565. https://doi.org/10.1016/j.str.2006.08.009