APPLICATIONS OF UNNATURAL AMINO ACID MUTAGENESIS FOR BIOCATALYSIS AND MOLECULAR RECOGNITION

Stefanie A. Baril

A dissertation submitted to the faculty at the University of North Carolina at Chapel Hill in partial fulfillment of the requirement for the degree of Doctor of Philosophy in the Department of Chemistry (Biological) in the College of Arts and Sciences.

Chapel Hill
2017

Approved by:

Eric Brustad

Gary Pielak

Brian Kuhlman

Matt Redinbo

Marcey Waters

ABSTRACT

Stefanie A. Baril: Applications of Unnatural Amino Acid Mutagenesis for Biocatalysis and Molecular Recognition
(Under the direction of Eric Brustad)

Proteins are diverse biopolymers constructed from 20 canonical amino acids, affording innumerable amino acid combinations. This combinatorial diversity allows for a variety of essential cellular functions ranging from enzyme catalysis, molecular recognition, signal transduction, structure, and beyond. Although nature has successfully evolved proteins for numerous purposes, protein function is limited by the restricted chemical functionality of the canonical amino acids. For example, natural amino acids include only a handful of residues capable of participating directly in catalysis. In addition, amino acid mutations rarely allow for fine-tuned changes to the amino acid side chain. Although information may be discovered from natural amino acid mutations, it is often difficult to determine how protein function would change with more subtle modulations of the chemical properties of the amino acid. With the introduction of *in vivo* unnatural amino acid (UAA) mutagenesis, amino acid sidechains may be expanded into previously unavailable chemical functionalities. The purpose of this dissertation is to expand the applications of unnatural amino acid mutagenesis into two diverse fields: 1) biophysical investigations of molecular recognition motifs that exploit cation-π interactions, and 2) biocatalysis using cofactor-like UAAs. For our biophysical studies, we examined cation-π interactions between epigenetic reader proteins and their cationic ligands. UAAs with different electron-withdrawing or –donating

groups were incorporated into the binding site, thereby tuning the electronics of the cation-π interaction. This methodology was successfully applied to Heterochromatin Protein 1, and we have engineered a new expression system to next study CBX5 and CBX7, which are implicated in multiple cancers. For biocatalysis, we have made progress towards genetically encoding a thiamine-like unnatural amino acid. Thiamine, a potent N-heterocyclic carbene cofactor, is required for cell metabolism in all domains of life. By incorporating a thiamine-like UAA *in vivo*, we are one step closer to transforming any binding pocket into a powerful N-heterocyclic carbene-containing enzyme as well as conferring active site stereoselectivity to existing N-heterocyclic carbene small molecule catalysts. It is our hope that our approaches will help the field of unnatural amino acid mutagenesis expand the tools available to protein chemists studying the innumerable proteins in our world.

This dissertation is dedicated to the memory of William Tickey, Sr.

ACKNOWLEDGEMENTS

There are many people who have been invaluable during the course of my graduate school career. Smooth seas never made a skilled sailor, and I am so happy to have had all of your support during this learning experience.

First, I would like to thank my mentor Eric Brustad all his help over the years and for choosing me as the founding member of his lab. We have come such a long way from our first summer hauling Dewars up nine flights of stairs. Thank you for understanding, but never judging, my caffeine addiction.

I would like to thank my committee members: Dr. Matthew Redinbo, Dr. Marcey Waters, Dr. Brian Kuhlman, and Dr. Gary Pielak, as well as Dr. Bo Li and Dr. Jeff Johnson, who served on my orals committee. Thank you so much for the guidance, support, and aptly timed wisecracks over the years. This experience would not be the same without all of you. To Dr. Mike Miley, thank you for all your help with crystallography and giving me a place to set up trays undisturbed. Thank you to Dr. Ash Tripathy for all his help with so many of our ITC runs. And to Dr. Laurie Betts, thank you for all of your help over the years with crystallography and for all of your support.

I would also like to thank my collaborators for all their help. Thank you to Dr. Marcey Waters, for all of her support and patience and for always believing that I could produce absurdly large amounts of protein. I would have never discovered my knack for protein expression engineering if you never requested all that HP1. To Dr. Amber Koenig, thank you so much for trusting and supporting me on this project. You were my absolute

favorite person to work with, and not just because of the mandatory cupcake and coffee breaks after a failed ITC run. To Mack Krone, thank you for being your cheerful, inquisitive self when our readers misbehaved (so many times) and for being understanding when I was not so cheerful (which was often). Thank you to Katherine Albanese, for being the first one in and the last one out on our team. I have no idea how you do it (or why you do it…get some sleep!), but thank you for doing it. To Kelsey Lamb, thank you for your synthetic expertise, for always having inhibitor when we needed it, and for always being excited when I bring baked goods. And to Dr. Dale Wilger, for being such a great synthetic chemistry mentor and resource. Chapter 4 would not be possible without you!

I would like to thank the members of the Brustad Lab, both past and present, in order of appearance: Josh, Evan, Adrienne, Richard, Amy, Tim, Mack, and Nolan and all the undergrads. It has been wonderful working with all of you. Considering I trained most of you, I am so proud of how far you come (in spite of my training). Please accept this as a formal apology for every time I have forgotten to do a lab chore or have stunk up the lab with cell pellets. I was the worst. Special thanks to Josh and Evan, who literally helped with the heavy lifting in the beginning of the lab. And to my undergrad, Will Huggins, for sticking it through the first year. I still regret torturing you with so much biochemistry, now knowing your synthesis skills. I am glad you are happily past our cell culture days. To members of the red and blue pods, thank you for making our shared space truly shared. You have saved my science and my sanity so many times over the years.

I would also like to thank my closest friends, from near and far, for being there for me during all of this: Lydia, Billy, Kennan, Rebecca, Will, Bertucci, Hamilton, Domby, Meghan,

Hannah, Katia, Laura, Emily, Tom, and Satusky. You are truly the best. Now you can argue amongst yourselves as to why I put you all in that order.

And finally, I would like to thank my family for their love, support, and frequent flyer miles over the years. To my parents, I cannot thank you enough for all you have done to help me get here. Thank you for never putting limits on my opportunities and for accepting my choice to move so far away from home. Who knew all those times I trashed the kitchen baking would lead to this? And hopefully that makes up for the fact that I never did the dishes.

For anyone else I may have left out, thank you so much. I clearly left you out as to not make anyone jealous. And to you, reader, for actually reading the acknowledgements. Either you are one of the people listed in this section, or just incredibly thorough. Either way, thank you for reading!

TABLE OF CONTENTS

LIST OF MAIN TABLES

LIST OF MAIN FIGURES

LIST OF SUPPLEMENTARY TABLES

LIST OF SUPPLEMENTARY FIGURES

LIST OF ABBREVIATIONS AND SYMBOLS

1MeHis	1-methylhistidine
3MeHis	3-methylhistidine
aaRS	Amino acyl tRNA synthetase
Ala, A	Alanine
Amp	Ampicillin
Arg, R	Arginine
Asn, N	Asparagine
Asp, D	Aspartate
BocK	Boc-lysine
BzTAZ	Benzyl thiazolylalanine
Cam	Chloramphenicol
cAMP	Cyclic adenosine monophosphate
Cat	Chloramphenicol acyltransferase gene
CBX	Chromobox
CH_3I, MeI	Methyl iodide
Cys, C	Cysteine
DCM	Dichloromethane
DIPEA	Diisopropylethylamine
DMF	N,N-dimethylformamide
DMH	Dimethyl histidine
DMNB	4,5-dimethoxy-2-nitrobenzylserine
DTT	Dithiothreitol

EPL	Expressed protein ligation
ESI-LCMS	Electrospray ionization liquid chromatography – mass spectrometry
EtOH	Ethanol
Gln, Q	Glutamine
Glu, E	Glutamate
Gly, G	Glycine
H3K27me3	Trimethyllysine at position 27 of histone 3
H3K9me3	Trimethyllysine at position 9 of histone 3
HBTU	O-benzotriazole-N, N, N', N',-tetramethyluronium hexafluorophosphate)
HCl	Hydrochloric acid
His, H	Histidine
HOBt	N-hydroxybenzotriazole
HP1	Heterochromatin Protein 1
Ile, I	Isoleucine
ITC	Isothermal titration calorimetry
K_a	Acid dissociation constant
Kan	Kanamycin
K_d	Dissociation constant
Kme	Monomethyllysine
Kme2	Dimethyllysine
Kme3	Trimethyllysine
Leu, L	Leucine

LFER	Linear free energy relationship
Log P	Hydrophobicity parameter
Lys, K	Lysine
MeCN	Acetonitrile
Met, M	Methionine
MeTAZ	Methyl thiazolylalanine
mmPylRS	Pyrrolysine amino acyl tRNA synthetase from *M. Mazei*
mTNF	Murine tumor necrosis factor
Na^+	Sodium ion
NaCl	Sodium Chloride
$NaHCO_3$	Sodium bicarbonate
NaOET	Sodium ethoxide
NaOH	Sodium hydroxide
NCL	Native chemical ligation
NHC	N-heterocyclic carbene
NNK	Randomized library site where N = any base and K= G or T
PBS	Phosphate-buffered saline
pCF_3F	*p*-Trifluoromethylphenylalanine
pCH_3F	*p*-Methylphenylalanine
*p*ClF	*p*-Chlorophenylalanine
*p*CNF	*p*-Cyanophenylalanine
*p*CNPheRS	*p*-Cyanophenylalanine amino acyl tRNA synthetase
Phe, F	Phenylalanine

pNO_2F	*p*-Nitrophenylalanine
PRC1	Polycomb repressive complex 1
PRC2	Polycomb repressive complex 2
Pro, P	Proline
PTM	Post-translational modification
pylHRS	Histidine synthetase evolved from PylRS
PylRS	Pyrrolysine amino acyl tRNA synthetase
SDS-PAGE	Sodium dodecyl sulfate polyacrylamide gel electrophoresis
Ser, S	Serine
sfGFP	Superfolder green fluorescent protein
$SOCl_2$	Thionyl chloride
Spec	Spectinomycin
SPPS	Solid phase peptide synthesis
Strep	Streptomycin
TAG	Amber stop codon
TAZ	Thiazolylalanine
TCEP	Tris(2-carboxyethyl)phosphine)
Tet	Tetracycline
ThDP	Thiamine diphosphate
Thr, T	Threonine
TIC	Total ion count
TIPS	Triisopropylsilane
tRNA	Transfer RNA

Trp, W	Tryptophan
Tyr, Y	Tyrosine
TyrRS	Tyrosine amino acyl tRNA synthetase
UAA	Unnatural amino acid
Val, V	Valine
VDW	van der Waals forces
WT	Wild type
ΔG_b	Gibbs free energy of binding
σ_{meta}	Sigma meta constant

CHAPTER 1: AN INTRODUCTION TO UNNATURAL AMINO ACID MUTAGENESIS

1.1 Nature is Diverse, but Chemically Limited

Proteins are diverse biopolymers constructed from 20 canonical amino acids. With these 20 building blocks, innumerable amino acid combinations are possible. This combinatorial diversity allows for a variety of essential cellular functions ranging from enzyme catalysis, molecular recognition, signal transduction, structure, and beyond.[1,2] Although nature has successfully evolved proteins for numerous purposes, protein function is limited by the restricted chemical functionality of the canonical amino acids. For example, canonical amino acids include only a handful of residues capable of participating directly in catalysis (i.e. a few acids, bases, or nucleophiles, Figure 1.1). As a result, enzyme catalysis often requires multiple residues working in a highly coordinated fashion, such as the catalytic triad of aspartate, histidine, and serine found in many proteases.[3] In addition, in order to study the role of amino acids in protein function, it would be desirable to make fine-tuned mutations to the side chain. But only a few mutations are conservative with respect to side chain structure. Asp → Asn, Glu → Gln, and Ser → Cys, for example, change only one atom of the amino acid side chain. For the remaining amino acids, mutations to another canonical amino acid may introduce significant changes to the amino acid side chain. Although information may be discovered from such mutations, it is often difficult to determine how protein function would change with more subtle modulations of the chemical properties of the side chain.

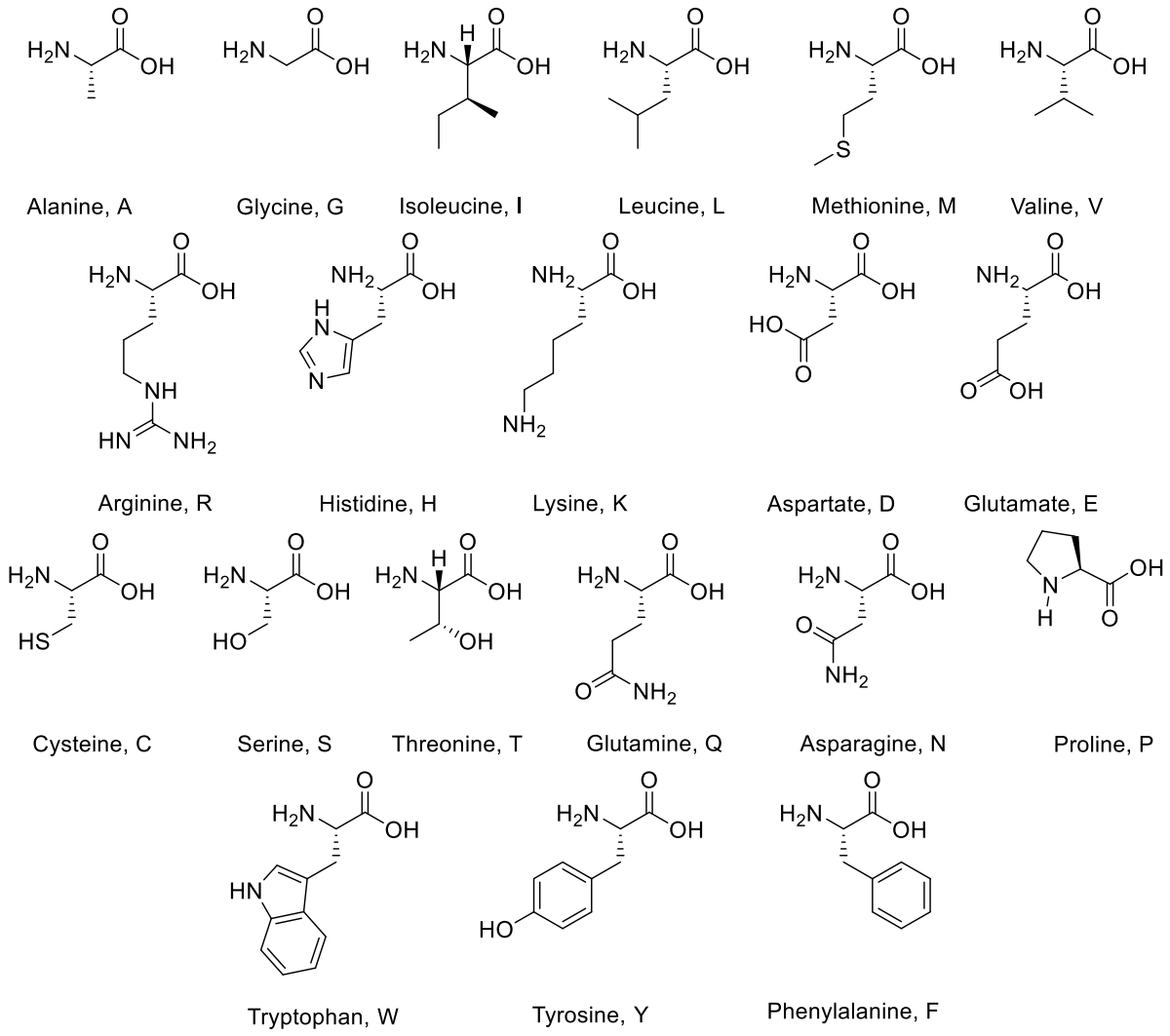


Figure 1.1 Structures of the canonical amino acids.

Introduction of unnatural amino acids (UAAs) to the genetic code can greatly expand the ability to tailor protein scaffolds at the atomic level, similar to synthetic chemists' precise control over the structure of small molecules they create. Introduction of designed and chemically synthesized UAAs into a protein allows for new functional manipulation of macromolecules that cannot be achieved through traditional protein mutagenesis.

A variety of methods have been used to incorporate UAAs into peptides and proteins. For example, solid phase peptide synthesis (SPPS) has proven to be a robust technique to

chemically synthesize short polypeptides (tens of amino acids).[1,4,5] Proteins made by SPPS can be made from any combination of both natural and unnatural amino acids. This method, however, is limited to small peptides and proteins, since longer peptide chains are more difficult to produce and purify in good yield due to even small inefficiencies in amino acid coupling. Nevertheless, SPPS has gained traction for building unnatural proteins through the assembly of peptide fragments via thioester-mediated protein assembly methods including native chemical ligation (NCL) and expressed protein ligation (EPL).[1,2]

Chemical approaches have also been used to incorporate UAAs into proteins in a manner that utilizes the biosynthetic machinery of the cell (i.e. the ribosome). Chemically-aminoacylated transfer RNAs (tRNAs) provide a mechanism to exploit the natural translation apparatus for UAA incorporation, and UAA-charged tRNAs can be used in cell free translation systems or used in cells (typically *Xenopus laevis* oocytes) via microinjection.[6,7] These methods typically use an amber stop codon-decoding tRNA that has been aminoacylated with an UAA using chemical approaches. Amber stop codons are used in the lowest frequency in bacteria, and are therefore less likely to be used in cell housekeeping proteins. Repurposing the stop codon to incorporate UAAs limits competition with "sense" canonical amino acid incorporation by exploiting read-through of "nonsense" codons; however, it may compete in some degree with translational termination.

The ribosome and translation elongation factors are highly promiscuous; the chemical diversity of the 20 canonical amino acids is a testament to this fact. Fortunately, this native promiscuity allows for incorporation of diverse unnatural amino acid side chains. While chemical aminoacylation is powerful – hypothetically, any amino acid could be incorporated if it can be chemically attached to a tRNA – the method requires a stoichiometric amount of

aminoacylated tRNA for protein production, which severely limits UAA-protein yield. As a result, this approach is often used for protein investigations that can be carried out under very low protein concentrations such as single cell or single molecule experiments.

1.2 Incorporating UAAs *In Vivo*

UAAs became more widely used after the development of *in vivo* UAA mutagenesis. First introduced by Peter Schultz and coworkers, this *in vivo* approach took advantage of an enzymatic method to attach UAAs to tRNAs, in lieu of synthetic aminoacylation.[1,2,8] In nature, the enzymes responsible for aminoacylating or "charging" a tRNA with its respective amino acid are known as amino acyl tRNA synthetases (aaRSs). There are twenty synthetases in the cell, each responsible for one of the canonical amino acids. Each synthetase has two essential domains: one for tRNA and anticodon recognition and another for amino acid recognition and catalysis (Figure 1.2). Synthetases are highly specific for their canonical amino acid and are well conserved across nature. However, the molecular recognition motifs that link the synthetase and corresponding tRNA vary enough that there is often little to no cross-reactivity when aaRS/tRNAs from one domain of life are introduced into another domain. For example, a tyrosine tRNA from archaea will not be charged with tyrosine by a bacterial TyrRS, and vice versa. This backwards-evolutionary compatibility allowed Schultz and coworkers to develop "orthogonal" aaRS-tRNA pairs. By engineering the catalytic domain of aaRSs to recognize new amino acids, Schultz and coworkers overcame the stoichiometric limitation of previous biosynthetic approaches. Once an UAA has been incorporated into the growing peptide chain, the uncharged tRNA may be recharged

iteratively, as long as free UAA remains available, in a manner that mimics native tRNA synthetases.

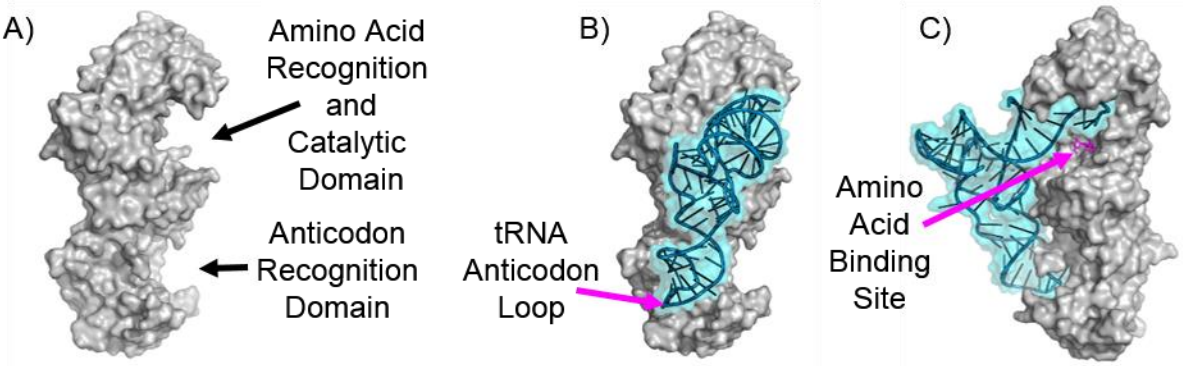


Figure 1.2 Structure of an aaRS-tRNA complex. A) Structure of the glutaminyl-aaRS (PDB ID: 1GTR).[9] B) Glutaminyl aaRS (gray) complexed with cognate tRNA (cyan). C) Glutaminyl aaRS-tRNA complex rotated to show amino acid binding pocket containing ATP (magenta).

While a number of orthogonal aaRS/tRNA pairs have been developed, only a few systems have found widespread use. Historically, the most widely used system derives from the tyrosine synthetase and corresponding tRNA isolated from the archaea *Methanocaldococcus jannaschii*.[1,2] Small mutations were required in these aaRS/tRNA pairs to achieve complete orthogonality. In order for an aaRS-tRNA pair to be orthogonal, it must meet the following requirements: 1) the orthogonal tRNA is not charged with any native amino acids, 2) the orthogonal aaRS does not charge native tRNAs with the UAA, and 3) the orthogonal tRNA does not have same anticodon as any native tRNAs.[1,2,10] These requirements ensure that UAAs are incorporated site specifically. Non-specific UAA incorporation, due to errors in any of these requirements, causes a mixture of protein products. Therefore, orthogonality is an important feature to permit the homogenous production of target proteins bearing UAAs at select locations, and to prevent cytotoxicity due to misincorporation of UAAs throughout the proteome.

To enable orthogonal UAA activation, the TyrRS amino acid binding site was mutated (Figure 1.2A) to permit recognition of tyrosine-like UAAs while at the same time discriminating against tyrosine.[1,2] Using this approach, a large number of tyrosine like unnatural amino acids have been incorporated into proteins (a select panel is shown in Figure 1.3). The anticodon recognition domain (Figure 1.2A) was also mutated to recognize the amber stop codon (UAG) rather than a tyrosine codon (UAU or UAC) so that native tyrosine tRNAs were not charged with UAAs. The amber stop codon was chosen instead of the opal and ochre stop codons (UGA and UAA, respectively) due to the infrequency of amber codon, limiting competition between the UAA incorporation and termination of translation. Finally, the tRNA's anticodon loop (Figure 1.2B) was also changed to an amber codon, thus creating an orthogonal pair.

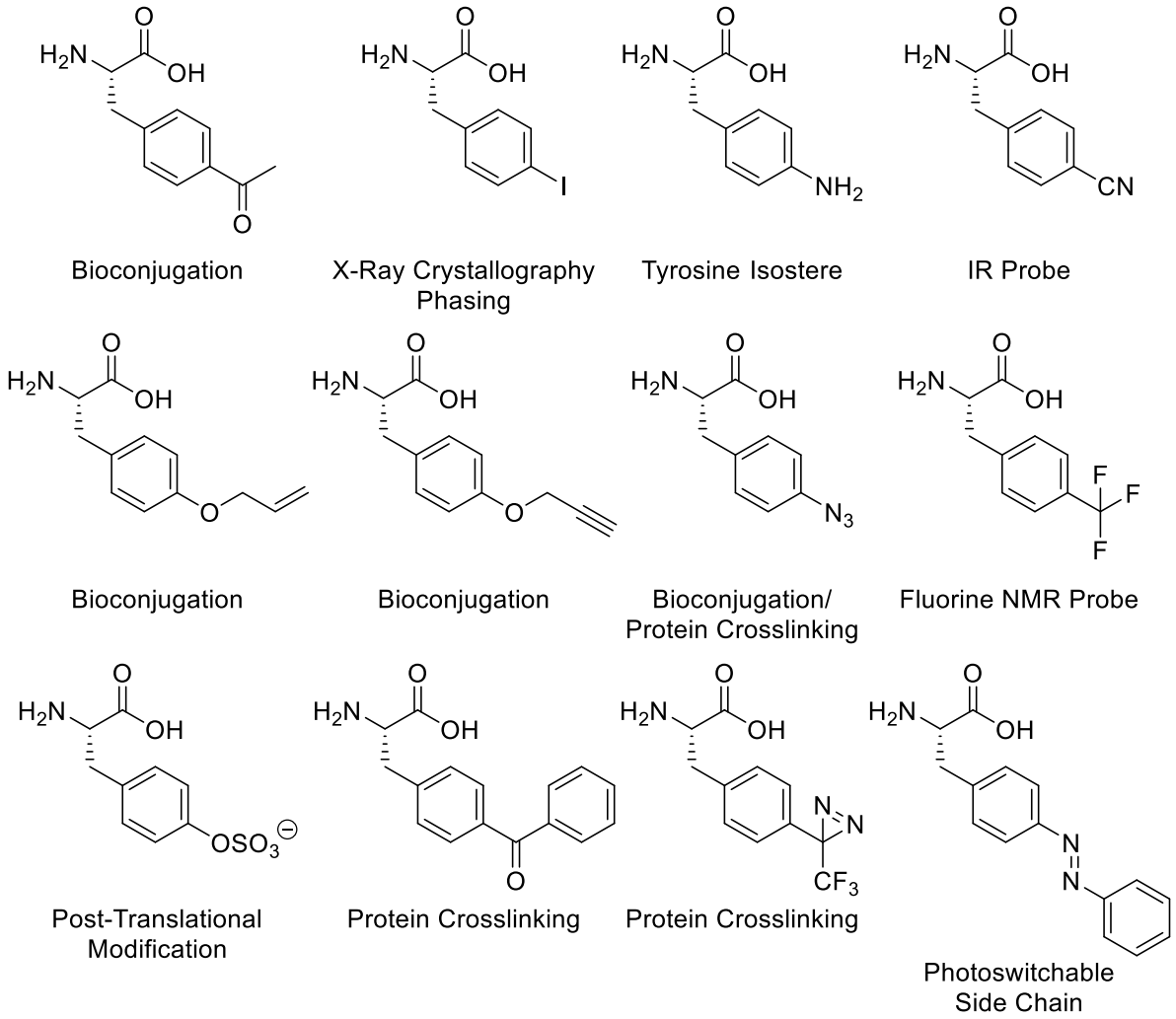


Figure 1.3 Structure of some previously incoroprated tyrosine-derived UAAs with their respective applications.[1,2]

In practice, once an orthogonal UAA-aaRS/tRNA pair has been generated, application to UAA protein production is relatively straightforward. To incorporate UAAs into proteins in *E. coli*, an amber stop codon is cloned into the gene at the site of interest. The vector containing the TAG-variant of the gene can then be co-transformed with an auxiliary vector containing the orthogonal tRNA-synthetase pair. Proteins are expressed by supplementing traditional growth media with the UAA of choice, and UAA-tagged protein can be purified using standard methods. This *in vivo* UAA method has now been applied to a

variety of proteins and UAAs. With simple cloning and use of an accessory plasmid, UAAs can feasibly be incorporated into any protein with a known gene sequence.

1.3 Limitations of *In Vivo* UAA Mutagenesis

In vivo unnatural amino acid mutagenesis' most prominent limitation is synthetase specificity. Although many unnatural amino acids have been incorporated *in vivo*, most are tyrosine, phenylalanine, or lysine derivatives.[1,2] Since specific synthetases do not yet exist for certain UAAs, synthetases must be evolved or screened for incorporation of a novel UAA. Synthetase selection and screening for UAA incorporation is often non-trivial, requiring large amounts (grams) of expensive or synthetically challenging UAA and requiring considerable experimental investment – construction of tRNA synthetase libraries followed by multiple rounds of mutation and selection for orthogonality and improved UAA incorporation. In addition, once a synthetase has been chosen, the performance of each synthetase is highly variable and yields can fluctuate wildly for different UAAs in a manner that is dependent on the UAA, aaRS, or even on the plasmid/vector system used for expression. Certain synthetases perform significantly better in one plasmid over another, and as a result multiple vectors must be screened in the process of unnatural protein optimization. Since Schultz's *in vivo* approach introduces chemically synthesized UAAs into cells, it is hard to predict how these molecules will behave inside the cell. Some UAAs are toxic, and others can be degraded through various cellular mechanisms. In the end, there is no guarantee that a designed UAA can be incorporated *in vivo*, and therefore UAA mutagenesis can be a risky, time intensive, and costly scientific endeavor that may be highly successful, or yield no tangible results.

1.4 Applications of UAA Mutagenesis

Since the advent of *in vivo* UAA mutagenesis in 2001, there has been growing literature demonstrating the use of this technology to study or manipulate protein structure and function. While a general review of this field is beyond the scope of this dissertation, the following sections serve to highlight a few of the exciting applications that are attainable through this methodology.

1.4.1 Control of Protein Phosphorylation Using a Photocaged UAA

UAAs have been used to control protein translocation within the cell.[1,11] Pho4 is a transcription factor that plays a major role in the yeast signaling pathway that allows for growth under different inorganic phosphate concentrations.[11] In low phosphate conditions, Pho4 is localized to the nucleus and is hypophosphorylated. In high phosphate conditions, Pho4 is phosphorylated by a cyclin-cyclin dependent kinase complex. Pho4 can be phosphorylated at multiple serine residues, which activates transcription, triggers export of Pho4 to the cytoplasm, or prevents reuptake of Pho4 into the nucleus. Preliminary studies identified two serine residues in Pho4 (Ser114 and Ser128) as important sites for phosphorylation to control nuclear export; however, the mechanism of action and role for each of these residues remained unknown and it was unclear whether phosphorylation of Ser114 or Ser128 or both could cause distinct Pho4 responses.

To determine how phosphorylation of these two serine residues affected Pho4 activity, a photocaged serine (4,5-dimethoxy-2-nitrobenzylserine (DMNB-Serine), Figure 1.4) was encoded in place of each serine residue of interest in Pho4. A GFP fusion tag was also appended to the protein to allow for Pho4 to be tracked within the cell using

fluorescence. The DMNB protecting group on serine blocks the site of Pho4 phosphorylation, causing Pho4-GFP to build up within the nucleus (Figure 1.4B before [1]). Treatment of the caged protein with blue light led to photolysis of the DMNB group from serine (Figure 1.4B, [1]), generating wild type Pho4-GFP. Upon decaging, the Pho4 serine residue can be phosphorylated leading to export from the nucleus (Figure 1.4B, [2]), as evidenced by a decrease in nuclear fluorescence and a concomitant increase cytoplasmic fluorescence. The disappearance of fluorescence in the nucleus was measured as a function of time, allowing kinetic data to be acquired for each serine, independently. Intriguingly, distinct export kinetics were observed for the different phosphoserine Pho4 isoforms. This result suggested that Pho4 function is dynamically regulated, and that the strength or role of effector function derived from the post-translational modification is highly dependent on the location of the modification.

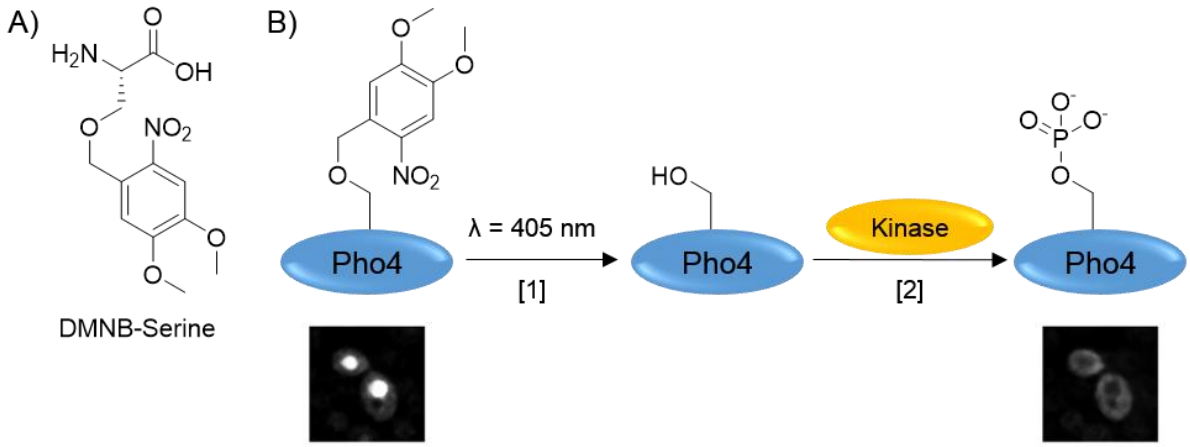


Figure 1.4 Application of a photocaged UAA. A) Structure of DMNB-Serine. B) Photolysis of DMNB group allows for tracking of Pho4 within the cell. Figure adapted from Chang and Schultz, 2010.[1]

1.4.2 Purification of Native Proteins Using a Boronate UAA

Scarless affinity purification, or purification without additional tags or "scars," has also been achieved using UAAs. When *p*-boronophenylalanine, a boronate-containing

unnatural amino acid, is incorporated into proteins, the UAA can be chemically oxidized or reduced, to afford a native amino acid, tyrosine or phenylalanine, respectively (Figure 1.5).[12] This unique chemical handle therefore provides a chemically-responsive masked analog of these two amino acids. In addition, in aqueous solutions and at physiological pH, boronic acids form strong, covalent interactions with polyhydroxylated compounds, such as N-methylglucamine and other sugars. *p*-Boronophenylalanine was incorporated at a tyrosine residue of the Z-domain of staphylococcal protein A using *in vivo* UAA mutagenesis. Boronophenylalanine-tagged proteins were then purified selectively in response to the boronic acid functional group using an N-methylglucamine conjugated resin with yields comparable to a traditional 6His-TAG/Ni-NTA purification. Proteins containing the boronate UAA were eluted off of the resin either using excess polyol to give the UAA-containing protein or with hydrogen peroxide to give a native protein sequence baring a tyrosine in place of the UAA. This method allows for protein isolation based on the unique functional groups of the UAA without the need for affinity tags.

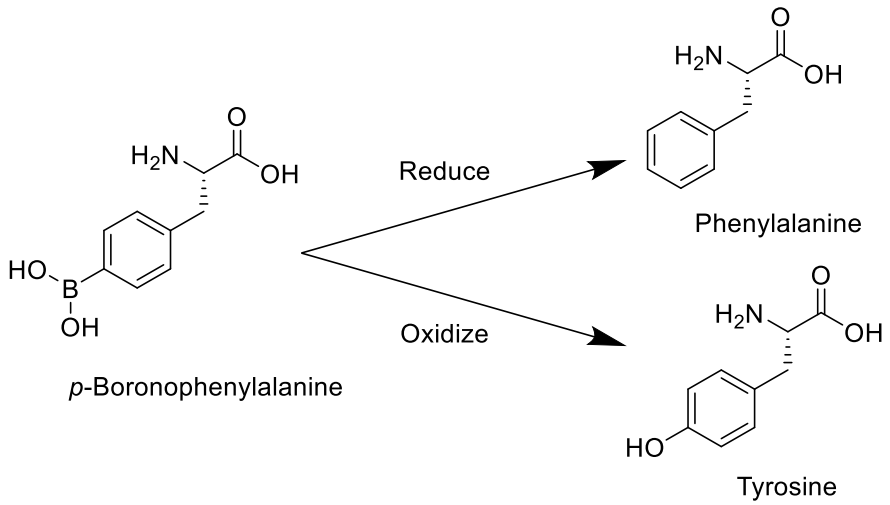


Figure 1.5 *p*-Boronophenylalanine may be oxidized or reduced to native amino acids

1.4.3 Termination of Immunological Self-Tolerance Using Nitroaryl UAAs

UAAs are finding growing use in the field of protein therapeutics. Immunotherapeutics, which activate the immune system to fight disease, can face major challenges when immunological targets are self-proteins that traditionally do not elicit immune responses. This immune tolerance of self-proteins is an important natural mechanism to prevent auto-immune disease, however, in illnesses of inflammation or abnormal cell division (cancer), the ability to convince the immune system to target host proteins could be provide a useful therapeutic strategy.[1,13]

Previously, rabbits immunized with a rabbit thyroglobulin that had been nonspecifically modified with a reactive diazonium compound produced cross-reactive antibodies to native thyroglobulin.[13] This self-protein immune response suggested that chemical modification of proteins may lead to novel immunogenic epitopes that induce high-titers of cross-reactive antibodies, even against a self-protein. Historically, nitroaryl groups have been used to elicit immune responses against small molecule antigens. In light of the fact that UAAs containing nitroaryl groups (such as *p*-nitrophenylalanine (*p*NO_2F), Figure 1.6A) have been site specifically incorporated into proteins in response to the amber TAG codon, Schultz and coworkers asked whether modifications of proteins with UAAs might lead to a general strategy to break immunological self-tolerance.

Murine tumor necrosis factor (mTNF-α) was selected as the protein of study due to its well-characterized involvement in the regulation of infectious, inflammatory, and autoimmune responses.[13] The signaling mechanisms, structure, and function of mTNF-α have been comprehensively studied and mTNF-α knockout mice show no apparent abnormalities. When mice are challenged with bacterially-derived lipopolysaccharide, mTNF-α leads to

inflammation responses that are cytotoxic. As a test system, auto-antibodies that are able to sequester mTNF-α might lead to attenuation of this inflammation response and resulting toxicity. Mammalian TNFs share a highly conserved tyrosine residue, Tyr86. Mutations to this residue lead to significant decrease in cytotoxicity with no effect on protein folding or trimer formation.

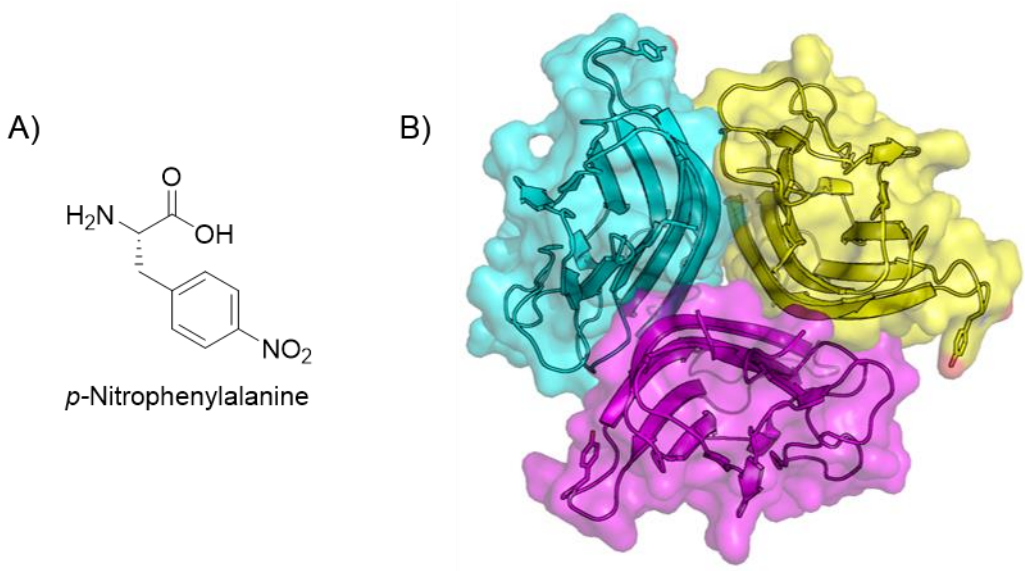


Figure 1.6 Structure of A) *p*-Nitrophenylalanine and B) mTNF-α (PDB ID: 2TNF). Tyr86 residues are shown as sticks.

To determine the effect of the *p*NO_2F mutation, mice were immunized with either wild type, Y86*p*NO_2F, or PBS buffer. Mice immunized with wild type protein or PBS buffer had no significant serum IgG titers against Y86*p*NO_2F nor wild type mTNF. This is expected since the wild type mTNF is a self-protein and should not elicit an immune response. Mice immunized with Y86*p*NO_2F had noticeably high IgG titers. Importantly, these antibodies showed cross-reactivity against both Y86*p*NO_2F mTNF and the wild type protein. To confirm the immune response resulted from the nitro group of the UAA, the Y86F mutant of mTNF was generated. Mice immunized with Y86F showed no significant anti-mTNF titers, suggesting the nitroaryl group is requisite in breaking immunological tolerance. The

vaccination of mice with Y86*p*NO_2F was next tested in an mTNF-α-dependent severe endotoxemia mouse model. It was found that Y86*p*NO_2F immunized mice had a significantly higher chance of survival than mice immunized with wild type mTNF (87.5% vs. 12.5%, respectively, Figure 1.7). This increased survival rate suggests the UAA-modified self-protein induced a robust cross-reactive immunological response against the native protein that conferred a protective advantage in a disease model. The introduction of *p*NO_2-containing UAAs to self-proteins may provide a general method of inducing immune responses for therapeutics.

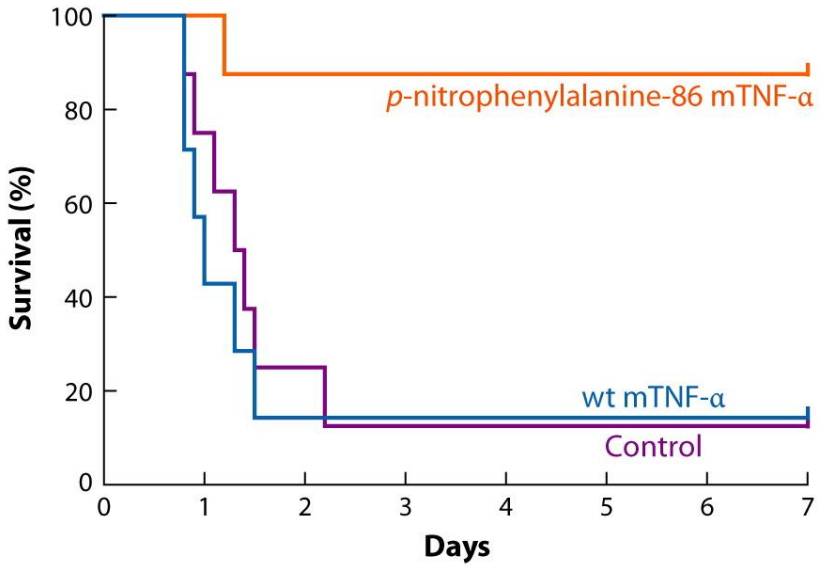


Figure 1.7 Y86*p*NO_2F confers a survival advantage in an mTNF-α-dependent disease model. Figure adapted from Chang and Schultz, 2010.[1]

1.5 Areas for Development

Although UAAs have brought great advances to the field of protein science, there are still applications waiting to be explored. One notable deficiency in UAA mutagenesis research is the application of UAAs to generate new enzymes that carry out chemistry not available to the 20 canonical amino acids. While UAAs have been encoded into enzyme

active sites in order to alter substrate specificity or reaction stereoselectivity,[1,2,12,14] a catalytically active UAA that enables new chemistry has yet to be incorporated into proteins *in vivo*. Nevertheless, the fruits of natural selection suggest that increased chemical diversity in proteins is an essential approach to augment biocatalysis. For example, nature has already expanded its chemical toolkit past the 20 amino acids by recruiting cofactors to the active site.[1] These cofactors are small molecules or metal ions that are required for diverse chemical reactions (e.g. electron transfer, oxidation, and reduction reactions) that are not genetically encoded. Some cofactors can be synthesized by a given organism, other must be included in diet or taken in from the environment. Genetically encoding a cofactor-like UAA could give the functional power of the catalytically active UAA along with the defined stereochemistry conferred by the enzyme active site.

Unnatural amino acids are also underutilized in biophysical characterization of proteins. For example, ligand binding often requires multiple noncovalent interactions with proximal amino acid residues that work in concert to recognize a ligand. However, due to the limited repertoire of canonical amino acids, traditional mutagenesis to binding sites can cause large changes in the biding pocket that are difficult to interpret. In short, very few natural mutations (e.g. Asp → Asn, Ser → Cys) can produce subtle changes to a side chain. On the other hand, UAA mutagenesis allows highly tailored modifications of amino acid side chains, permitting even single atoms on the amino acid side chain to be added, removed, or substituted. This molecular precision allows for a more in depth investigation into interactions such as hydrogen bonding, cation-π interactions, hydrophobic effects, and ionic interactions. Although UAAs have been used to investigate binding interactions, a

comprehensive study of this type has not yet been undertaken using *in vivo* UAA mutagenesis.

The purpose of this dissertation is to expand the applications of unnatural amino acid mutagenesis into two diverse fields: 1) biophysical investigations of molecular recognition motifs that exploit cation-π interactions, and 2) biocatalysis using cofactor-like UAAs. In chapter 2 and 3, we will outline efforts to examine protein-ligand interactions in epigenetic reader proteins. Cation-π interactions have been implicated as a driving force behind recognition of cationic ligands by this family of proteins. UAAs with different electron-withdrawing –donating groups were incorporated into the binding site, thereby tuning the electronics of the cation-π interaction. In chapter 2, we describe our efforts on a model methyllysine reader system, Heterochromatin Protein 1. In chapter 3, we will expand our approach into two mammalian readers, CBX5 and CBX7, which are implicated in multiple cancers. In chapter 4, we will describe our efforts to incorporate a thiamine-like amino acid. Thiamine is thought to be the first evolved cofactor and is required for many important carboxylation and decarboxylation processes of the cell.[15] Thiamine's catalytic core contains an N-heterocyclic carbene (NHC), which is a class of potent small molecule catalysts in synthetic chemistry. Despite the availability of histidine, a canonical NHC-containing amino acid, all life requires thiamine, and plants, fungi, and bacteria have evolved biosynthetic pathways for this cofactor. Harnessing the catalytic potential of a NHC in the form of an UAA allows us to introduce NHC catalysis to new active sites, thereby expanding the catalytic potential of enzymes. These studies will expand the applications to the already flourishing field of *in vivo* UAA mutagenesis.

REFERENCES

(1) Liu, C. C.; Schultz, P. G. *Annu. Rev. Biochem.* **2010**, *79*, 413–444.

(2) Wang, L.; Xie, J.; Schultz, P. G. *Annu. Rev. Biophys. Biomol. Struct.* **2006**, *35*, 225–249.

(3) Polgár, L. *Cell. Mol. Life Sci.* **2005**, *62* (19–20), 2161-2172.

(4) Bianco, A.; Townsley, F. M.; Greiss, S.; Lang, K.; Chin, J. W. *Nat. Chem. Biol.* **2012**, *8* (9), 748–750.

(5) England, P. M. *Biochemistry* **2004**, *43* (37), 11623–11629.

(6) Noren, C. J.; Anthony-Cahill, S. J.; Griffith, M. C.; Schultz, P. G. *Science* **1989**, *244* (4901), 182–188.

(7) Beene, D. L.; Dougherty, D. A.; Lester, H. a. *Curr. Opin. Neurobiol.* **2003**, *13* (3), 264–270.

(8) Chin, J. W.; Schultz, P. G. *Chembiochem* **2002**, *3* (11), 1135–1137.

(9) Rould, M. A.; Perona, J. J.; Steitz, T. A. *Nature* **1991**, *352* (6332), 213–218.

(10) Wang, L.; Brock, A.; Herberich, B.; Schultz, P. G. *Science* **2001**, *292* (5516), 498–500.

(11) Lemke, E. A.; Summerer, D.; Geierstanger, B. H.; Brittain, S. M.; Schultz, P. G. *Nat. Chem. Biol.* **2007**, *3* (12), 769–772.

(12) Brustad, E.; Bushey, M. L.; Lee, J. W.; Groff, D.; Liu, W.; Schultz, P. G. *Angew. Chem. Int. Ed. Engl.* **2008**, *47* (43), 8220–8223.

(13) Grünewald, J.; Tsao, M.-L.; Perera, R.; Dong, L.; Niessen, F.; Wen, B. G.; Kubitz, D. M.; Smider, V. V; Ruf, W.; Nasoff, M.; Lerner, R. A.; Schultz, P. G. *Proc. Natl. Acad. Sci. U. S. A.* **2008**, *105* (32), 11276–112780.

(14) Plass, T.; Milles, S.; Koehler, C.; Schultz, C.; Lemke, E. A. *Angew. Chemie Int. Ed.* **2011**, *50* (17), 3878–3881.

(15) Frank, R.; Leeper, F.; Luisi, B. *Cell. Mol. Life Sci.* **2007**, *64* (7-8), 892–905.

CHAPTER 2: PROBING CATION-π INTERACTIONS OF HETEROCHROMATIN PROTEIN 1 USING IN VIVO UNNATURAL AMINO ACID MUTAGENESIS

2.1 Epigenetic Control of Gene Expression

Although the human body contains many diverse tissues, the vast majority of cells contain identical DNA. With each cell having the same DNA template, regulation of gene expression is critical to cell differentiation, development, and function. Epigenetics – or the study of heritable changes in gene expression or function that are not caused by DNA sequence – studies how our genes are turned on and off inside of cells. In recent years, links between gene expression and cell dysfunction and disease have become better established, leading to increased interest in the mechanisms underlying epigenetic control.

Gene expression (or lack thereof) is significantly governed by the way DNA is compacted within the cell. DNA is packaged and condensed using histones, proteins in the nucleus that act as a structural bobbin around which DNA is wound and packaged.[1] Two copies of each of the four histone proteins (H2A, H2B, H3, and H4) are wrapped with ~146 base pairs to form the nucleosome (Figure 2.1).[2] DNA coils around multiple histones along the same DNA strand to eventually condense to form a chromatin fiber (Figure 2.2A). Chromatin can be further classified into two subcategories: heterochromatin and euchromatin (Figure 2.3). Heterochromatin is composed of DNA tightly wrapped around histones and contains genes that are mostly inactive.[1,3] Conversely, euchromatin is characterized by loosely coiled DNA that better allows for gene transcription.[1,3]

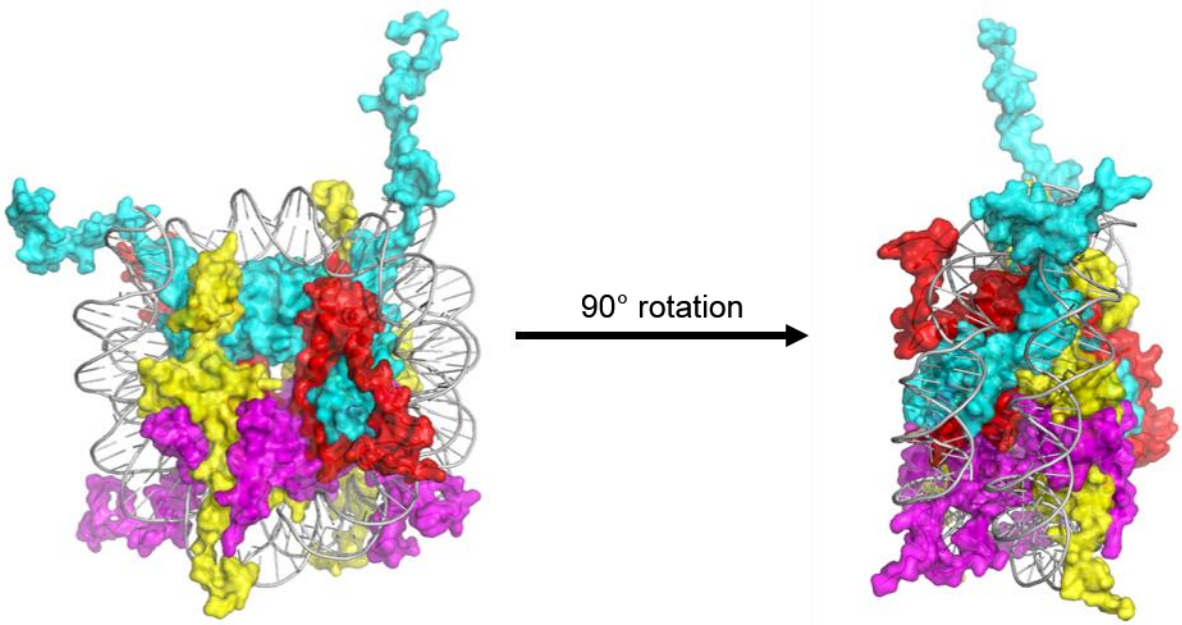


Figure 2.1 Structure of a DNA-histone complex (PDB ID: 1KX5). The histone is comprised of two of each of subunit: H4 (red), H3 (cyan), H2A (yellow), and H2B (magenta). The DNA strand is shown in gray.

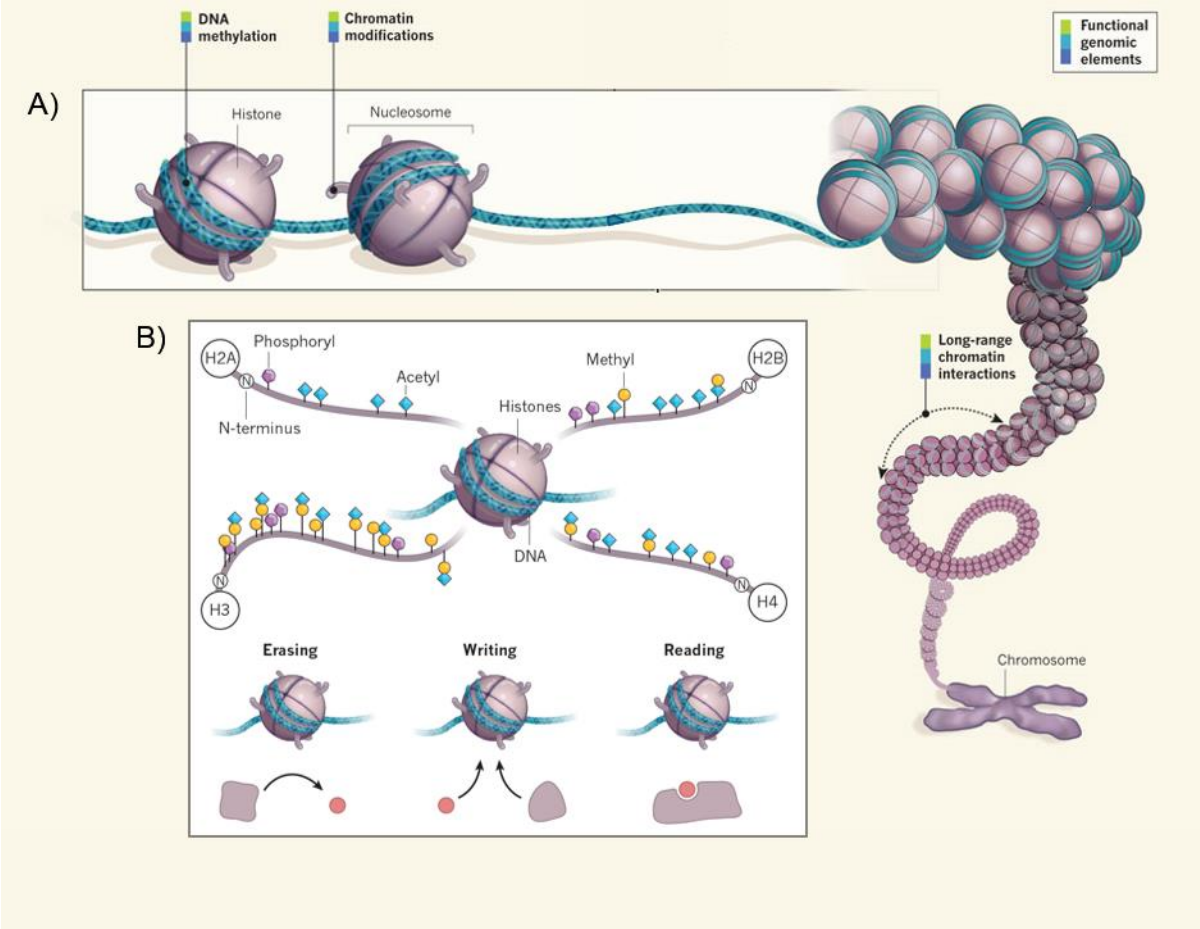


Figure 2.2 Histone interactions inside the cell. A) DNA (teal) is wound around histone proteins (purple) to form the nucleosome. Tightly packed nucleosomes condense to form chromatin fibers, which are further compacted to form chromosomes. B) The protein tails of histone proteins have extensive post translational modifications (PTMs). These PTMS may be removed by eraser proteins (erasing), added by writer proteins (writing), or interpreted by reader proteins (reading). Adapted from and reprinted by permission from Macmillan Publishers Ltd: NATURE Helin, K.; Dhanak, D. *Nature* **2013**, *502* (7472), 480–488, copyright 2013, and NATURE Ecker, J. R.; et al. *Nature* **2012**, *489* (7414), 52–55, copyright 2012.

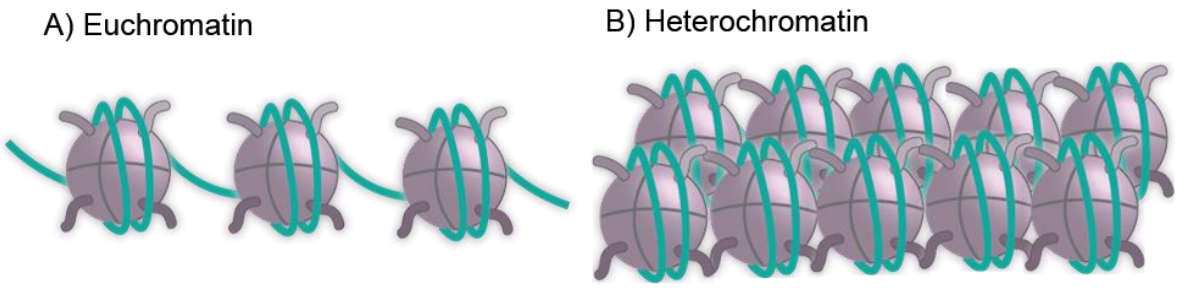


Figure 2.3 Structure of euchromatin (A) and heterochromatin (B).

Histone proteins are extensively post-translationally modified, especially at the histone tails (Figure 2.2B).[1,2,4] These post-translational modifications (PTMs) are covalent modifications of the amino acid side chains that are added after ribosomal translation. PTMs are installed by proteins known as writer proteins, are removed by eraser proteins, and are recognized and interpreted by reader proteins.[4] The most common examples of post-translational modifications are phosphorylation, acetylation, methylation, and ubiquitination.[5] PTMs have two major effects on chromatin: altering interactions of histone proteins with DNA and providing a chemical tag for the recruitment of other protein factors that lead to epigenetic regulation.[1,2]

2.2 Methyllysine Reader Proteins

Lysine methylation of histone proteins is a well-recognized means of epigenetic control. Lysine can be post-translationally modified with one, two, or three methyl groups, to create mono-, di-, or trimethyllysine, respectively (Figure 2.4).[6–8] Lysine methylation is an interesting PTM because methylation does not alter the charge of lysine.[9] Lysine reader proteins are able to discriminate methyllysines from unmodified lysine, which illustrates their exacting specificity for molecular recognition. In humans, these readers utilize a cage of aromatic residues that interact with the N-methyl groups of methylated lysine (Kme_n)

residues (Figure 2.5). Not surprisingly, cation-π interactions have been implicated as major contributors to Kme_n recognition.[10]

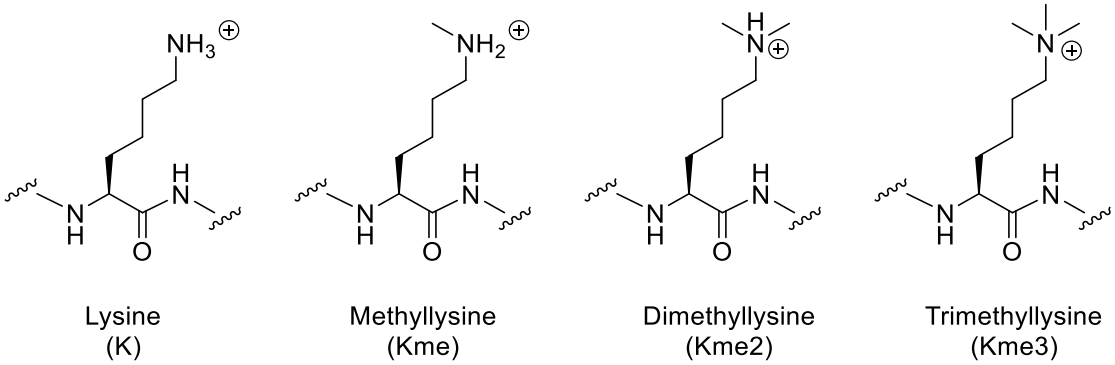


Figure 2.4 Methylation states of lysine.

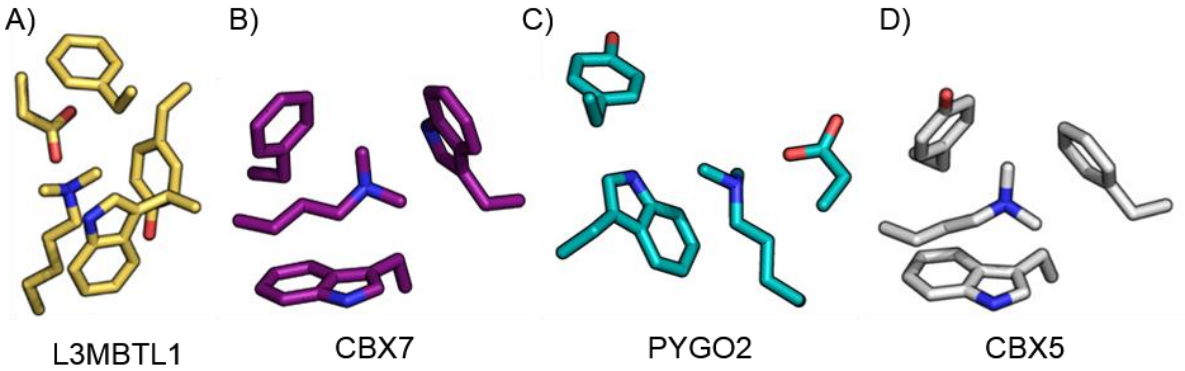


Figure 2.5 Structures of methyllysine reader proteins. A) Lethal (3) malignant brain tumor-like protein 1, L3MBTL1 (PDB ID: 2PQW). B) Chromobox protein homolog 7, CBX7 (PDB ID: 4X3K). C) Pygopus homolog 2, Pygo2 (PDB ID: 4UP0). D) Chromobox protein homolog 5, CBX5 (PDB ID: 3FDT).

2.3 The Cation-π Interaction

The non-covalent cation-π interaction arises from the attractive force between a positively charged cation and the negatively charged face of a π-system.[11] The $C^{\delta-}$–$H^{\delta+}$ dipoles of aromatics such as benzene hybridize, forming an area of negative electrostatic potential (ESP) parallel to the face of the ring.[10] Positively charged cations are drawn to the hybridized π-cloud by electrostatic attraction. For cations like tetramethylammonium, cation-

π interactions occur between the methyl groups and the π-system. Although the formal charge is placed on the nitrogen atom, carbon is still more electropositive than nitrogen and therefore the positive charge is located on the methyl groups (Figure 2.6). The cation-π interaction decreases with increasing cation size. As the positive charge is dispersed over a larger volume, the interaction between the cation and π-system weakens. Cation-π interactions are not affected by polarizability, as cyclohexane binds cations with a significantly weaker affinity than benzene, despite being more polarizable.

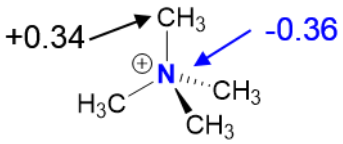


Figure 2.6 Charge distributions in the tetramethylammonium cation.[10]

Cation-π interactions also vary based on the characteristics of the π-system.[10,12,13] Electron-withdrawing substituents on the ring (-CF_3, -CN, NO_2) reduce the cation affinity, while electron donating substituents (-NH_2, -CH_3) increase affinity. These changes to the π-system can be easily visualized in the form of ESP maps (Figure 2.7). Substituent effects on cation-π binding are normally rationalized using electrostatic models, although numerous factors can contribute to binding.[10,12,13] Cation-π binding affinities do not correlate well based on resonance effects of the substituent.[13] Binding affinities do, however, roughly correlate with Hammet's σ_{meta} values, suggesting inductive effects are important in cation-π interactions.[12,13] Although substituent effects through polarization have been proposed, recent studies by Wheeler and Houk have suggested that the polarization model is inaccurate and substituent effects are due to direct through-space interactions of the cation with substituents.[12]

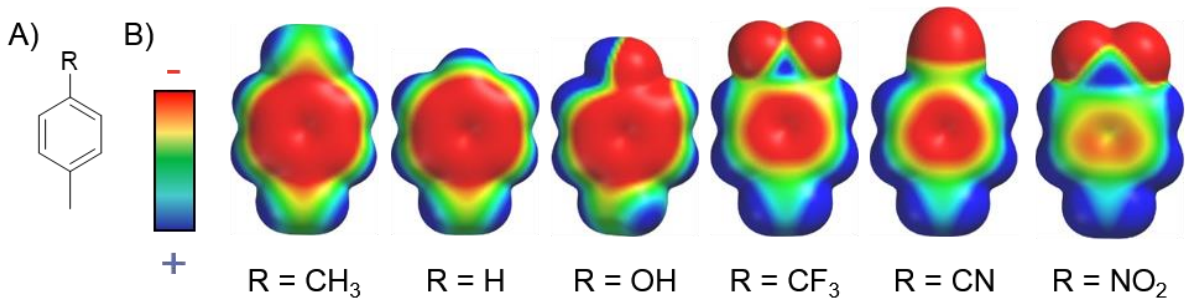


Figure 2.7 Substituent effects on π-system of toluene. A) General structure of substituted toluene derivatives. B) ESP maps of corresponding substituted toluene derivatives. Red signifies increased negative ESP, whereas blue signifies more positive ESP, as denoted. ESP maps provided courtesy of Dr. Marcey Waters.

Cation-π interactions are relatively new discoveries in both biological and physical organic chemistry.[10] Cation-π interactions were initially discovered in 1981 by Kebarle after gas phase experiments revealed that potassium ions preferred benzene over water (-19.2 kcal/mol vs. -17.9 kcal/mol).[14] Cation-π interactions have been implicated as major contributors to molecular recognition and they remain energetically significant even in biological conditions.[10] One-third of homodimers and half of protein complexes contain at least one cation-π interaction, which can contribute 2-5 kcal/mol to binding energies.[10,15] Since their initial discovery, both protein and physical organic chemists and have been working to better understand and predict the energetic contribution of this interaction.

2.4 Cation-π Interactions in Proteins

In the past, unnatural amino acids (UAAs) have been employed for mechanistic studies of cation-π interactions. First introduced by Dougherty and coworkers, many of these investigations used ligand-gated ion channels.[16–18] These proteins also contain an aromatic cage that binds cationic ligands such as acetylcholine and nicotine (Figure 2.8A and B). In these studies, aromatic residues in the cage were replaced with UAAs with increasing

fluorine substitution, thereby decreasing the electrostatic potential of the aromatic ring (Figure 2.8C and D). Residues involved in cation-π interactions showed decreased receptor activity with increasing fluorination.[18] A linear free energy relationship (LFER) between the log of receptor activity and calculated cation-π interaction strength was observed. In a number of cases, Dougherty found a single aromatic residue in the binding pocket exhibited a LFER with acetylcholine binding, which was unexpected given that multiple aromatic residues were available in the binding pocket.[17,19] These results imply that for aromatic residues, geometry with respect to the cation, distance with respect to the cation, and other residue-dependent contexts will control which residues interact with a cation and how strongly they interact with a cation. However, due to lack of structural information on the wild type or UAA-mutant ion channels, no further information could be gained on the single LFER residue. In other words, while these studies unequivocally show the presence and importance of cation- π effects in protein-ligand interactions, they are limited in terms of the ability to provide a complete structural understanding of these interactions.

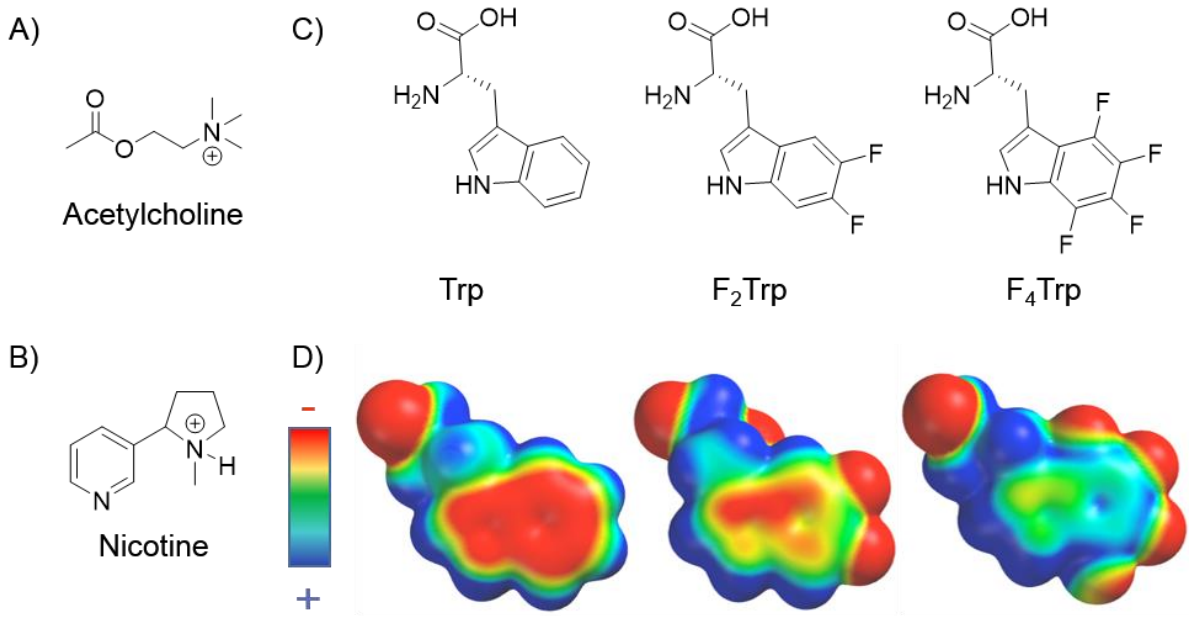


Figure 2.8 Structures of interest for Dougherty's study of ligand-gated ion channels. Structure of receptor ligands, A) acetylcholine and B) nicotine. C) Structure of fluorinated UAAs and D) corresponding ESP maps. Red signifies increased negative ESP, whereas blue signifies more positive ESP, as denoted. ESP maps provided courtesy of Dr. Marcey Waters.

In addition, although Dougherty and coworkers pioneered the study of cation-π interactions in proteins, their approach is not amenable to all cation-π systems. UAAs were site specifically incorporated into suppressor tRNAs that decode an engineered TAG stop codon.[20,21] The UAAs were covalently attached to the tRNA using chemical synthesis and then introduced to the cell using microinjection or electroporation. Since the aminoacylated tRNA acted as the limiting (stoichiometric) reagent in UAA-protein production, UAA-protein yields were low.[16] Conveniently, Dougherty's ion channels required little protein for experiments due to the sensitivity of the analytical technique used to infer ligand binding (e.g. single channel or single cell patch clamp techniques).[17,22] However, the limited protein yields of this approach have excluded many other cation-π systems from being studied using Dougherty's approach and have precluded cation-π investigations of ligand-binding using

more direct procedures for binding measurements including isothermal titration calorimetry (ITC), surface plasmon resonance (SPR), or fluorescence anisotropy.

2.5 Heterochromatin Protein 1 (HP1)

For our investigation of cation-π binding in reader proteins, we chose Heterochromatin Protein 1 (HP1) of *Drosophila melanogaster*, a well-characterized reader protein (Figure 2.9). HP1 recognizes trimethyllysine (Kme3) PTMs of the H3 histone tail at the K9 position (traditionally abbreviated as H3K9me3). The interaction between H3K9me3 and HP1 contributes to heterochromatin assembly and gene silencing.[23] HP1's binding pocket contains a traditional aromatic cage comprised of one tryptophan and two tyrosine residues. When these aromatic cage residues are mutated to alanine, HP1 binds Kme3 with significantly lower affinity.[12] Binding between HP1 and a neutral tert-butyl isostere of Kme3 also showed a reduced affinity of more than 2 kcal/mol, thus indicating that the cation is important for HP1's recognition of Kme3.

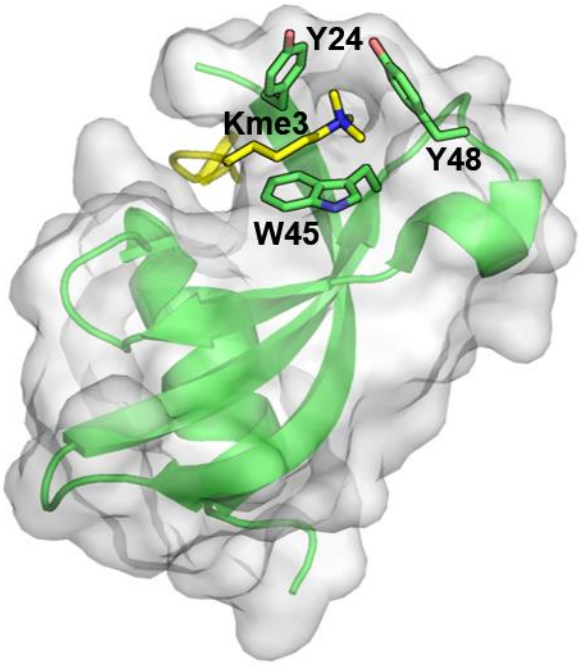


Figure 2.9 Structure of wild type HP1 (PDB ID: 1KNE).[24] Residues in the aromatic cage are shown as sticks. H3K9me3 peptide shown in yellow.

2.6 *In Vivo* Unnatural Amino Acid Mutagenesis in HP1

For this study, we chose to expand upon Dougherty's approach to studying cation-π interactions using the orthogonal tRNA/tRNA synthetase pairs first introduced by Peter Schultz and colleauges.[25–28] With Schultz's method, UAAs can be incorporated selectively *in vivo* with increased yields over the chemically aminoacylated tRNAs used by Dougherty.[29] Although Schultz's approach allows for incorporation of diverse UAAs, fluorinated amino acids are not well incorporated.[30,31] Fluoro-UAAs are similar in size and shape to canonical amino acids and as a result are often accepted by native synthetases, leading to non-specific incorporation and/or cell death.[30,32,33] Selective synthetases for fluorinated tyrosines (F_n-Ys) have been evolved by recognizing the decreased pK_a of the F_n-Ys' phenol groups.[31,34] With increasing fluorination, a larger percentage of the F_n-Ys is anionic at physiological pH. The

introduction of these potentially ionic UAAs would complicate the study of cation-π interactions, thereby disqualifying this family of UAAs for use in cation-π studies.

Fortunately, tyrosine is not the only aromatic amino acid with electron-withdrawing analogs incorporated *in vivo*.[35–38] Many of the UAAs incorporated in *E. coli* using Schultz's method include phenylalanine derivatives bearing electron-withdrawing or –donating groups (EWGs or EDGs, respectively). Many of these UAAs are commercially available and contain EWGs or EDGs commonly used for analysis of substituent group effects via traditional Hammet plot analyses. For the HP1 cation-π system, we chose to incorporate *para*-substituted phenylalanine derivatives. Based on the structure, substitutions at this position were least likely to interfere sterically with the binding of the Kme3 peptide ligand or the global protein structure. We chose a set of *para*-substitutions (CN, CH_3, CF_3, NO_2) that are commonly used to study substituent effects on aromatic compounds via LFER measurements and that have well-studied Hammett values and calculated ESP values. Calculated ESP densities on the rings of these substituted side chains (shown as substituted toluenes) are shown in Figure 2.7.

The *p*-cyanophenylalanine synthetase (*p*CNPheRS) is a particularly promiscuous synthetase that has been shown to exclude tyrosine or phenylalanine yet incorporate various *p*-EWG-substituted phenylalanine derivatives including *p*-cyanophenylalanine (*p*CNF), *p*-nitrophenylalanine (*p*NO_2F), and *p*-chlorophenylalanine (*p*ClF).[38] Our screening has found that this synthetase can also incorporate the EDG-containing *p*-methylphenylalanine *(p*CH_3F*)* and EWG-containing *p*-trifluoromethylphenylalanine *(p*CF_3F*)* (Figure 2.10). These UAAs have been previously incorporated by other synthetases, however now the *p*CNPheRS-system can now allow us to express all of our desired UAA-HP1s with high purity using a

single synthetase.[39] To our knowledge, this is the first demonstration of promiscuous incorporation of *p*CF_3F and *p*CH_3F by *p*CNPheRS.

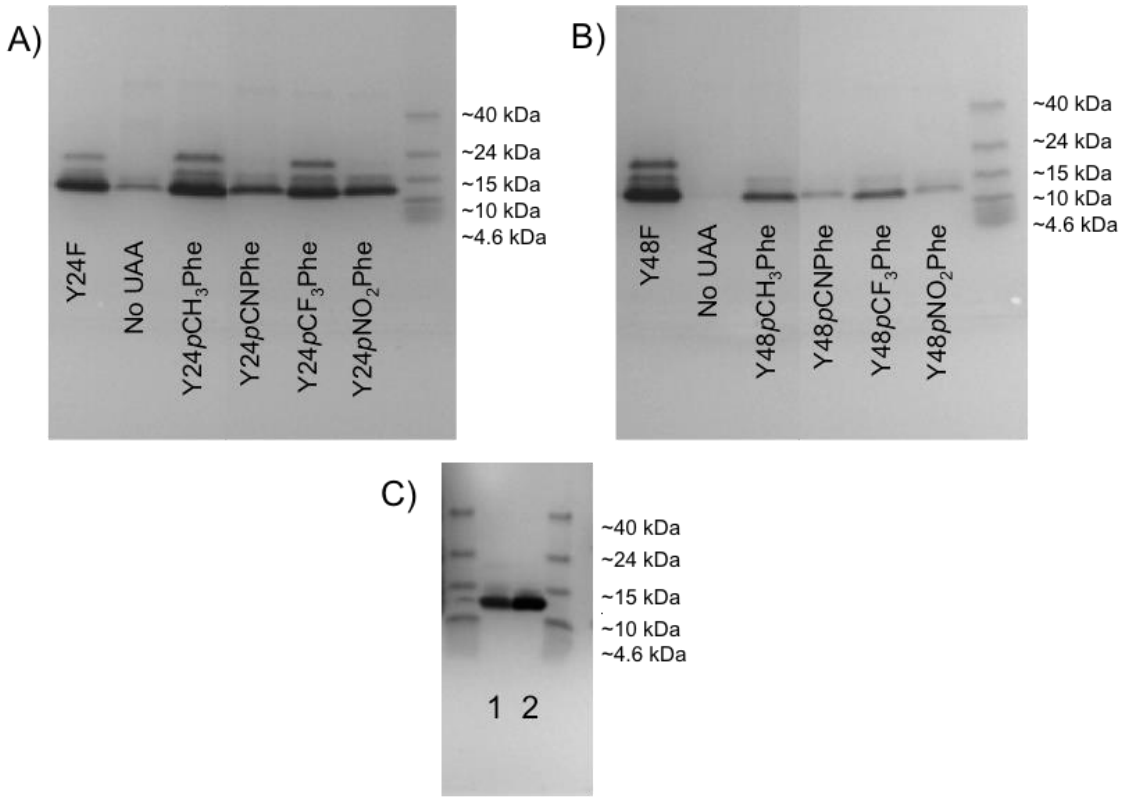


Figure 2.10 SDS-PAGE analysis of purified HP1 mutants. 6XHis-tagged purification of Tyr24 (A) and Tyr48 (B) mutants. C) Impurities present after his-tag purification (1) are removed after size-exclusion chromatography (2).

To incorporate UAAs into HP1, TAG stop codons were cloned into the gene at the tyrosine 24 or 48 positions. These mutant HP1s were cloned into a pET expression vector that was used along with an accessory plasmid containing the orthogonal tRNA/synthetase pair (pUltra-*p*CNPheRS).[40] Expressions in the absence of amino acid yielded little to no HP1; however, addition of 5–20 mM UAA produced high levels of UAA containing protein (Figure 2.10). To further increase yields of UAA-incorporation *in vivo*, multiple vector systems were screened. The system with the best yields was a pET expression vector paired with a pUltra vector, which has been previously shown to increase UAA-incorporation.[40]

HP1 expression was induced using autoinduction media giving yields 5–18 fold over traditional IPTG induction.[41,42] UAA-incorporation was confirmed by ESI-LCMS, and canonical amino acid contamination was not detected (SI Figure 2.1–2.12). Circular dichroism showed UAA-mutations did not affect folding of HP1 (SI Figure 2.14).

2.7 Recognition of Kme3 by UAA-HP1s

Previously, HP1 binding to Kme3 had been studied using fluorescence anisotropy.[23,43–45] However, anisotropy initially required high concentrations for the more weakly-binding UAA-HP1s. Due to low yields and the poor behavior of HP1 at these high concentrations, fluorescence anisotropy was not feasible. Instead, we studied binding using isothermal titration calorimetry (ITC), which has been previously used to measure Kme3 binding to HP1.[6,23,24] In ITC, the Kme3 peptide is slowly titrated into a cell containing the HP1 protein. With each injection, the instrument measures the heat release inside of the cell. The protein sample is titrated with peptide until the binding sites have saturated and no heat release is observed. The heat release per mole of peptide is then plotted against the molar ratio of peptide to protein to create a binding curve. ITC provides a wealth of information including direct measurement of ΔH and K_d (and therefore ΔG) and calculation of ΔS. The binding constants between each HP1-UAA variant and an H3K9me3 peptide (amino acids 1–15) were measured using ITC. Data from HP1 ITC experiments can be found in Table 2.1. Y24*p*CNF was poorly behaved at the high concentrations necessary for ITC, and thus we were unable to obtain reliable binding data for this mutant.

Table 2.1 Binding constants for HP1 mutants measured by ITC

Protein	Cation-π Energy[a] (kcal/mol)	K_d (μM)[b]	ΔG (kcal/mol)
Wild type	26.6	14.4 ± 1.9	-6.6 ± 0.1
Y24*p*CH$_3$Phe	28.3	19.5 ± 1.3[c]	-6.4 ± 0.1
Y24F	26.9	19.0 ± 0.6	-6.4 ±0.1[d]
Y24*p*CF$_3$Phe	19.4	51.8 ± 5.2	-5.9 ± 0.1[d]
Y24*p*CNPhe	16.0	*n.d.*	*n.d.*
Y24*p*NO$_2$Phe	14.0	91.7 ± 0.1[c]	-5.5 ± 0.1[d]
Y48*p*CH$_3$Phe	28.3	16.7 ± 3.0	-6.5 ± 0.1
Y48F	26.9	15.8 ± 2.2	-6.5 ± 0.1
Y48*p*CF$_3$Phe	19.4	24.0 ± 0.8	-6.3 ± 0.1[d]
Y48*p*CNPhe	16.0	44.2 ± 1.7	-5.9 ± 0.1[d]
Y48*p*NO$_2$Phe	14.0	44.9 ± 14.3[c]	-5.9 ± 0.2

[a]Values taken from Wheeler et al.[12] [b]Values are an average of 3 runs unless otherwise noted. Errors are calculated from standard deviation. [c]Average of 2 runs. [d]Errors are calculated from error in fit given by Origin software. *n.d.* = not determined

When the free energy of binding (ΔG_b) is plotted against calculated cation-π binding energies based on gas phase interactions of substituted benzene with Na^+,[12] a LFER is observed (Figure 2.11), revealing the presence of a tunable cation-π interaction at each position. High correlations were also observed when plotted against other methods that have been used to calculate cation-π energies (Figure 2.12–2.14). Intriguingly, comparison of the relationship between ΔG_b and calculated cation-π binding energies suggests a ~1.6 fold difference in magnitude of the effect when comparing the two tyrosine positions; Tyr24 participates in a stronger cation-π interaction with Kme3 than Tyr48.

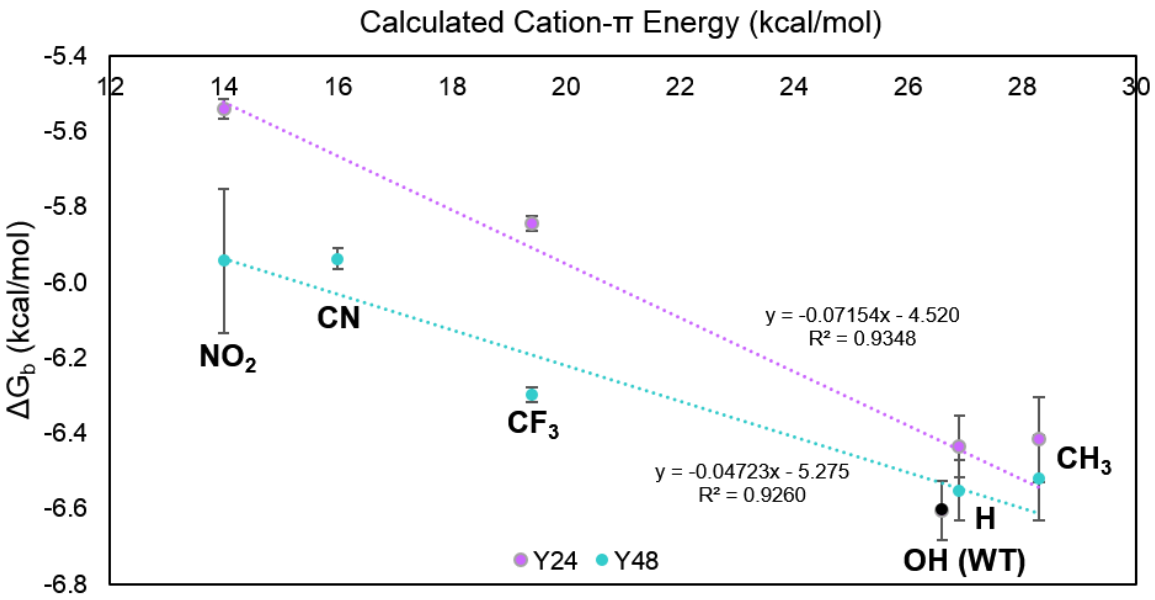


Figure 2.11 Relationship between ΔG_b of HP1 variants/Kme3 and calculated gas-phase cation-π binding energies between C_6H_5R and Na^+.[12] Y24 variants are shown in purple and Y48 variants are shown in teal. Both data sets share the wild type point (WT), shown in black. ΔG_b values were determined at 25°C.

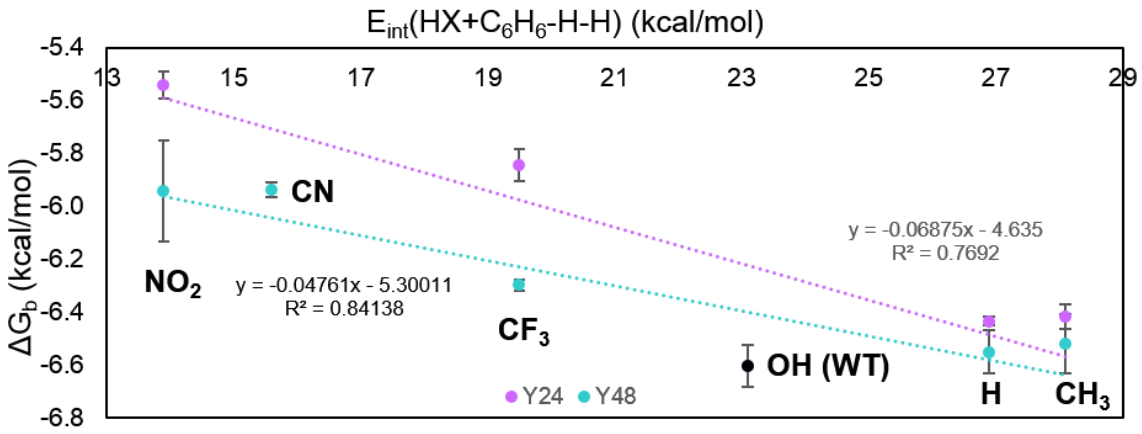


Figure 2.12 Relationship between ΔG_b of HP1 variants/Kme3 and sum of through space interaction of substituent (HX) plus benzene. Y24 variants are shown in purple and Y48 variants are shown in teal. Both data sets share the wild type point (WT), shown in black. ΔG_b values were determined at 25°C.

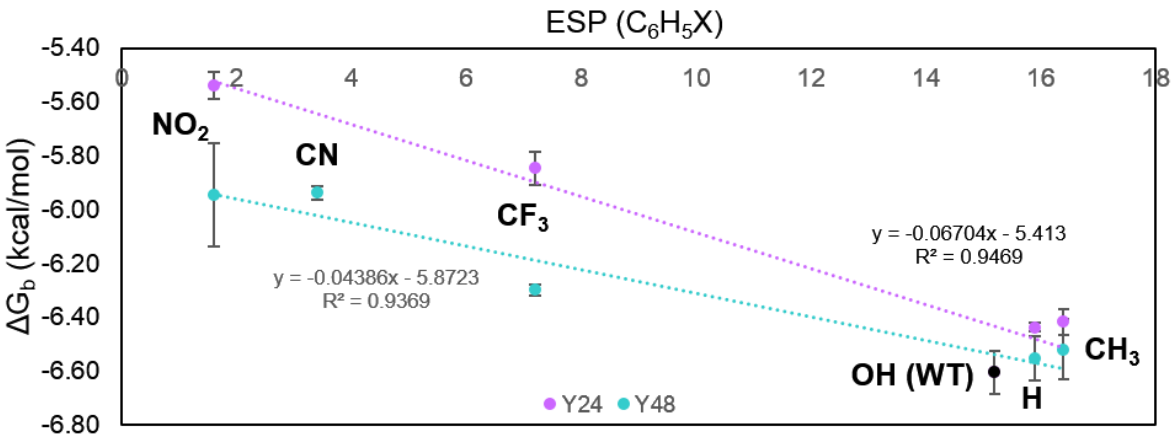


Figure 2.13 Relationship between ΔG_b of HP1 variants/Kme3 and electrostatic potential (ESP). Y24 variants are shown in purple and Y48 variants are shown in teal. Both data sets share the wild type point (WT), shown in black. ΔG_b values were determined at 25°C.

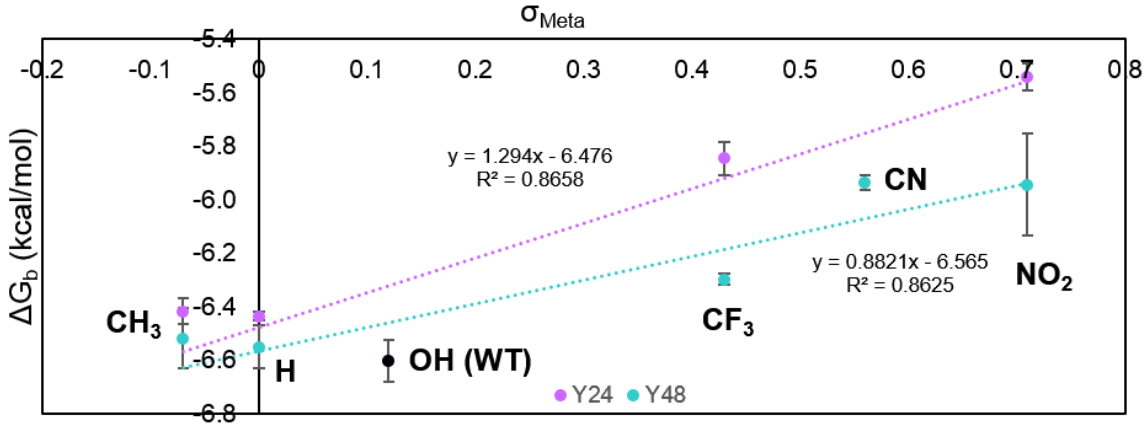


Figure 2.14 Relationship between ΔG_b of HP1 variants/Kme3 and sigma meta (σ_{meta}). Y24 variants are shown in purple and Y48 variants are shown in teal. Both data sets share the wild type point (WT), shown in black. ΔG_b values were determined at 25°C.

Substituents to the phenyl ring may alter other physical properties that contribute to binding, such as polarizability or hydrophobicity. Substituents with increased polarizability can better contribute to van der Waals (VDW) interactions. If VDW interactions contributed significantly to Kme3 recognition, binding affinity would increase with polarizability and a positive correlation would be expected. However, a weak negative correlation was observed between binding affinity and polarizability (Figure 2.15A). Log P is a parameter used to

measure hydrophobicity based on partition coefficients between octanol and water. Higher log P values correspond to higher hydrophobicity (as more compound has partitioned into the octanol layer), and therefore a positive correlation would be expected if the hydrophobic effect contributed to Kme3 binding. However, no significant correlation between ΔG_b and log P was observed (Figure 2.15B). These results indicate that VDW interactions and the hydrophobic effect are not strong drivers of the observed substituent effects and also support our observation that changes in binding affinity result from modification of the cation-π interaction.

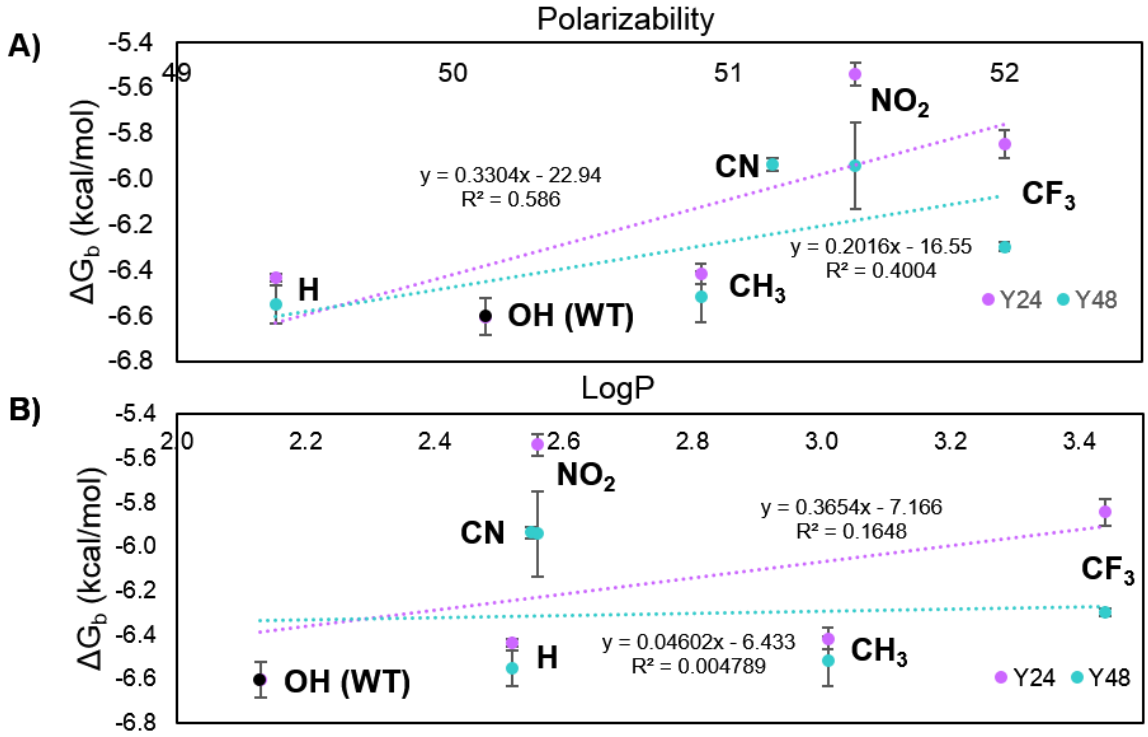


Figure 2.15 Relationship between ΔG_b of HP1 variants/Kme3 and other physical properties effected by substituent.[28] A) Relationship between ΔG_b and polarizability. B) Relationship between ΔG_b and hydrophobicity parameter (log P). Y24 variants are shown in purple and Y48 variants are shown in teal. For all plots, both data sets share the wild type point (WT), shown in black. ΔG_b values were determined at 25°C.

2.8 Structural Investigation of HP1 Mutants

To ensure that changes in binding were not due to structure, X-ray crystal structures of two HP1 variants were determined. Since the *p*NO_2F and phenylalanine have the largest and smallest *p*-substituents (NO_2 and H, respectively), these two structures could show the extremes of structural effects on the aromatic cage. *p*NO_2F and phenylalanine also represent opposite ends of the ESP spectrum; their structures could show any ESP-induced differences in structure. Since the Y24-position shows more pronounced effect on binding than the Y48 position, the Y24F and Y24*p*NO_2F mutants were crystallized. Structures for both variants were determined at 1.52 and 1.28 Å resolution, respectively, using the wild type HP1 crystal structure as a model for molecular replacement. Changes in binding affinity do not appear to be the result of changes in protein structure as wild type, Y24F, and Y24pNO_2F crystal structures overlay with an RMSD of less than 0.3 Å (Figure 2.16), and the distances between Kme3 atoms and each phenyl ring do not significantly change (Figure 2.17).

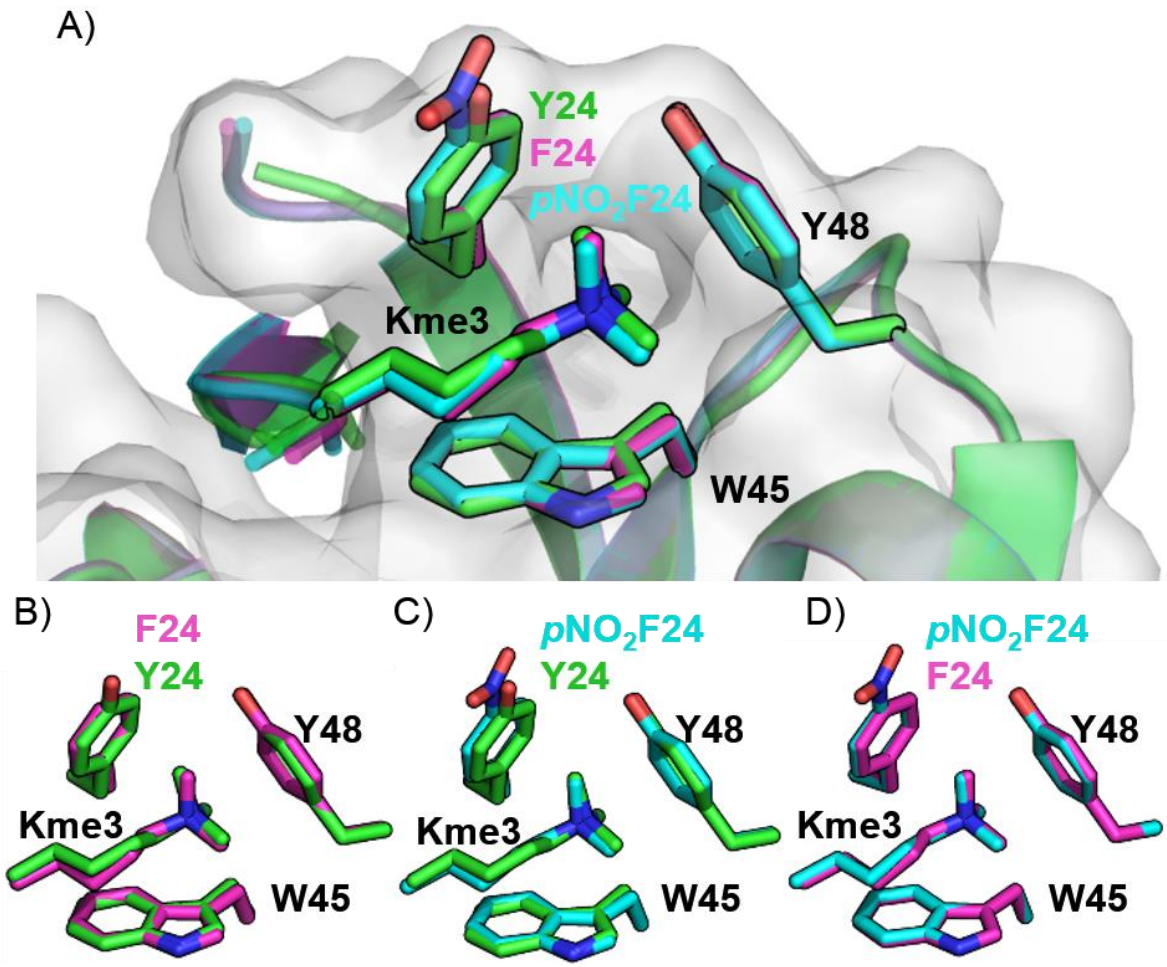
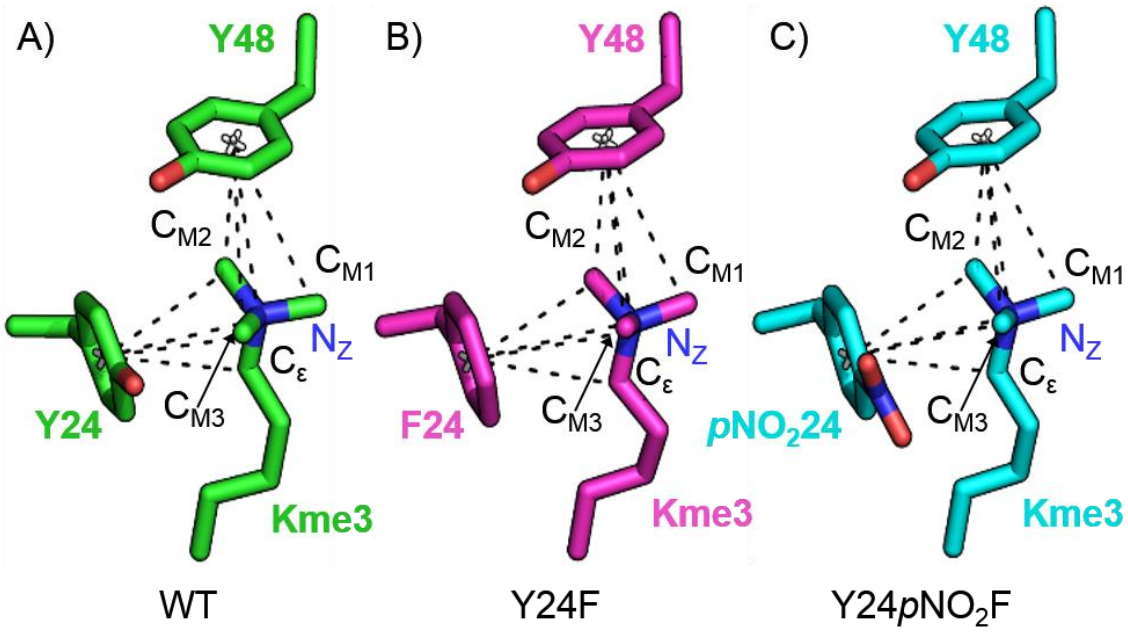


Figure 2.16 Overlays of the aromatic cage of various HP1 mutants. A) Overlay of wild type (green), Y24F (magenta), and Y24*p*NO_2F (cyan) shows minimal perturbation of the aromatic cage. Wild type surface shown for orientation. B) Wild type and Y24F cage overlay, RMS = 0.222 Å, C) Wild type and Y24*p*NO_2F cage overlay, RMS = 0.211 Å, D) Y24F and Y24*p*NO_2F cage overlay, RMS = 0.058 Å.



D)

Mutant	Distance to 24 Position (Å)				Distance to 48 Position (Å)				
	C_{M2}	C_{M3}	C_ε	N_Z	C_{M1}	C_{M2}	C_{M3}	C_ε	N_Z
WT	4.0	3.6	4.1	4.2	4.5	3.3	4.6	5.5	4.2
Y24F	3.9	4.0	4.1	4.3	4.5	3.5	4.5	5.8	4.4
Y24pNO_2F	4.1	4.0	4.2	4.4	4.4	3.3	4.3	5.6	4.2

Figure 2.17 Cation-π distances between Kme3 and 24- and 48- position substituents. A) HP1 wild type (PDB 1KNE), B), HP1 Y24F (PDB ID: 6ASZ), and C) HP1 Y24*p*NO_2F (PDB ID: 6AT0). D) Measured cation-π distances between Kme3 and center of the aromatic ring at positions 24 or 48.

2.9 Computational Investigation into HP1 Kme3 Binding

Inspection of the X-ray structures provides a qualitative assessment of the differences in measured binding affinity upon UAA mutagenesis at the 24- and 48- positions (Figure 2.18). Two methyl groups and the methylene of Kme3 make van der Waals contact with Tyr24 (< 4.5 Å, Figure 2.18 A and B), whereas only a single methyl group makes close contact with Tyr48 (Figure 2.18 C). Computational studies by Dougherty have predicted that

a 3-point contact of tetramethylammonium with benzene is about 1.67-fold stronger when compared to a single methyl group in gas-phase calculations.[46] These calculations are consistent with our data in Figure 2.11 in which the slope for the Y24 position is 1.6-fold greater than the slope of the Y48 position.

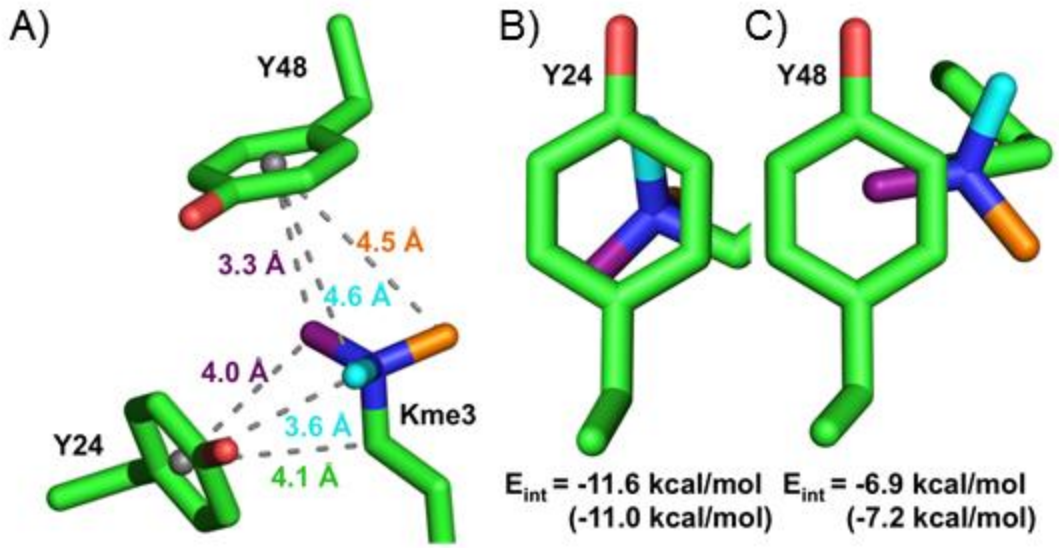


Figure 2.18 Interaction differences for the Y24 and Y48 residue of HP1. A) Measured distances between the center of Y24 and Y48 with respect to atoms on Kme3 (PDB: 1KNE). Distances for Y24*p*NO_2F and Y24F are provided in Figure 2.17. B and C) Contact surface of Kme3 with Tyr24 (B) and Tyr48 (C) viewed normal to the plane of the ring. Kme3 atoms are colored as in panel A. Interaction energies (E_{int}) for Kme3 and each tyrosine are shown and were calculated by M06/6-31G(d,p) or (M06/6-311+G(d,p)).

To further support this observation, interaction energies (E_{int}) between Kme3 and Tyr24 or Tyr48 were calculated using geometries obtained from the wild type Kme3-bound HP1 crystal structure (1KNE) by our collaborators in Kendall Houk's lab at the University of California, Los Angeles. E_{int} values were calculated at the M06/6-31G(d,p) level of theory, which has previously been shown to predict cation-π strength in good agreement with experimental gas-phase measurements for tetramethylammonium interactions with benzene.[46] Calculated E_{int} values predict a stronger interaction with Kme3 for Tyr24

(E_{int} = -11.6 kcal/mol) when compared to Tyr48 (E_{int} = -6.9 kcal/mol; Figure 2.18). Furthermore, calculations at the M06 level performed using the larger 6-311+G(d,p) basis set provided similar results (Tyr24-E_{int} = -11.0 kcal/mol; Tyr48-E_{int} = -7.2 kcal/mol). Our experimental results are consistent with both levels of theory, which predict that the magnitude of the substituent effect differs by a factor of 1.5 and 1.7, respectively.

2.10 HP1 Recognition of Kme2

HP1 also binds dimethyllysine ligands (Kme2), albeit with slightly lower affinity.[45] In the crystal structure of wild type HP1 bound to H3K9me2, the Kme2 ligand shifts slightly away from the Y24 residue due to a salt bridge between dimethyllysine and the Glu52 residue of the binding pocket (Figure 2.19).[43] Based on our initial findings with Kme3, we would expect Kme2 binding to still exhibit a preference for the Y24 position, but not as strongly as observed for the Kme3 ligand. Although initial results have been consistent with our hypothesis, this avenue of study does not seem feasible due to the behavior of the ligand when compared to Kme3 peptides. The H3K9me2 ligand is significantly harder to synthesize and purify than its trimethyl analog, making ligand production the rate-limiting step in the progress of this project. Instead, we will focus future endeavors on probing cages of other methyllysine reader proteins of therapeutic interest.

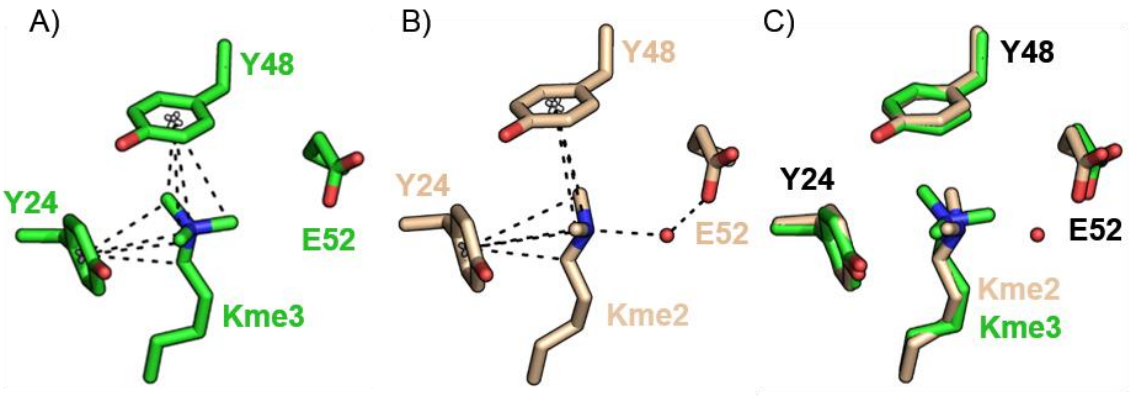


Figure 2.19 Structure of wild type HP1 bound to Kme3 and Kme2 ligands.[24] A) Structure of wild type HP1 and Kme3 ligand (PDB ID: 1KNE). B) Structure of wild type HP1 and Kme3 ligand (PDB ID: 1KNA). C) Overlay of Kme3 and Kme2 shows slight shift of ligand inside the binding pocket.

2.11 Discussion

In summary, we have developed a method for detailed mechanistic and structural investigations of cation-π interactions in proteins, which we have applied here to a methyllysine reader protein. This work provides a rare direct measurement of the electronic tunability of discrete cation-π binding interactions in aqueous solutions.[47,48] Interestingly, while our data demonstrate that both Tyr24 and Tyr48 of HP1 contribute to Kme3 binding via a cation-π interaction, our combined experimental and computational results indicate that these positions do not participate to the same degree, with Y24 exhibiting a greater influence on Kme3 binding. ITC binding analyses and X-ray crystal structures provide the first experimental data demonstrating that the distance and degree of contact influence the magnitude of the cation-π interaction, as had been predicted computationally.[46] Few examples exist of different magnitudes of cation-π interactions within the same binding pocket, and these studies lack structural insight into the molecular basis of such differences.

As computational modeling has become a tool for more efficient drug design, this work also highlights the importance of accurately modeling cation-π interactions in therapeutic targets. The combined binding and structural information from this work provides an experimental benchmark for validating computational methods. Furthermore, as many methyllysine reader proteins share an aromatic cage motif in their binding pocket,[20,23] this work suggests that differences in degree of contacts among reader proteins may be exploited to enhance selective inhibition. By understanding how a protein recognizes its natural substrate, we provide a new framework for the study and design of probes with the necessary affinity and selectivity for therapeutic use.

2.12 Experimental

2.12.1 Cloning, DNA Sequences, and Protein Sequences

pULTRA-*p*CNPheRS[38] was obtained from the lab of Dr. Peter Schultz and is also available from addgene (Plasmid # 48215). HP1 was cloned into a pET11a vector using NdeI and BamHI restriction sites. Mutations to the HP1 gene were generated using standard overlap PCR. Oligonucleotides for PCR were obtained from Integrated DNA Technologies and enzymes and reagents used for cloning were obtained from New England BioLabs Inc. DNA sequences of cloned HP1 mutants from NdeI and BamHI restriction sites are shown below. The underlined portion of the sequence is the HP1 coding sequence, the 6X His tag is italicized, and the 24 and 48 positions have been bolded for clarity. Mutations to the Tyr24 position are shown in red and mutations to the Tyr48 position are shown in blue.

HP1 wild type DNA sequence:

CAT ATG AAA AAA *CAC CAC CAC CAC CAC CAC* GCC GAA GAG GAG GAG GAG GAG **TAC** GCC GTG GAA AAG ATC ATC GAC AGG CGG GTG CGC AAG GGA ATG GTG GAG TAC TAT CTG AAA TGG AAG GGC **TAT** CCC GAA ACT GAG AAC ACG TGG GAG CCG GAG AAC AAT CTC GAC TGC CAG GAT CTT ATC CAG CAG TAC GAG GCG AGC CGC AAG GAT TAA GGA TCC

HP1 wild type protein sequence:

MKKHHHHHHAEEEEEE**Y**AVEKIIDRRVRKGMVEYYLKWKG**Y**PETENTWEPENNLD CQDLIQQYEASRKD

HP1 Y24F DNA sequence

CAT ATG AAA AAA *CAC CAC CAC CAC CAC CAC* GCC GAA GAG GAG GAG GAG GAG **TTC** GCC GTG GAA AAG ATC ATC GAC AGG CGG GTG CGC AAG GGA ATG GTG GAG TAC TAT CTG AAA TGG AAG GGC **TAT** CCC GAA ACT GAG AAC ACG TGG GAG CCG GAG AAC AAT CTC GAC TGC CAG GAT CTT ATC CAG CAG TAC GAG GCG AGC CGC AAG GAT TAA GGA TCC

HP1 Y24F protein sequence:

MKKHHHHHHAEEEEEE**F**AVEKIIDRRVRKGMVEYYLKWKG**Y**PETENTWEPENNLD CQDLIQQYEASRKD

HP1 Y24TAG DNA sequence

CAT ATG AAA AAA *CAC CAC CAC CAC CAC CAC* GCC GAA GAG GAG GAG GAG GAG **TAG** GCC GTG GAA AAG ATC ATC GAC AGG CGG GTG CGC AAG GGA ATG GTG GAG TAC TAT CTG AAA TGG AAG GGC **TAT** CCC GAA ACT GAG AAC ACG TGG GAG CCG GAG AAC AAT CTC GAC TGC CAG GAT CTT ATC CAG CAG TAC GAG GCG AGC CGC AAG GAT TAA GGA TCC

HP1 Y24TAG protein sequence: (* represents an UAA)

MKKHHHHHHAEEEEEE*****AVEKIIDRRVRKGMVEYYLKWKGYPETENTWEPENNLD CQDLIQQYEASRKD

HP1 Y48F DNA sequence

CAT ATG AAA AAA *CAC CAC CAC CAC CAC CAC* GCC GAA GAG GAG GAG GAG GAG **TAC** GCC GTG GAA AAG ATC ATC GAC AGG CGG GTG CGC AAG GGA ATG GTG GAG TAC TAT CTG AAA TGG AAG GGC **TTT** CCC GAA ACT GAG AAC ACG TGG GAG CCG GAG AAC AAT CTC GAC TGC CAG GAT CTT ATC CAG CAG TAC GAG GCG AGC CGC AAG GAT TAA GGA TCC

HP1 Y48F protein sequence:

MKKHHHHHHAEEEEEEYAVEKIIDRRVRKGMVEYYLKWKG**F**PETENTWEPENNLD CQDLIQQYEASRKD

HP1 Y48TAG DNA sequence:

CAT ATG AAA AAA *CAC CAC CAC CAC CAC CAC* GCC GAA GAG GAG GAG GAG GAG **TAC** GCC GTG GAA AAG ATC ATC GAC AGG CGG GTG CGC AAG GGA ATG GTG GAG TAC TAT CTG AAA TGG AAG GGC **TAG** CCC GAA ACT GAG AAC ACG TGG GAG CCG GAG AAC AAT CTC GAC TGC CAG GAT CTT ATC CAG CAG TAC GAG GCG AGC CGC AAG GAT TAA GGA TCC

HP1 Y48TAG protein sequence: (* represents an UAA)

MKKHHHHHHAEEEEEEYAVEKIIDRRVRKGMVEYYLKWKG*PETENTWEPENNLD CQDLIQQYEASRKD

2.12.2 Protein Expression and Optimization

For UAA-HP1 variants, pET11a-HP1-Y24TAG or -48TAG was co-transformed with pUltra-*p*CNPheRS into Bl21-Gold(DE3) competent cells (Agilent Technologies). For HP1 wild type, Y24F and Y48F, pET11a-HP1, -Y24F or -Y48F were transformed into BL21-Gold(DE3) competent cells. Cells were rescued with 1 mL SOC broth and then incubated for 45 min at 37°C with shaking. 50 uL of each rescue was plated as follows: wild type/Y24F/Y48F on LB ampicillin (100 mg/L) agar plates; Y24TAG/Y48TAG co-transformed with pUltra-*p*CNPheRS on LB ampicillin (100 mg/L) and streptomycin (50 mg/L) agar plates. Plates were incubated overnight at 37°C. Single colonies from the transformation plates were used to inoculate LB with appropriate antibiotic in baffled flasks (flask volume <4X larger than LB volume). Cultures were grown to saturation overnight at 37°C with shaking at 225 RPM.

Initially, proteins were expressed in using LB media and traditional IPTG induction as reported previously.[24] However, UAA-HP1 yields were low and UAA was precious, and therefore expression conditions were optimized. First, expressions were optimized based on UAA concentration. Traditionally, media is supplemented to give a final UAA concentration of 1 mM. However, based on screening, 5 mM of the *p*-substituted phenylalanine derivatives gave the best yields with diminishing returns at >5mM. *p*NO_2F required an even higher concentration of 20 mM, presumably due to the decreased affinity of the UAA for the synthetase because of the altered phenyl ring. Optimized expression length was found to be 48 hours, but new expression conditions still gave very poor yields for *p*CNF and *p*NO_2F proteins.

Since yields could not be increased using simple changes to the expression protocol, other expression systems were explored. Since UAA-protein yield is largely dependent on the vector of the orthogonal pair, pET/pUltra system was compared to pBad/pDule and pBk/pEvol expression systems. The pBad/pDule expression system has been previously found amenable to autoinduction, which can increase yields up to 10-fold.[41,42] After testing each system with traditional induction or autoinduction, the system with the best yields was the pET/pUltra system, which had been previously shown to increase UAA-incorporation.[40] Screening various media supplements, flask and seal types, and shake speeds helped increase HP1 yields to 5–18 fold over the initial expression conditions.

After optimization, all proteins were expressed in 2.5 L Ultra Yield Flasks™ (Thompson Instrument Company) containing 500 mL of ZYP-5052 autoinduction media[41,42] supplemented with 5 mM $MgCl_2$, 5 mM $MgSO_4$, and 1:5000 dilution of Antifoam 204 to increase oxygen uptake and prevent foaming over. Each flask also contained appropriate

antibiotics (100 mg/L ampicillin (pET-HP1s), 50 mg/L streptomycin (pUltra-*p*CNPheRS)). For wild type, Y24F, and Y48F expressions, media was inoculated with 2.5 mL of saturated overnight culture. For Y24TAG and Y48TAG expressions, autoinduction media was inoculated with 5 mL of saturated overnight culture to account for the slower initial growth in the presence of two antibiotics. After inoculation, cultures were incubated at 37°C with 310-350 RPM shaking until reaching an OD_{600} between 1–2. Dry UAA (Chem Impex International) was added to the appropriate TAG cultures (2.5 mmol UAA for 5 mM final concentrations. *p*NO_2F was increased to 10 mmol for 20 mM final concentration to compensate for lower affinity of the *p*CNPheRS for *p*NO_2Phe). Incubator temperature was then dropped to 18°C and the cultures were left to express for 24-48 hours. For expressions containing Y24*p*NO_2Phe, the incubator was covered with aluminum foil to prevent light degradation of *p*NO_2Phe.

After expression, cultures were pelleted at 4500 RPM for 10 min and the supernatant was decanted. Cell pellets were frozen overnight at -20°C and resuspended in 20 mL lysis buffer (50 mM Tris, pH 8, 150 mM NaCl, 30 mM imidazole, 0.25 mg/mL lysozyme, 1mM phenylmethanesulfonyl fluoride, with cOmplete EDTA-Free Protease Inhibitor Cocktail Tablets (Roche)). The resuspended pellet was incubated at 37°C with 225 RPM shaking for 30 min and cooled on ice for 10 min. Pellets were sonicated on ice for 7.5 min (20% amplitude, 0.5 seconds on, 0.5 seconds off) until the lysate appeared homogenous. Lysate was clarified by centrifugation (19,000 RPM, Sorvall SS-34 rotor) for 45 min. Supernatant was decanted and filtered through a 0.45 um syringe filter.

2.12.3 Protein Purification

Filtered lysate was purified on an ÄKTAPurifier UPC 10 (GE) equipped with a HisTrap-5mL HP column (GE). HP1 was 6XHis-tag purified using the buffers previously described[24] and eluted using a step gradient from 0–55 % buffer B. Eluted fractions were pooled and concentrated on a 3 kDa Amicon Ultra-15 Centrifugal filter. The concentrated sample was purified by size exclusion chromatography using a Superdex 200 10/300 GL size exclusion column equilibrated in SEC buffer (50 mM sodium phosphate, pH 8, 25 mM NaCl, 2 mM DTT). Eluted fractions (eluted at 15.5–18 mL) were pooled, concentrated, and quantified using a Cary 100 UV/Vis Spectrophotometer (Agilent Technologies). Extinction coefficients for UAA proteins were calculated by measuring the extinction coefficient of each free amino acid in solution and adding the free UAA extinction coefficient to the extinction coefficient of wild type HP1 with one tyrosine removed. The extinction coefficient of wild type HP1 with a tyrosine removed was calculated using the Scripps Protein Calculator (http://protcalc.sourceforge.net/). Extinction coefficients are provided in SI Table 2.1. Protein purity and UAA incorporation was assessed using SDS-PAGE and ESI-LCMS.

2.12.4 ESI-LCMS Confirmation of UAA Incorporation

1 mL of a 10 μM solution of each protein was exchanged into HPLC-grade water using a 3 kDa Amicon Ultra-15 centrifugal filter and then filtered through glass wool. The samples were run on an Agilent 6520 Accurate-Mass Q-TOF ESI positive LCMS (SI Table 2.2) using one of two methods: A) or B). Details of each method can be found in SI Table 2.3 and SI Table 2.4. All LCMS chromatograms show evidence of the appropriate UAA incorporation with no detectible canonical amino acid contamination (SI Table 2.5). LCMS

chromatograms from each HP1 variant can be found in SI Figure 2.1–2.13. Although incorporation of tyrosine or phenylalanine can be detected in TAG mutants expressed in the absence of unnatural amino acid (SI Figure 2.7 and SI Figure 2.13), no evidence of tyrosine or phenylalanine incorporation is detected in the presence of UAA.

2.12.5 Peptide Synthesis

Peptides were synthesized by Dr. Amber Koenig, Mack Krone, and Katherine Albanese of Dr. Marcey Waters' lab. H3K9me3 (ARTKQTARK$(Me)_3$STGGKAY) was synthesized using Fmoc-protected amino acids and Rink Amide AM resin on a 0.5 mmol scale. The amino acid residues were activated with HBTU (O-benzotriazole-N, N, N', N',-tetramethyluronium hexafluorophosphate) and HOBt (N-hydroxybenzotriazole) in the presence DIPEA (diisopropylethylamine) in DMF (N,N-dimethylformamide). 4 equivalents of the amino acid, HBTU, and HOBt were used for each coupling step, along with 8 equivalents of DIPEA. Double couplings of 30 minutes were used for each residue. Deprotections of Fmoc were carried out in 20% piperidine in DMF, twice for 15 minutes each.

Trimethyllysine was generated during the synthesis of the H3 peptide by first coupling Fmoc-Lys$(Me)_2$-OH·HCl for 5 hours with HBTU/HOBt activation. 2 equivalents of dimethyllysine, HBTU, and HOBt were used, along with 4 equivalents of DIPEA. Immediately after coupling, the resin was washed with DMF and the residue was methylated to form trimethyllysine with 7-methyl-1,5,7-triaza-bicyclo[4.4.0]dec-5-ene (MTDB, 1.2 equivalents) and methyl iodide (10 equivalents) in DMF for 6 hours. The resin was washed with DMF and peptide synthesis was continued with aforementioned conditions.

Peptides were cleaved with 95:2.5:2.5 trifluoroacetic acid (TFA):water:triisopropylsilane (TIPS) for 4 hours. The TFA was evaporated and products were precipitated with cold diethyl ether. The resulting peptides were extracted with water and lyophilized. Crude peptide material were purified by reversed phase HPLC using a C-18 semipreparative column and a gradient of 0 to 100% B in 60 minutes, where solvent A was 95:5 water:acetonitrile, 0.1% TFA and solvent B was 95:5 acetonitrile:water, 0.1% TFA. The purified peptides were lyophilized. The peptide was desalted for ITC using a Sephadex G-24 column from GE in water and lyophilized to a powder. Identity was confirmed by MALDI mass spectrometry. Calculated $M+H^+$: 1765.02 Da, Observed: 1765.95 Da.

2.12.6 Circular Dichroism (CD) of HP1 Mutants

CD experiments were performed using an Applied Photophysics Chiroscan Circular Dicroism Spectrophotometer. Spectra were obtained with 30 μM HP1 protein in 10 mM sodium phosphate buffer, pH 7.4 with 2 mM dithiothreitol (DTT) at 20°C. All scans were corrected with buffer subtraction. The mean residue ellipticity was calculated using the equation $\theta = \frac{signal}{10lc}\frac{1}{r}$ where θ is MRE, signal is CD signal, l is path length, c is protein concentration, and r is the number of amino acid residues. Difference in CD spectra are likely due to error in extinction coefficients used to quantitate protein concentration. CD spectra can be found in SI Figure 2.14.

2.12.7 Isothermal Titration Calorimetry (ITC) Binding Measurements

ITC experiments were performed by titrating H3K9me3 peptide (2.5-7.47 mM) into HP1 mutants (160-290 μM) in 50 mM sodium phosphate, pH 7.4, 150 mM NaCl, 2 mM

TCEP at 25°C using a Microcal AutoITC200. Peptide and protein concentrations were determined by measuring absorbance at 280 nm on a Cary 100 UV/Vis Spectrophotometer (Agilent Technologies). Heat of dilution was accounted for by subtracting the endpoint ΔH value from each prior injection. Data was analyzed using the One-Site binding model supplied in the Origin software. While the binding stoichiometry is known to be 1:1, at the high concentrations used here active protein concentration may differ from measured concentration. When ITC experiments were run under low c-value conditions ($c \leq 4$, $c = \frac{[protein]}{K_d}$), the stoichiometry parameter (N) of the non-linear fitting function was fixed to 1.[49,50] ITC binding curves can be found in SI Figure 2.15–2.24.

2.12.8 Data from LFER Plots

All plots were generated using the data found in tables SI Table 2.6 and SI Table 2.7.

2.12.9 Protein Crystallography

HP1 Y24F and Y24*p*NO_2F protein was diluted to a concentration of 10 mg/mL in 10 mM potassium phosphate, pH 7, 2mM TCEP. The diluted protein was then spiked with 8.6 mg/mL H3K9me3 peptide (~70% pure) in a 4:1 peptide:HP1 ratio. Crystals were grown by sitting drop vapor diffusion at 4°C. Cryschem Plates (Hampton Research) were set up on ice by mixing 1 uL of the protein-peptide dilution and 1 uL of reservoir solution. Crystal growth was typically observed within 12–72 hours. Crystals were harvested and flash-frozen in liquid nitrogen with no supplementary cryoprotectant necessary.

Reservoir solution for Y24Phe: 0.1 M MES, pH 6.3; 3.4 M $(NH_4)_2SO_4$

Reservoir solution for Y24*p*NO_2Phe: 0.1 M MES, pH 5.8; 3.0 M $(NH_4)_2SO_4$

2.12.10 X-ray Data Collection and Protein Structure Determination

X-ray diffraction data were collected at Southeast Regional Collaborative Access Team (SER-CAT) at the Advanced Photon Source (Argonne National Laboratory) using beamline 22-ID and a MAR300HS CCD detector. Data were collected at 100 K. Statistics for data collection and refinement are listed in SI Table 2.8. Diffraction data sets were integrated and scaled with the automated data processing software KYLIN provided by SER-CAT.[51] Initial phases were determined by molecular replacement against the wild type HP1 structure (PDB accession code 1KNE)[24] using Phenix Phaser.[52] Refinement was accomplished by iterative cycles of manual model building with *Coot*[53] and automated refinement using Phenix Refine.[52] Model quality was assessed with the Phenix Validation tool. All of the protein structure figures and alignments were generated using PyMOL software (The PyMOL Molecular Graphics System, Version 1.8, Schrödinger LLC.). Data collection and refinement statistics can be found in SI Table 2.8.

2.12.11 Verification of the Y24*p*NO_2F Mutation in Protein Structure

For the Y24*p*NO_2F structure, a phenylalanine was first modeled in at the Y24 residue. After refinement, the mFo-DFc map showed extra electron density near the *para*-position of the phenylalanine ring (SI Figure 2.25 A). When the phenylalanine is mutated to *p*NO_2Phe, the mFo-DFc density fits the UAA's R-group well (SI Figure 2.25 B). Once the Y24*p*NO_2F mutation model is refined, the 2mFo-DFc density fits the UAA well (SI Figure 2.25 C). The 2mFo-DFc density from the Y24F structure also matches the Y24F mutation, but lacks the *para*-electron density of the Y24*p*NO_2F structure (SI Figure 2.25 D).

2.12.12 Determining Changes to HP1 Variant Binding Pockets

The center of each tyrosine residue was modeled in using the pseudoatom command in PyMOL. The distance between each atom in Kme3 was measured using PyMOL's distance command. All calculated cation-π distances can be found in Figure 2.17.

2.12.13 Computational Methods for E_{int} Calculations Between the Wild Type Protein and Trimethyllysine (Kme3)

The structure of the Y24–Y48–Kme3 complex was extracted and truncated from the crystal structure of the wild type protein (PDB: 1KNE). Each terminus of the fragments was capped with a hydrogen atom at 1.09 Å. The cation-π interaction for each of the two tyrosine residues with the lysine ammonium ion was computed by single-point energy calculations at the M06/6-31G(d,p) level of theory.[54] M06/6-31G(d,p) was recently shown to model cation/π interactions well by Dougherty, et al.[46] The interaction energy is defined as the energy difference between the dimer and each amino acid monomers: $E_{int} = E_{dimer} - (E_{Kme3} + E_{Y})$. All quantum chemical calculations were performed using *Gaussian 09*.[55] All graphics on optimized structures were generated with *CYLview*.[56]

2.13 Supplementary Information

2.13.1 Supplementary Tables

SI Table 2.1 Extinction coefficients for UAAs and HP1-UAA variants

Mutant Name	Extinction Coefficients ($cm^{-1}M^{-1}$)	Molecular Weight (Da)
Wild type	17780.0	8569.4
Y24F or Y48F	16500.0	8553.4
Y24*p*CNF or Y48*p*CNF	17169.4	8578.4
Y24*p*NO_2F or Y48*p*NO_2F	24817.3	8598.4
Y24*p*CH_3F or Y48*p*CH_3F	16632.8	8567.5
Y24*p*CF_3F or Y48*p*CF_3Phe	16504.4	8621.4
UAA	**Free UAA Extinction Coefficients**	**Molecular Weight (Da)**
***p*CNPhe**	669.4	190.2
***p*NO_2Phe**	8317.3	210.2
***p*CH_3Phe**	132.8	179.2
***p*CF_3Phe**	4.4	233.2

SI Table 2.2 ESI-LCMS instrument information

Column	Restek Viva C4 5 µm 150 x 2.1 mm
Solvent A	0.1 % formic acid in water
Solvent B	0.1 % formic acid in acetonitrile
Temperature	35°C
Ion Source	Dual ESI
Ion Polarity	Positive
Abs. Threshold	200
Rel threshold (%)	0.01
Cycle Time	1 s
Gas Temp	350 °C
Drying gas	12 l/min
Nebulizer	50 psig
Fragmentor	200 V
Skimmer	65 V
OCT 1 RF VPP	750
Min Mass Range	100 m/z
Max Mass Range	3200 m/z
Acquisition Rate	1 spectra/s
Acquisition time	1000.2 ms/spectrum
Transients/spectrum	9898

SI Table 2.3 ESI-LCMS method information for method A

Method A	
Solvent A	Water
Solvent B	Acetonitrile
Flowrate	0.4 mL/min
Gradient	
Time (min)	%B
0	5
2	5
8	30
22	60
23	60
35	70
40	95
42	95
44	5

SI Table 2.4 ESI-LCMS method information for method B

Method B	
Solvent A	Water
Solvent B	Acetonitrile
Flowrate	0.3 mL/min
Gradient	
Time (min)	%B
0	5
15	95
20	95
20.01	5
25	5

SI Table 2.5 ESI-LCMS data verifies UAA-incorporation

Sample	Expected Mass (Da)	Observed Masses (Da)	Difference (Da)	% Difference
Wild type	8569.30	8569.67	0.29	3.4×10^{-5}
24F	8553.31	8553.78	0.29	3.4×10^{-5}
24*p*CH$_3$Phe	8567.46	8567.11	0.35	4.1×10^{-5}
24*p*CNPhe	8578.43	8578.96	0.55	6.4×10^{-5}
24*p*CF$_3$Phe	8621.42	8622.13	0.71	8.2×10^{-5}
24*p*NO$_2$Phe	8598.39	8598.80	0.41	4.8×10^{-5}
48F	8553.31	8553.85	0.54	6.3×10^{-5}
48*p*CH$_3$Phe	8567.46	8568.10	0.64	7.5×10^{-5}
48*p*CNPhe	8578.43	8579.01	0.58	6.8×10^{-5}
48*p*CF$_3$Phe	8621.42	8622.09	0.67	7.8×10^{-5}
48*p*NO$_2$Phe	8598.39	8598.71	0.32	3.7×10^{-5}

SI Table 2.6 Binding constants from LFER plots

R Group	E_{int} (C_6H_5X), kcal/mol	Polarizability	logP	E_{int} ($HX+C_6H_6-H_2$), kcal/mol	σ_{Meta}	ESP (C_6H_5X), kcal/mol
OH	26.6	50.12	2.13	23.1	0.12	15.2
H	26.9	49.36	2.52	26.9	0.00	15.9
CH_3	28.3	50.9	3.01	28.1	-0.07	16.4
CF_3	19.4	52.0	3.44	19.5	0.43	7.2
CN	16.0	51.16	2.55	15.6	0.56	3.4
NO_2	14.0	51.46	2.56	13.9	0.71	1.6

SI Table 2.7 Binding data from HP1 mutants

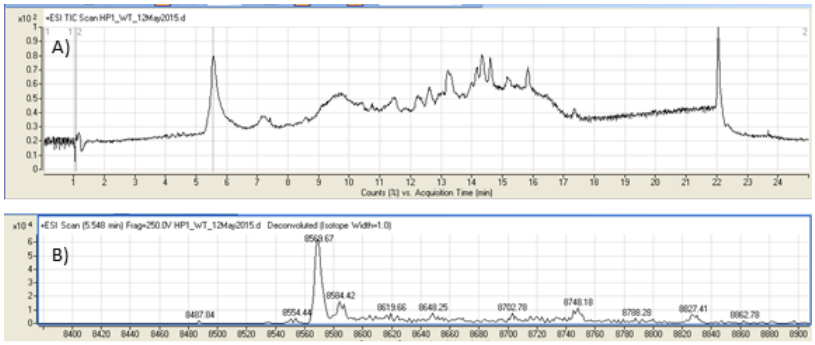
HP1 Mutant	Average ΔG_b, kcal/mol	Standard Deviation, kcal/mol
Wild type	-6.6	0.1
Y24F	-6.4	0.1
Y24*p*MeF	-6.4	0.1
Y24*p*CF_3F	-5.9	0.1
Y24*p*NO_2F	-5.5	0.1
Y48F	-6.5	0.1
Y48*p*MeF	-6.5	0.1
Y48*p*CF_3F	-6.3	0.1
Y48*p*CNF	-5.9	0.1
Y48*p*NO_2F	-5.9	0.2

SI Table 2.8 Data collection and refinement statistics for HP1 mutant crystals

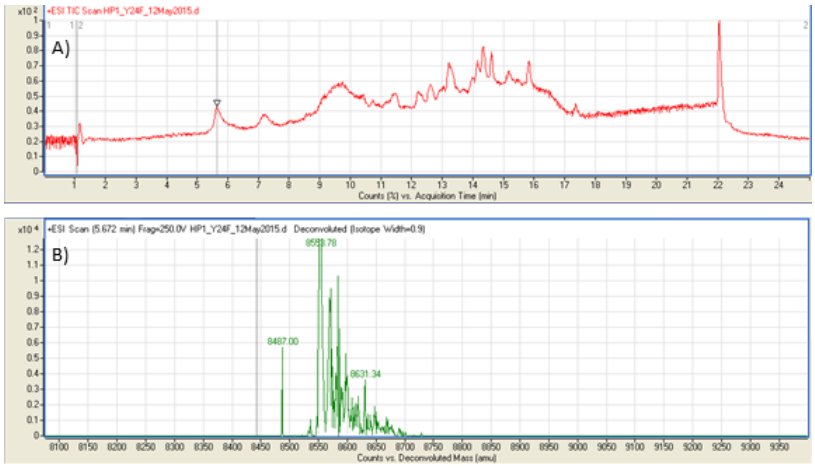
	HP1 Y24F	HP1 Y24*p*NO_2Phe
PDB accession #	6ASZ	6AT0
Data collection		
Space group	C 2 2 21	C 2 2 21
Wavelength	1.000	1.000
Cell dimensions		
a, *b*, *c* (Å)	34.52 76.78 75.51	34.42 76.86 76.48
a, b, g (°)	90	90
Resolution (Å)	10.89 − 1.52 (1.57 − 1.52)*	11.22 − 1.285 (1.33 − 1.285)*
R_{merge}	8.0(47.4)	4.5 (46.86)
I / σI	4.4(1.5)	12.8 (2.56)
Completeness (%)	98.2(99.8)	96.9 (93.9)
Redundancy	5.7(5.2)	5.36(4.14)
Refinement		
Resolution (Å)	1.57 − 1.52	1.33 − 1.285
No. reflections	15582	25417
R_{work} / R_{free}	0.25 / 0.27	0.24 / 0.26
No. atoms		
Protein	448	483
Ligand/ion	49	56
Water	19	40
B-factors		
Protein	26.6	27.9
Ligand/ion	29.5	30.8
Water	29.1	37.5
R.m.s. deviations		
Bond lengths (Å)	0.006	0.004
Bond angles (°)	0.76	0.75
Ramachandran outliers	0%	0 %

*All data sets were collected from single crystals. Highest-resolution shell is shown in parentheses.

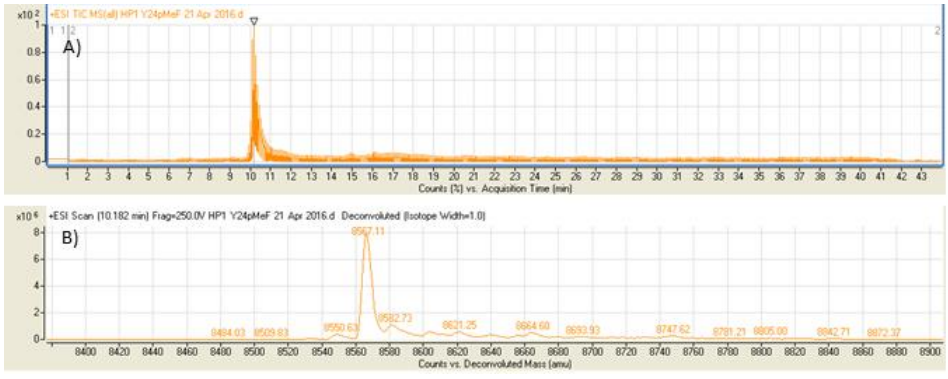
2.13.2 Supplementary Figures



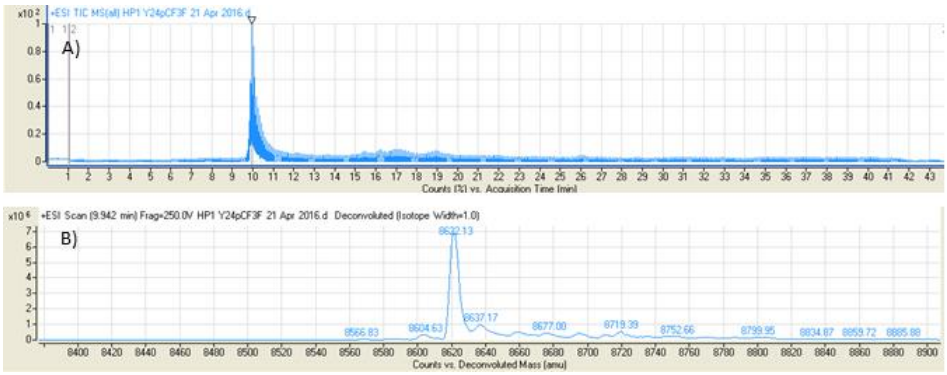
SI Figure 2.1 LCMS of HP1 wild type using method A.
TIC scan (A) and corresponding m/z deconvolution (B).



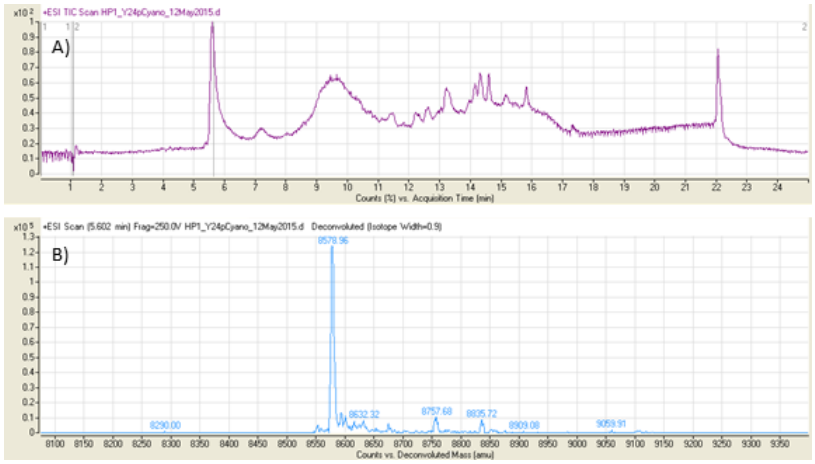
SI Figure 2.2 LCMS of HP1 Y24F using method A.
TIC scan (A) and corresponding m/z deconvolution (B).



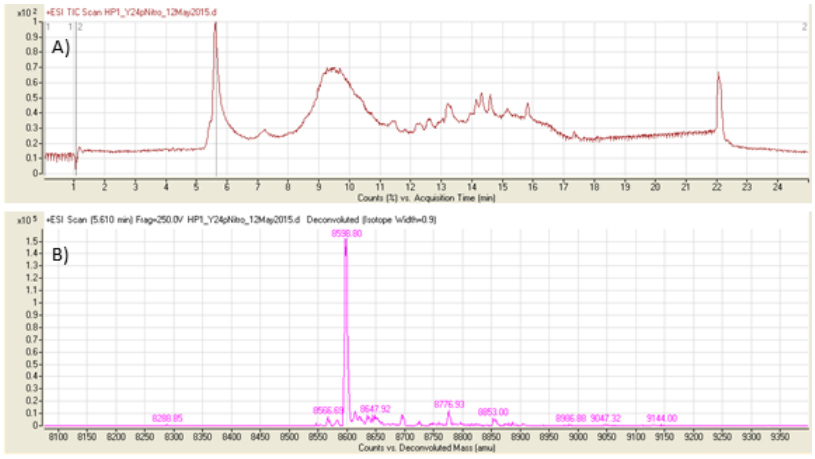
SI Figure 2.3 LCMS of HP1 Y24*p*CH_3F using method B.
TIC scan (A) and corresponding m/z deconvolution (B).



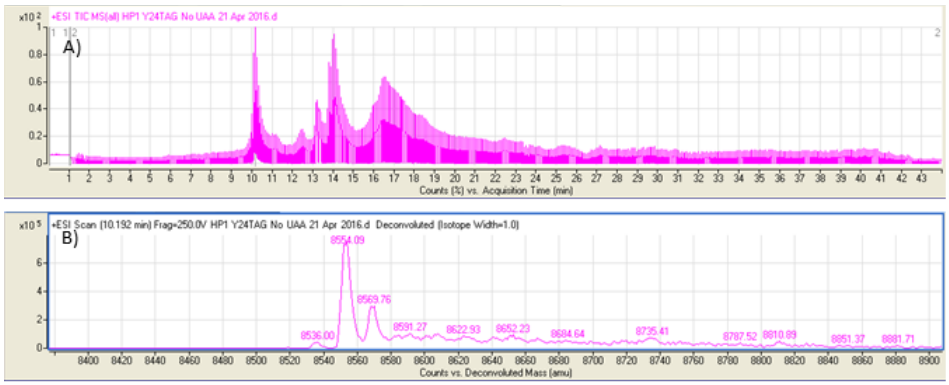
SI Figure 2.4 LCMS of HP1 Y24*p*CF_3F using method B.
TIC scan (A) and corresponding m/z deconvolution (B).



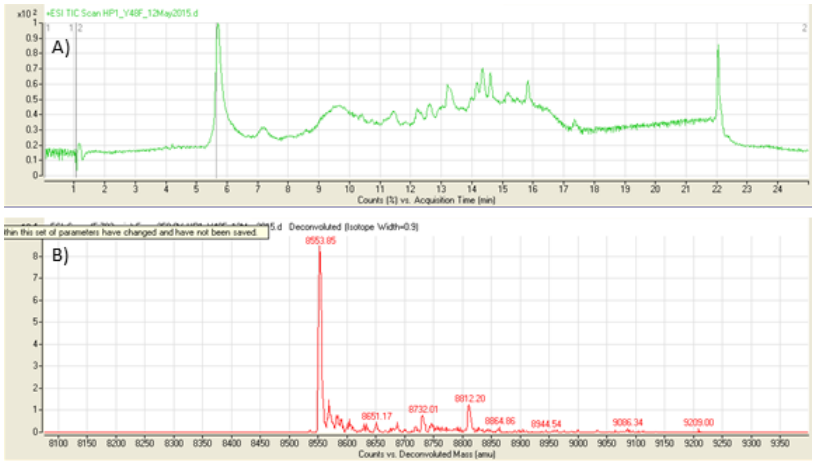
SI Figure 2.5 LCMS of HP1 Y24*p*CNF using method A.
TIC scan (A) and corresponding m/z deconvolution (B).



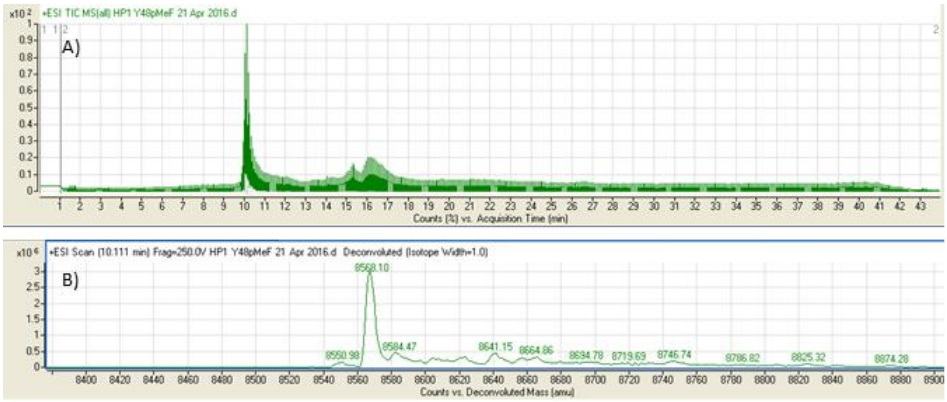
SI Figure 2.6 LCMS of HP1 Y24*p*NO_2F using method A.
TIC scan (A) and corresponding m/z deconvolution (B).



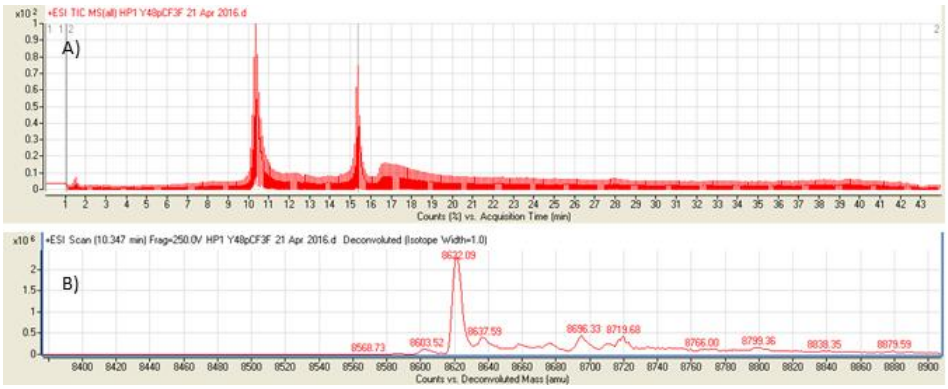
SI Figure 2.7 LCMS of HP1 Y24TAG with no UAA added using method B. TIC scan (A) and corresponding m/z deconvolution (B). Wild type HP1 and Y24F are produced in the absence of UAA, but when UAA is added wild type and Y24F are not detected.



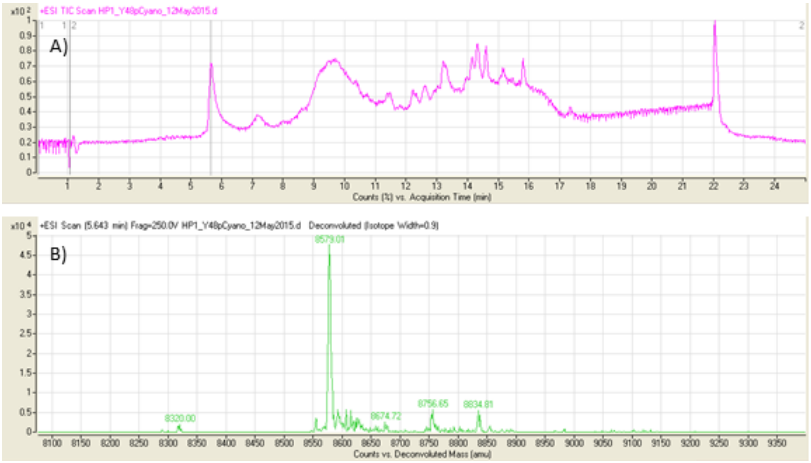
SI Figure 2.8 LCMS of HP1 Y48F using method A. TIC scan (A) and corresponding m/z deconvolution (B).



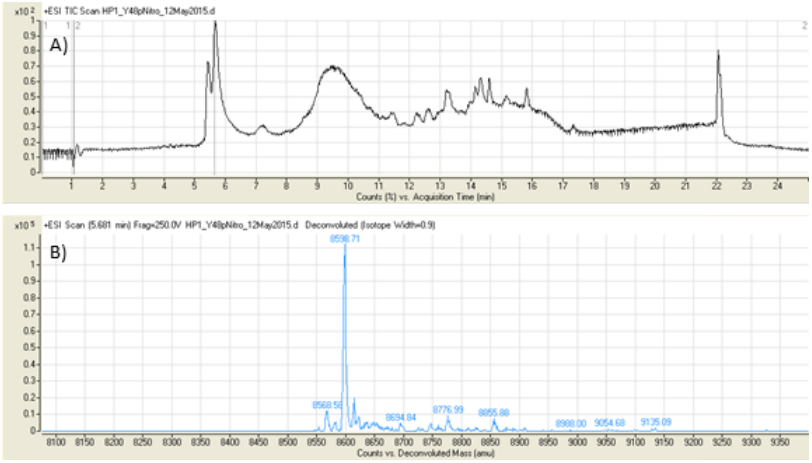
SI Figure 2.9 LCMS of HP1 Y48*p*CH_3F using method B. TIC scan (A) and corresponding m/z deconvolution (B).



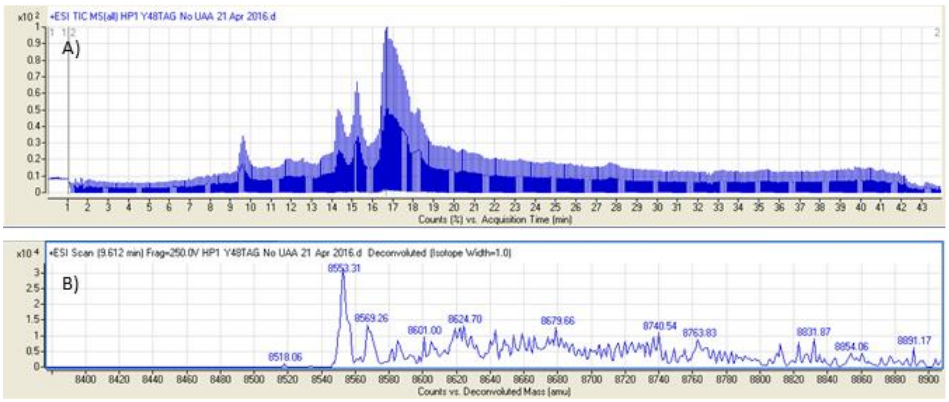
SI Figure 2.10 LCMS of HP1 Y48*p*CF_3F using method B.
TIC scan (A) and corresponding m/z deconvolution (B).



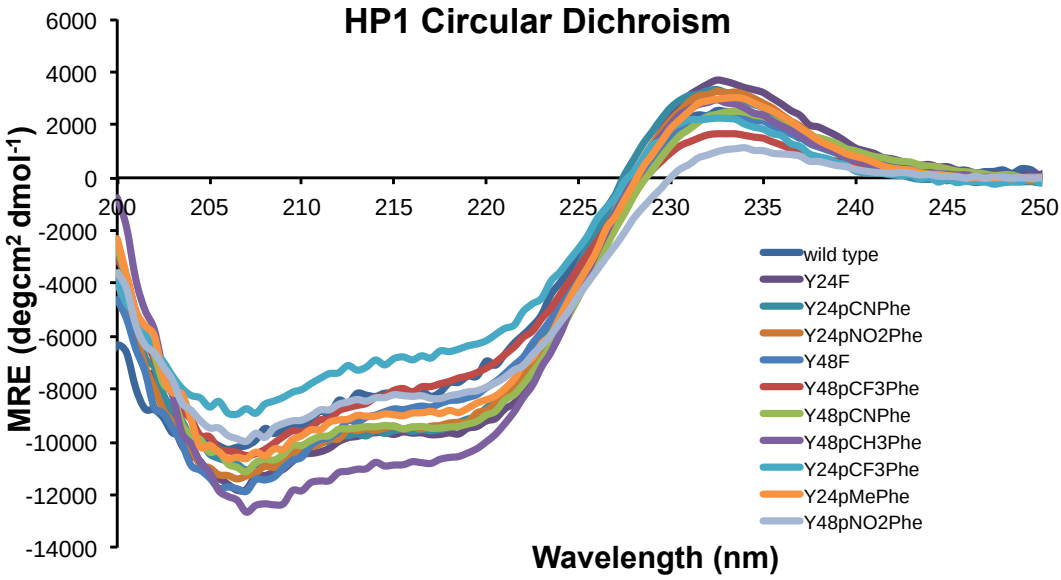
SI Figure 2.11 LCMS of HP1 Y48*p*CNF using method A.
TIC scan (A) and corresponding m/z deconvolution (B).



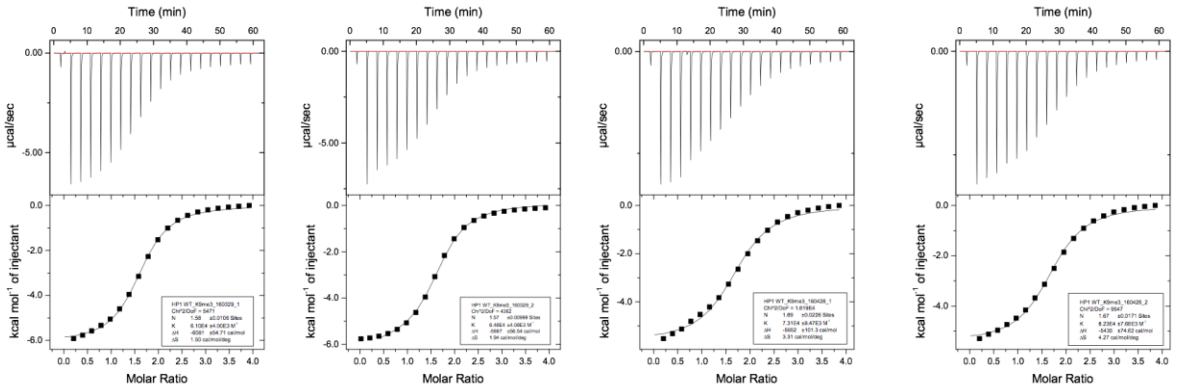
SI Figure 2.12 LCMS of HP1 Y48*p*NO_2F using method A.
TIC scan (A) and corresponding m/z deconvolution (B).



SI Figure 2.13 LCMS of HP1 Y48TAG with no UAA added using method B. TIC scan (A) and corresponding m/z deconvolution (B). Wild type HP1 and Y48F are produced in the absence of UAA, but when UAA is added wild type and Y48F are not detected.



SI Figure 2.14 Circular dichroism of HP1 mutants. Differences in spectra are likely due to calculated extinction coefficients used to quantitate protein concentration. Y24*p*MePhe is synonymous with Y24*p*CH_3Phe.



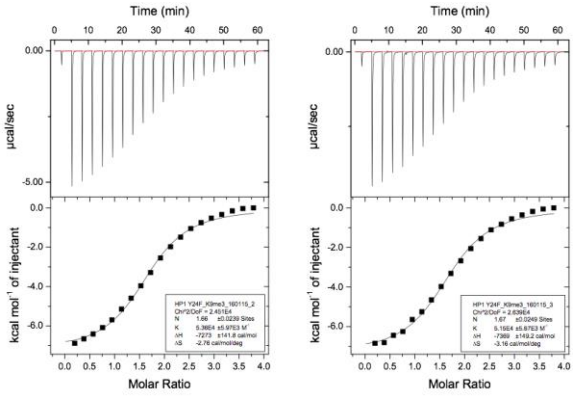
HP1 Wild Type - 1
[Peptide]: 4.95 mM
[Protein]: 255 µM
K_D: 16.39 µM
C: 15.6
N: 1.58

HP1 Wild Type - 2
[Peptide]: 4.95 mM
[Protein]: 255 µM
K_D: 15.43 µM
C: 16.5
N: 1.57

HP1 Wild Type - 3
[Peptide]: 2.86 mM
[Protein]: 150 µM
K_D: 13.68 µM
C: 11.0
N: 1.69

HP1 Wild Type - 4
[Peptide]: 2.86 mM
[Protein]: 150 µM
K_D: 12.15 µM
C: 12.3
N: 1.67

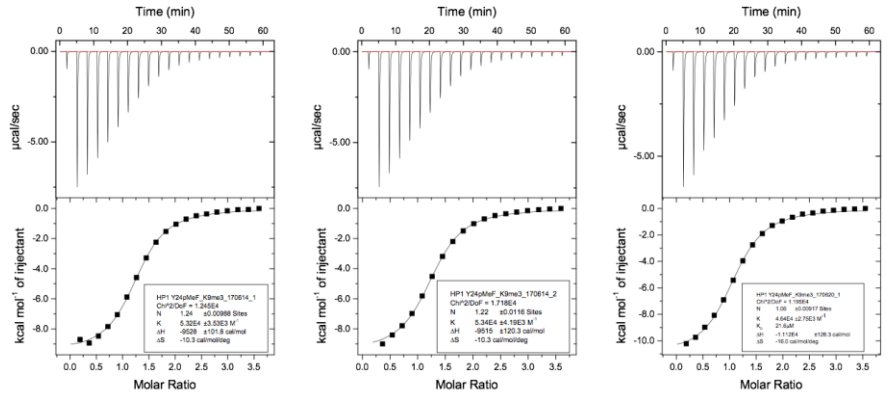
SI Figure 2.15 ITC curves of H3K9me3 peptide binding to wild type HP1.



HP1 Y24F - 1
[Peptide]: 3.00 mM
[Protein]: 160 µM
K_D: 18.66 µM
C: 8.6
N: 1.66

HP1 Y24F - 2
[Peptide]: 3.00 mM
[Protein]: 160 µM
K_D: 19.42 µM
C: 8.2
N: 1.67

SI Figure 2.16 ITC curves of H3K9me3 peptide binding to HP1 Y24F.

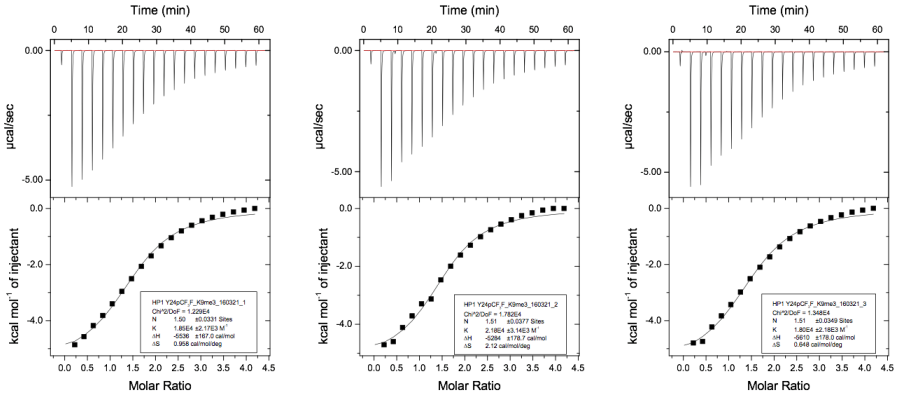


HP1 Y24pMeF - 1
[Peptide]: 4.75 mM
[Protein]: 267 µM
K_D: 18.80 µM
C: 14.2
N: 1.24

HP1 Y24pMeF - 2
[Peptide]: 4.75 mM
[Protein]: 267 µM
K_D: 18.73 µM
C: 14.3
N: 1.22

HP1 Y24pMeF - 3
[Peptide]: 4.25 mM
[Protein]: 242 µM
K_D: 21.55 µM
C: 11.2
N: 1.06

SI Figure 2.17 ITC curves of H3K9me3 peptide binding to HP1 Y24*p*CH_3F.

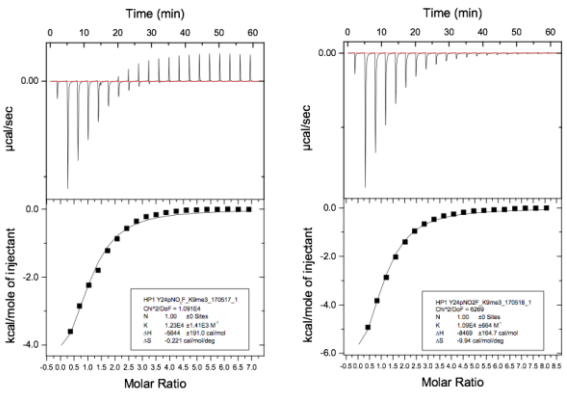


HP1 Y24pCF$_3$F - 1
[Peptide]: 5.18 mM
[Protein]: 250 µM
K_D: 54.05 µM
C: 4.6
N: 1.50

HP1 Y24pCF$_3$F - 2
[Peptide]: 5.18 mM
[Protein]: 250 µM
K_D: 45.87 µM
C: 5.5
N: 1.51

HP1 Y24pCF$_3$F - 3
[Peptide]: 5.18 mM
[Protein]: 250 µM
K_D: 55.56 µM
C: 4.5
N: 1.51

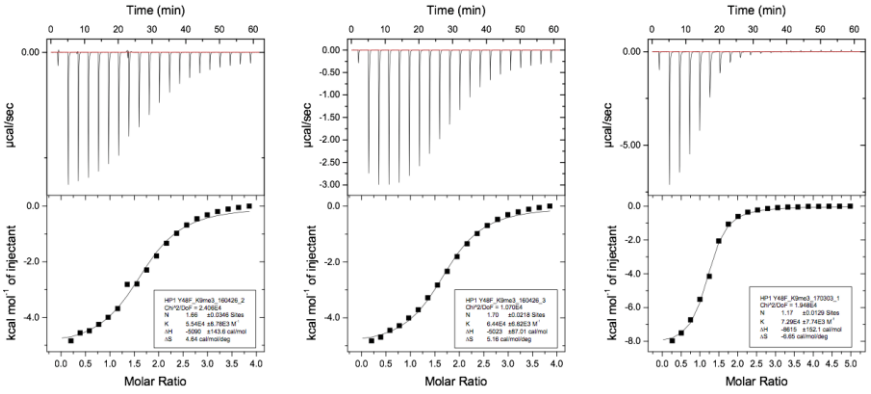
SI Figure 2.18 ITC curves of H3K9me3 peptide binding to HP1 Y24*p*CF_3F.



HP1 Y24pNO$_2$F - 1
[Peptide]: 7.00 mM
[Protein]: 205 µM
K_D: 91.74 µM
C: 2.5
N: 1.00

HP1 Y24pNO$_2$F - 2
[Peptide]: 7.47 mM
[Protein]: 187 µM
K_D: 91.74 µM
C: 2.0
N: 1.00

SI Figure 2.19 ITC curves of H3K9me3 peptide binding to HP1 Y24*p*NO$_2$F.

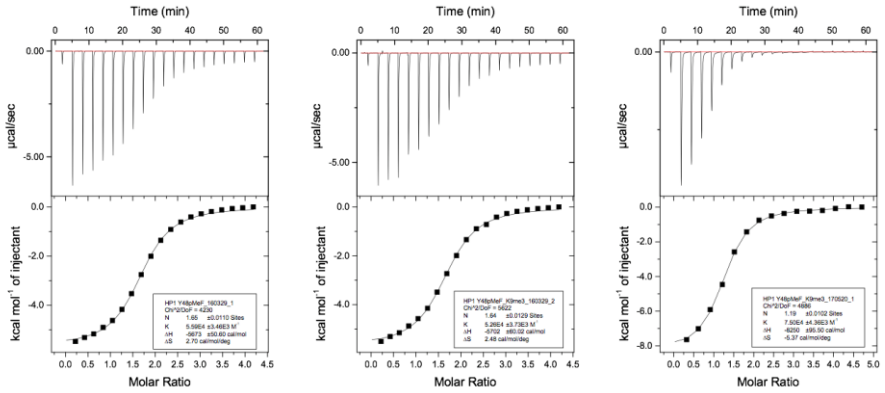


HP1 Y48F - 1
[Peptide]: 2.86 mM
[Protein]: 150 µM
K_D: 18.05 µM
C: 8.3
N: 1.66

HP1 Y48F - 2
[Peptide]: 2.86 mM
[Protein]: 150 µM
K_D: 15.53 µM
C: 9.7
N: 1.70

HP1 Y48F - 3
[Peptide]: 5.20 mM
[Protein]: 210 µM
K_D: 13.72 µM
C: 15.3
N: 1.17

SI Figure 2.20 ITC curves of H3K9me3 binding to HP1 Y48F.

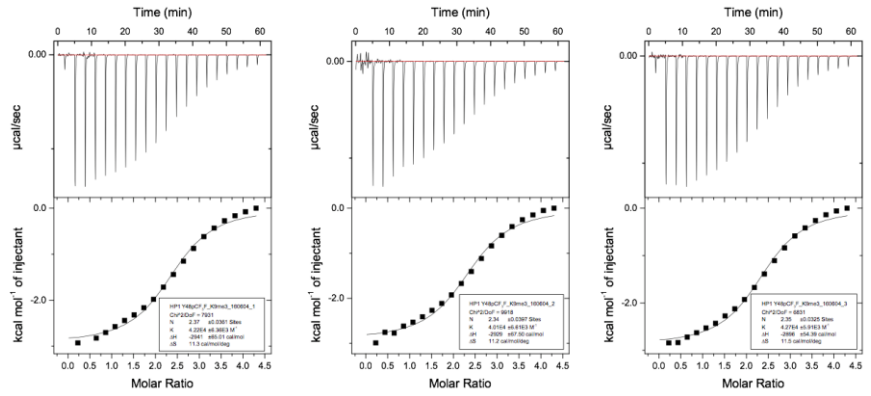


HP1 Y48pMeF - 1
[Peptide]: 4.95 mM
[Protein]: 240 µM
K_D: 17.89 µM
C: 13.4
N: 1.65

HP1 Y48pMeF - 2
[Peptide]: 4.95 mM
[Protein]: 240 µM
K_D: 19.01 µM
C: 12.6
N: 1.64

HP1 Y48pMeF - 3
[Peptide]: 4.60 mM
[Protein]: 153 µM
K_D: 13.33 µM
C: 11.5
N: 1.19

SI Figure 2.21 ITC curves of H3K9me3 peptide binding to HP1 Y48*p*CH_3F.

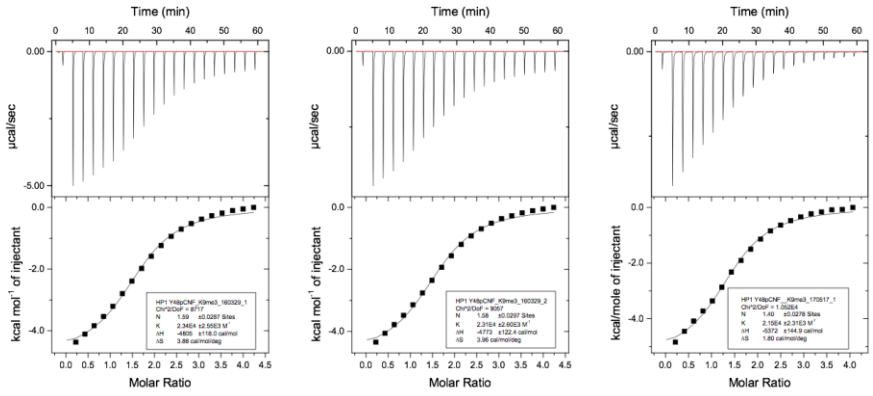


HP1 Y48pCF$_3$F - 1
[Peptide]: 5.00 mM
[Protein]: 235 µM
K_D: 23.70 µM
C: 9.9
N: 2.37

HP1 Y48pCF$_3$F - 2
[Peptide]: 5.00 mM
[Protein]: 235 µM
K_D: 24.94 µM
C: 9.4
N: 2.34

HP1 Y48pCF$_3$F - 3
[Peptide]: 5.00 mM
[Protein]: 235 µM
K_D: 23.42 µM
C: 10.0
N: 2.35

SI Figure 2.22 ITC curves of H3K9me3 peptide binding to HP1 Y48*p*CF_3F.

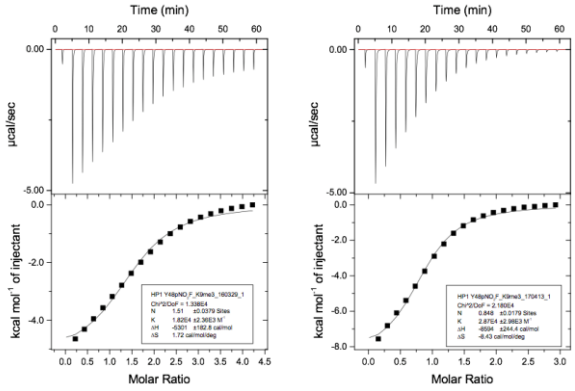


HP1 Y48pCNF - 1
[Peptide]: 4.95 mM
[Protein]: 240 µM
K_D: 42.74 µM
C: 5.6
N: 1.59

HP1 Y48pCNF - 2
[Peptide]: 4.95 mM
[Protein]: 240 µM
K_D: 43.29 µM
C: 5.5
N: 1.58

HP1 Y48pCNF - 3
[Peptide]: 5.21mM
[Protein]: 259 µM
K_D: 46.51 µM
C: 5.6
N: 1.40

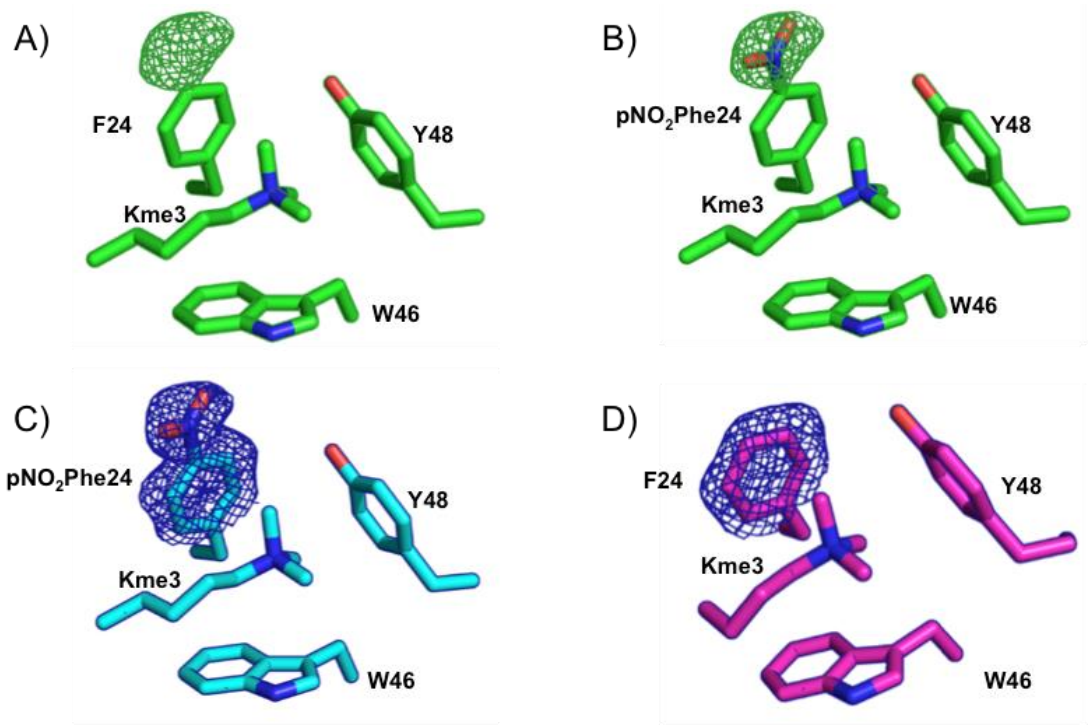
SI Figure 2.23 ITC curves of H3K9me3 peptide binding to HP1 Y48*p*CNF.



HP1 Y48pNO$_2$F - 1
[Peptide]: 4.95 mM
[Protein]: 240 µM
K_D: 54.95 µM
C: 4.4
N: 1.51

HP1 Y48pNO$_2$F - 2
[Peptide]: 4.10 mM
[Protein]: 280 µM
K_D: 34.8 µM
C: 6.8
N: 0.856

SI Figure 2.24 ITC curves of H3K9me3 peptide binding to HP1 Y24*p*NO$_2$F.



SI Figure 2.25 Density maps of the HP1 Y24 mutants. A) mFo-DFc map of Y24*p*NO_2F density with Y24F mutation shows additional density at *para*-position. B) When Y24*p*NO_2F mutation is modeled into the Y24*p*NO_2F mFo-DFc density, the nitro group fits the density well. C) 2mFo-DFc density map of the Y24*p*NO_2F density with Y24*p*NO_2F mutation shows *p*NO_2F mutation is present. D) 2mFo-DFc density of Y24F shows the differences in density for the F and Y24*p*NO_2F amino acids.

REFERENCES

(1) Dawson, M. A.; Kouzarides, T.; Huntly, B. J. P. *N. Engl. J. Med.* **2012**, *367* (7), 647–657.

(2) Di Croce, L.; Helin, K. *Nat. Struct. Mol. Biol.* **2013**, *20* (10), 1147–1155.

(3) Kwon, S. H.; Workman, J. L. *BioEssays* **2011**, *33* (4), 280–289.

(4) Helin, K.; Dhanak, D. *Nature* **2013**, *502* (7472), 480–488.

(5) Milosevich, N.; Hof, F. *Biochemistry* **2016**, *55* (11), 1570–1583.

(6) Kamps, J. J. A. G.; Huang, J.; Poater, J.; Xu, C.; Pieters, B. J. G. E.; Dong, A.; Min, J.; Sherman, W.; Beuming, T.; Matthias Bickelhaupt, F.; Li, H.; Mecinović, J. *Nat. Commun.* **2015**, *6*, 8911.

(7) Stuckey, J. I.; Simpson, C.; Norris-Drouin, J. L.; Cholensky, S. H.; Lee, J.; Pasca, R.; Cheng, N.; Dickson, B. M.; Pearce, K. H.; Frye, S. V.; James, L. I. *J. Med. Chem.* **2016**, *59* (19), 8913–8923.

(8) Kaustov, L.; Ouyang, H.; Amaya, M.; Lemak, A.; Nady, N.; Duan, S.; Wasney, G. A.; Li, Z.; Vedadi, M.; Schapira, M.; Min, J.; Arrowsmith, C. H. *J. Biol. Chem.* **2011**, *286* (1), 521–529.

(9) James, L. I.; Barsyte-Lovejoy, D.; Zhong, N.; Krichevsky, L.; Korboukh, V. K.; Herold, J. M.; MacNevin, C. J.; Norris, J. L.; Sagum, C. A.; Tempel, W.; Marcon, E.; Guo, H.; Gao, C.; Huang, X.-P.; Duan, S.; Emili, A.; Greenblatt, J. F.; Kireev, D. B.; Jin, J.; Janzen, W. P.; Brown, P. J.; Bedford, M. T.; Arrowsmith, C. H.; Frye, S. V. *Nat. Chem. Biol.* **2013**, *9* (3), 184–191.

(10) Dougherty, D. A. *Acc. Chem. Res.* **2013**, *46* (4), 885–893.

(11) Dougherty, D. A. *Science.* **1996**, *271* (5246), 163–168.

(12) Wheeler, S. E.; Houk, K. N. *J. Am. Chem. Soc.* **2009**, *131*, 3126–3127.

(13) Mecozzi, S.; West, A. P.; Dougherty, D. A. *J. Am. Chem. Soc.* **1996**, *118* (9), 2307–2308.

(14) Sunner, J.; Nishizawa, K.; Kebarle, P. *J. Phys. Chem.* **1981**, *85* (2), 1814–1820.

(15) Gallivan, J. P.; Dougherty, D. A. *Proc. Natl. Acad. Sci.* **1999**, *96* (17), 9459–9464.

(16) Herold, J. M.; Wigle, T. J.; Norris, J. L.; Lam, R.; Korboukh, V. K.; Gao, C.; Ingerman, L. A.; Kireev, D. B.; Senisterra, G.; Vedadi, M.; Tripathy, A.; Brown, P. J.; Arrowsmith, C. H.; Jin, J.; Janzen, W. P.; Frye, S. V. *J. Med. Chem.* **2011**, *54* (7), 2504–2511.

(17) Zürcher, M.; Diederich, F. *J. Org. Chem.* **2008**, *73* (12), 4345–4361.

(18) Zhong, W.; Gallivan, J. P.; Zhang, Y.; Li, L.; Lester, H. A.; Dougherty, D. A. *Proc. Natl. Acad. Sci. U. S. A.* **1998**, *95* (21), 12088–120933.

(19) Bacher, J. M.; Ellington, A. D. *J. Bacteriol* **2001**, *183* (18), 5414–5426.

(20) Adams-Cioaba, M. A.; Min, J. *Biochem. Cell Biol.* **2009**, *87* (1), 93–105.

(21) Perfetti, M. T.; Baughman, B. M.; Dickson, B. M.; Mu, Y.; Cui, G.; Mader, P.; Dong, A.; Norris, J. L.; Rothbart, S. B.; Strahl, B. D.; Brown, P. J.; Janzen, W. P.; Arrowsmith, C. H.; Mer, G.; McBride, K. M.; James, L. I.; Frye, S. V. *ACS Chem. Biol.* **2015**, *10* (4), 1072–1081.

(22) Sanchez, R.; Zhou, M. *Trends Biochem. Sci.* **2011**, *36* (7), 364–372.

(23) Jacobs, S. A.; Taverna, S. D.; Zhang, Y.; Briggs, S. D.; Li, J.; Eissenberg, J. C.; Allis, C. D.; Khorasanizadeh, S. *EMBO J.* **2001**, *20* (18), 5232–5241.

(24) Jacobs, S. A; Khorasanizadeh, S. *Science* **2002**, *295* (5562), 2080–2083.

(25) Dougherty, D. A. *Curr. Opin. Chem. Biol.* **2000**, *4* (6), 645–652.

(26) Saks, M. E.; Sampson, J. R.; Nowak, M. W.; Kearney, P. C.; Du, F.; Abelson, J. N.; Lester, H. A.; Dougherty, D. A. *J. Biol. Chem.* **1996**, *271* (38), 23169–23175.

(27) Lester, H. A. *Science* **1988**, *241* (4869), 1057–1063.

(28) Wang, L.; Brock, A.; Herberich, B.; Schultz, P. G. *Science* **2001**, *292* (5516), 498–500.

(29) Brustad, E.; Bushey, M. L.; Lee, J. W.; Groff, D.; Liu, W.; Schultz, P. G. *Angew. Chem. Int. Ed. Engl.* **2008**, *47* (43), 8220–8223.

(30) Wan, W.; Tharp, J. M.; Liu, W. R. *Biochim. Biophys. Acta* **2014**, *1844* (6), 1059–1070.

(31) Liu, C. C.; Schultz, P. G. *Annu. Rev. Biochem.* **2010**, *79*, 413–444.

(32) Santiago, C.; Nguyen, K.; Schapira, M. *J. Comput. Aided. Mol. Des.* **2011**, *25* (12), 1171–1178.

(33) Wang, L.; Magliery, T. J.; Liu, D. R.; Schultz, P. G. *J. Am. Chem. Soc.* **2000**, *122* (20), 5010–5011.

(34) Yoo, T. H.; Link, A. J.; Tirrell, D. A. *Proc. Natl. Acad. Sci. U. S. A.* **2007**, *104* (35), 13887–13890.

(35) Liu, W.; Alfonta, L.; Mack, A. V.; Schultz, P. G. *Angew. Chemie - Int. Ed.* **2007**, *46* (32), 6073–6075.

(36) Wang, J.; Xie, J.; Schultz, P. G. *J. Am. Chem. Soc.* **2006**, *128* (27), 8738–8739.

(37) Tharp, J. M.; Wang, Y.-S.; Lee, Y.-J.; Yang, Y.; Liu, W. R. *ACS Chem. Biol.* **2014**, *9* (4), 884–890.

(38) Young, D. D.; Young, T. S.; Jahnz, M.; Ahmad, I.; Spraggon, G.; Schultz, P. G. *Biochemistry* **2011**, *50* (11), 1894–1900.

(39) Miyake-Stoner, S. J.; Refakis, C. A.; Hammill, J. T.; Lusic, H.; Hazen, J. L.; Deiters, A.; Mehl, R. A.. *Biochemistry* **2010**, *49* (8), 1667–1677.

(40) Chatterjee, A.; Sun, S. B.; Furman, J. L.; Xiao, H.; Schultz, P. G. *Biochemistry* **2013**, *52* (10), 1828-1837.

(41) Hammill, J. T.; Miyake-Stoner, S.; Hazen, J. L.; Jackson, J. C.; Mehl, R. A. *Nat. Protoc.* **2007**, *2* (10), 2601–2607.

(42) Studier, F. W. *Protein Expr. Purif.* **2005**, *41* (1), 207–234.

(43) Hughes, R. M.; Wiggins, K. R.; Khorasanizadeh, S.; Waters, M. L. *Proc. Natl. Acad. Sci. U. S. A.* **2007**, *104* (27), 11184–11188.

(44) Fischle, W.; Wang, Y.; Jacobs, S. A.; Kim, Y.; Allis, C. D.; Khorasanizadeh, S. *Genes Dev.* **2003**, *17* (15), 1870–1871.

(45) Eisert, R. J.; Waters, M. L. *Chembiochem* **2011**, *12* (18), 2786–2790.

(46) Davis, M. R.; Dougherty, D. A. *Phys. Chem. Chem. Phys.* **2015**, *17* (43), 29262–29270.

(47) Lee, Y.-J.; Schmidt, M. J.; Tharp, J. M.; Weber, A.; Koenig, A. L.; Zheng, H.; Gao, J.; Waters, M. L.; Summerer, D.; Liu, W. R. *Chem. Commun.* **2016**, *52* (85), 12606–12609.

(48) Mahadevi, A. S.; Sastry, G. N. *Int. J. Quantum Chem.* **2014**, *114* (2), 145–153.

(49) Turnbull, W. B.; Daranas, A. H. *J. Am. Chem. Soc.* **2003**, *125* (48), 14859–14866.

(50) Sokkalingam, P.; Shraberg, J.; Rick, S. W.; Gibb, B. C. *J. Am. Chem. Soc.* **2016**, *138* (1), 48–51.

(51) Fu, Z. Q. *Acta Crystallogr. Sect. D Biol. Crystallogr.* **2005**, *61* (12), 1643–16438.

(52) Adams, P. D.; Afonine, P. V.; Bunkóczi, G.; Chen, V. B.; Davis, I. W.; Echols, N.; Headd, J. J.; Hung, L. W.; Kapral, G. J.; Grosse-Kunstleve, R. W.; McCoy, A. J.; Moriarty, N. W.; Oeffner, R.; Read, R. J.; Richardson, D. C.; Richardson, J. S.; Terwilliger, T. C.; Zwart, P. H. *Acta Crystallogr. Sect. D Biol. Crystallogr.* **2010**, *66* (2), 213–221.

(53) Emsley, P.; Lohkamp, B.; Scott, W. G.; Cowtan, K. *Acta Crystallogr. Sect. D Biol. Crystallogr.* **2010**, *66* (4), 486–501.

(54) Zhao, Y.; Truhlar, D. G. *Theor. Chem. Acc.* **2008**, *120* (1–3), 215–241.

(55) Gaussian 09, R. A., M. J. Frisch, G. W. Trucks, H. B. Schlegel, G. E. Scuseria, M. A. Robb, J. R. Cheeseman, G. Scalmani, V. Barone, G. A. Petersson, H. Nakatsuji, X. Li, M. Caricato, A. Marenich, J. Bloino, B. G. Janesko, R. Gomperts, B. Mennucci, H. P. Hratchian, J. V. Ortiz, A. F. Izmaylov, J. L. Sonnenberg, D. Williams-Young, F. Ding, F. Lipparini, F. Egidi, J. Goings, B. Peng, A. Petrone, T. Henderson, D. Ranasinghe, V. G. Zakrzewski, J. Gao, N. Rega, G. Zheng, W. Liang, M. Hada, M. Ehara, K. Toyota, R. Fukuda, J. Hasegawa, M. Ishida, T. Nakajima, Y. Honda, O. Kitao, H. Nakai, T. Vreven, K. Throssell, J. A. Montgomery, Jr., J. E. Peralta, F. Ogliaro, M. Bearpark, J. J. Heyd, E. Brothers, K. N. Kudin, V. N. Staroverov, T. Keith, R. Kobayashi, J. Normand, K. Raghavachari, A. Rendell, J. C. Burant, S. S. Iyengar, J. Tomasi, M. Cossi, J. M. Millam, M. Klene, C. Adamo, R. Cammi, J. W. Ochterski, R. L. Martin, K. Morokuma, O. Farkas, J. B. Foresman, and D. J. Fox, Gaussian, Inc., Wallingford CT, 2016.

(57) Legault, C. Y. C., version 1.0b, Université de Sherbrooke, 2009.

CHAPTER 3: PROBING CATION-π INTERACTIONS OF MAMMALIAN READER PROTEINS USING IN VIVO UNNATURAL AMINO ACID MUTAGENESIS

3.1 Heterochromatin Protein 1 is a Model System

Heterochromatin Protein 1 (HP1) was chosen for study in chapter 2 because it is a model system with a known crystal structure that expresses well in *Escherichia coli*, providing a convenient starting point for UAA incorporation. However, success in investigating HP1 prompted us to examine more interesting (and challenging) methyllysine readers that are relevant for human health and disease. In addition, the University of North Carolina at Chapel Hill is an ideal location to expand this work due to the large concentration of scientists pursuing epigenetics research. As a result of collaborations with Stephen Frye's lab in the Center for Integrative Chemical Biology and Drug Discovery, we have expanded our studies to mammalian readers of therapeutic interest. Here we describe our initial progress expanding efforts from chapter 2 towards the expression and engineering of two mammalian reader proteins, CBX7 and CBX5.

3.2 Chromobox Proteins in Mammals

Like their *Drosophila* counterparts, mammalian cells possess a diverse set proteins that decode PTMs on histone tails, including several families of methyllysine reader proteins. In humans, one family of these reader proteins are the chromobox (CBX) proteins. Human chromobox proteins can be further classified into two groups (Figure 3.1). The first group, containing CBX1, CBX3, and CBX5, are homologs of *Drosophila* HP1.[1] Because of their

homology, they are often referred to as HP1β, HP1γ, and HP1α, respectively.[2–4] HP1-like proteins are involved in the formation of heterochromatin and hence the control of gene expression through recognition of the H3K9me3 PTM.[1] HP1 proteins contain an N-terminal chromodomain, a C-terminal chromoshadow domain, and variable hinge region in between (Figure 3.2).[5] The hinge region interacts with H1 histones, nonspecific DNA and RNA, and chromatin.[5] The chromoshadow domain is important for protein-protein interactions with chromatin, chromatin-modifying proteins, replication factors, and transcriptional regulatory proteins.[5] The chromodomain of CBX5, of interest to this work, recognizes its H3K9me3 substrate and is responsible for localizing CBX5 to heterochromatin.[5]

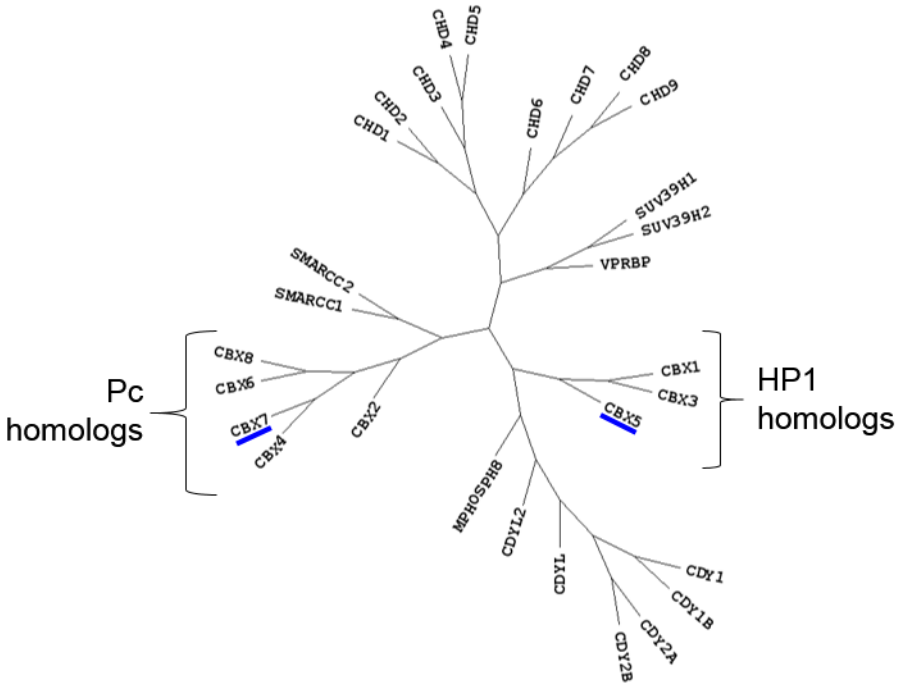


Figure 3.1 Phylogenetic tree of CBX proteins.[6]

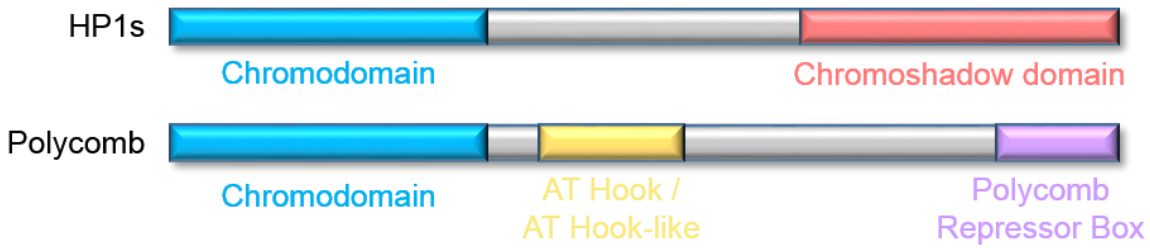


Figure 3.2 Domains of CBX proteins.

The second group of human chromodomains, comprised of CBX2, CBX4, CBX6, CBX7, and CBX8, are homologs of *Drosophila* polycomb (Pc) proteins.[1] Unlike HP1s, polycomb proteins target the H3K27me3 PTM.[4,7] Although both contain an N-terminal chromodomain, polycomb chromodomains tend to bind the respective H3K27me3 substrate with less affinity or specificity than their HP1 counterparts.[1] Polycomb proteins contain a DNA binding region known as an AT-hook or AT-hook-like motif as well as a C-terminal polycomb repressor box, which is important for association with protein partners (Figure 3.2).[4,8] Non-conserved regions in the center of the polycomb proteins are thought to contribute to binding specificity and localization.[4] The proteins have been implicated in the regulation of developmental genes.[1]

Gene expression is regulated by polycomb CBX proteins by targeting the polycomb repressive complex 1 (PRC1) to H3K27me3 site, a known repressive PTM.[7] In the 'canonical' polycomb signaling cascade, the H3K27me3 PTM is installed by the EZH1 and EZH2 methyltransferases of polycomb repressive complex 2 (PRC2) and then is recognized by the Kme3 reader's chromodomain of PRC1.[7,9,10] PRC1 reader domains from CBXs are thought to facilitate transcriptional repression by docking the complex to the H3K27me3 site. This docking allows for E3 ubiquitin ligase subunit in the PRC1 complex to monoubiquitinate Lys119 of H2A.[7,10] This ubiquitination is implicated in DNA methylation,

chromatin compaction, and gene repession.[7] Although a canonical model of the polycomb signaling pathway exists, not all PRC1 complexes fit this model, and the role that PRC1 diversity plays in gene regulation is not well understood.

3.3 Differences in CBX Binding are Due to Structural Variations in the Chromodomain

Chromobox proteins, like other methyllysine reader proteins, bind methyllysine ligands using aromatic cages.[1,10,11] The identity of the methyllysine ligands varies: HP1-like proteins bind H3K9me3 PTMs and Pc proteins bind H3K27me3 PTMs.[1,5,10] Ligand binding in CBX chromodomains utilizes two conserved elements: 1) a binding pocket complementary to an alanine residue of the histone "ARKS" motif where "K" is the methylation site, and 2) formation of a continuous β-sheet between the protein and the extended β-strand of the ligand.[1] Chromobox proteins all recognize ligands containing the "ARKS" motif, however it is the amino acids downstream of this sequence that are important for protein binding, suggesting specificity differences are largely due to the differences in the surface in contact with the ligand.[1] HP1s have conserved residues that act as "polar fingers" (Figure 3.3A). These residues, Glu3 and Asp42 in CBX5, help sandwich the ligand (specifically the threonine directly before the ARKS motif) into the binding groove. Polycomb proteins do not have "polar fingers" but instead use a hydrophobic clasp mechanism (Figure 3.3B). In CBX7, Val10 and Leu49 form a hydrophobic ring that allows for interaction with hydrophobic residues. This hydrophobic clasp is not as sterically restrictive as the polar fingers, and can accommodate various amino acids. The hydrophobic clasp's tolerance of multiple residues contributes to the diminished specificity of the polycomb chromodomains.[1]

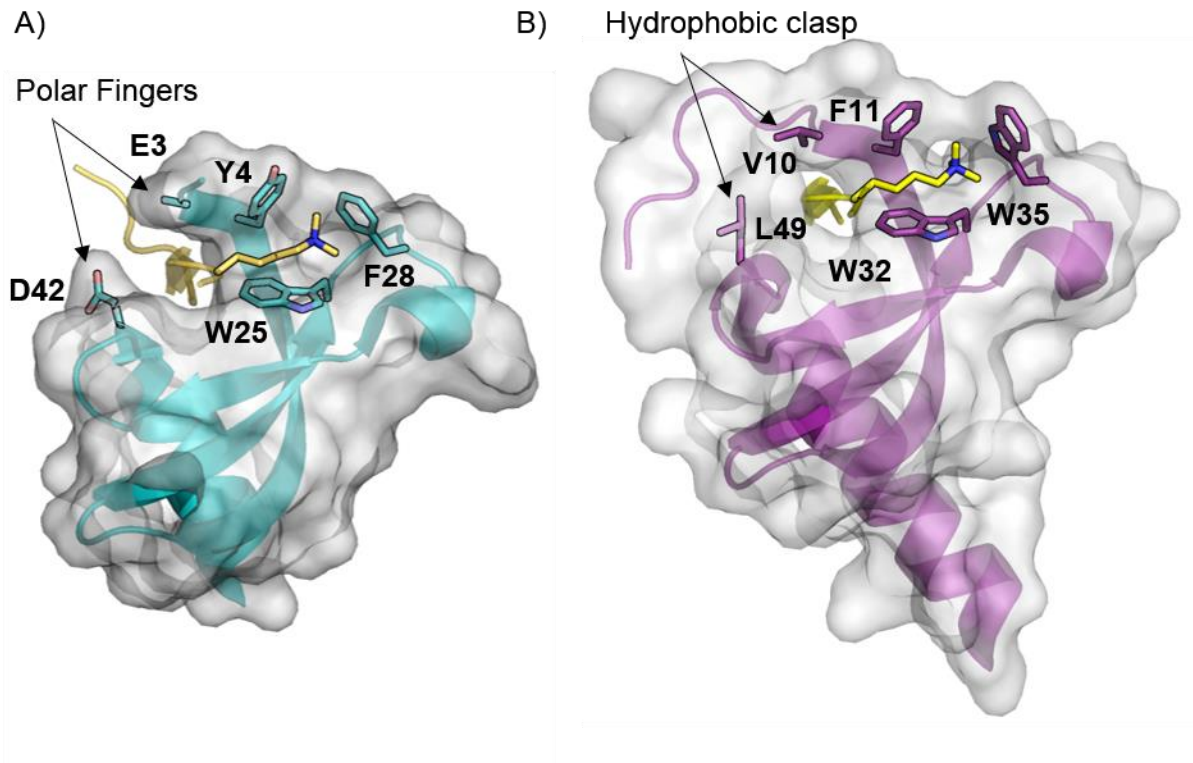


Figure 3.3 Structural features of chromodomains of HP1s and polycomb proteins. A) Structure of CBX5 (HP1α) with "polar finger" motif (PDB ID: 3FDT). B) Structure of CBX7 (Pc) with hydrophobic clasp (PDB ID: 4X3K).

3.4 Chromobox Proteins are Implicated in Disease

Methylation marks can literally be a death sentence: specific histone methylation events are associated with more aggressive cancers and their corresponding poor survival rates.[12] CBX7 is a well-studied CBX protein that has been implicated in multiple cancers.[7,12] In some gastric, prostate, leukemia, and lymphoma cell lines, CBX7 knockdowns lead to growth inhibition.[7] Overexpression of CBX7 in these same cell lines provides a proliferation advantage. Intriguingly, CBX7 performs a tumor suppression function in thyroid, colon, lung, and pancreatic tissues.[7] These conflicting functions are common to CBX proteins and other epigenetic readers whose activity is largely context- and tissue-dependent.[7,12]

In recent years, the control or modulation of H3K27me3 signaling via small molecules has become of significant target of pharmaceutical interest. Inhibition of the EZH2 methyltransferase of PRC2 reduced high H3K27me3 levels in lymphomas marked by activating mutations.[13] Inhibitors of certain H3K27me3 demethylases have shown activity against diffuse intrinsic pontine glioma tumors, an aggressive brain cancer with a <1% survival rate that affects mostly children.[14] Based on these developments, inhibition of Kme3 recognition of PRC1 may also be of therapeutic interest.[7,12] Although ligands for PRC1 chromodomains have been reported, very few act as high-quality, bioavailable and bioactive chemical probes.[7] Such probes that inhibit PRC1's reader function would allow for deconvolution of PRC1 cascades in the cell and potential roles in disease as well as determination of common characteristics with reported therapeutic inhibitors.[7]

3.5 Chromobox Proteins are Challenging Drug Targets

Although methyllysine reader proteins have been suggested as targets for therapeutics, the development of inhibitors targeting chromodomains of methyllysine readers lags far behind molecules that inhibit methyltransferases (PTM writer proteins).[12] This lag in progress is likely due to the challenging nature of methyllysine readers as drug targets: they are highly conserved in structure, they perform overlapping functions with other CBX proteins, and the downstream effects of inhibition are still largely unknown.[1,10,12,15] In addition, this challenge also reflects the difficulty in targeting protein-protein interaction motifs (reader proteins) when compared to the design of traditional small molecules that target enzyme active sites (writer proteins). Chromobox proteins use surface groove recognition to bind their methyllysine ligands.[1,12] This binding pocket is shallow, and binding

relies on hydrogen bonding between the protein and histone tail in addition to cation-π interactions between methyllysine adducts and side chains of the residues in the CBX aromatic cage.[12] The CBX chromodomains have high sequence similarity, especially in regions interacting with the H3 histone.[7] The similarity of CBX proteins suggests that an inhibitor of one CBX chromodomain could likely inhibit others.[7] Since competition and interaction between CBX proteins is still being elucidated, off target effects on other CBX proteins may complicate the screening process. CBX proteins function as part of complex signaling cascades and multi-subunit complexes, and so it is unclear how targeting one component in the complex cellular machine will effect biophysical functions.

By incorporating UAAs into HP1, we discovered the inequivalent nature of the two tyrosine residues and the added influence of the Y24 position. We theorize that the residues in the aromatic cages in CBX7 and CBX5 also do not contribute equally to cation-π binding, as was observed in HP1. By incorporating UAAs into the aromatic cage of CBX7 and CBX5, we may elucidate the contributions of each phenylalanine and tyrosine residue in the aromatic cage. Each residue's influence on binding affinity may be used by medicinal chemists to aid in the design of better inhibitors of these therapeutic targets.

3.6 CBX7: A Polycomb Protein

CBX7 is a well-studied chromobox protein believed to play a role in regulation of cell differentiation.[7] CBX7-PRC1 represses genes involved in early lineage commitment.[10] In embryonic stem cells of mice, CBX7 must bind to H3K27me for PCRC1 targeting. In differentiating cells, CBX2 and CBX4 replace CBX7 in the PRC1 complex, helping to repress pluripotency genes required for cell differentiation.[10] Development of agonists or

chemical probes for CBX7 has proven difficult due to the dynamic nature of the Kme3 biding pocket.[7,12]

In the absence of substrate, the aromatic cage of CBX7 is unformed (Figure 3.4).[7] The binding H3K27me3 peptide substrate results in the formation of the aromatic cage through induced-fit using the peptide's β-sheet. Structural modeling studies of CBX7 have shown that H3K27me3 binding is similar to the mechanism of β-hairpin folding between two antiparallel β-sheets.[7] One β-sheet is provided by the histone peptide, the other by residues 8-13 of CBX7.[1,7,12] Hairpin formation requires contacts between the N-terminal end of the histone peptide and CBX7. Once these contacts are made, a β-turn is formed, allowing for backbone hydrogen bonding between the β-sheets to "zip" the complex into its folded state.[7] This folded state also orients the Phe11 residue towards the Kme3 residue, allowing for the formation of the aromatic cage.[7]

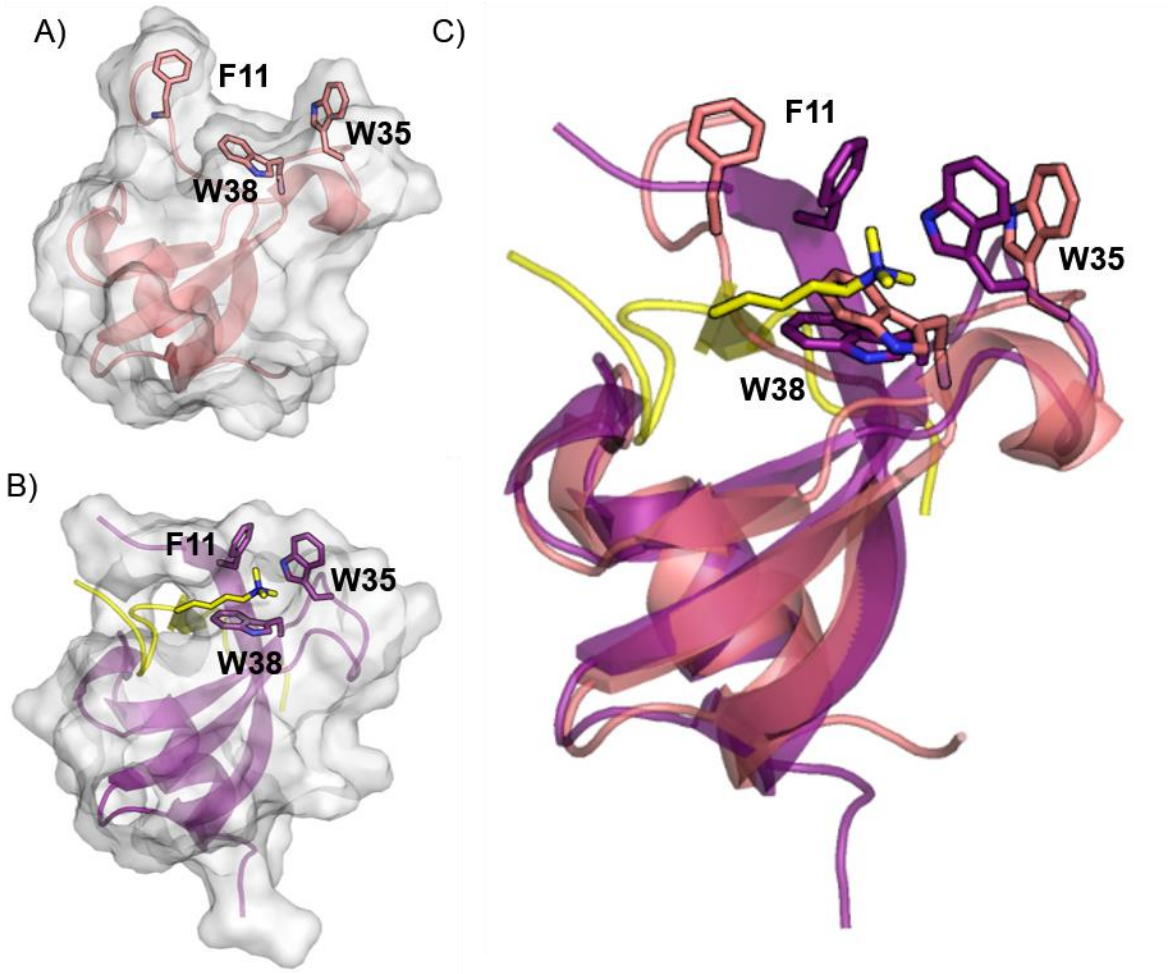


Figure 3.4 Aromatic cage of CBX7 forms upon ligand binding. A) Apo-CBX7 (PDB ID: 2K1B). B) CBX7 bound to the H3K27me3 ligand (PDB: 2L1B).

3.7 Development of CBX7 Inhibitors

The Frye lab, at UNC Chapel Hill, was one of the first to report a potent and bioactive inhibitor of PRC1 chromodomains. This inhibitor, UNC3866, was found to be the most potent agonist for CBX7 with a K_d of 97 ± 2.4 nM. After co-crystallizing CBX7 and UNC3866 (Figure 3.5), it was found that UNC3866's amide backbone forms multiple hydrogen bonds to CBX7, while CBX7's Asp50, Arg52, and Leu53 residues interact with the tert-butylbenzoyl N-terminus of UNC3866. These contacts are similar to the contacts between CBX7 and its native substrate that facilitate loop closure and formation of the

aromatic cage.[7] Notably, a diethyllysine moiety (a one-carbon extension of natural dimethyllysine PTMs) of UNC3866 buries its cationic head group in the aromatic cage of CBX7, binding that is similar to native Kme3 and Kme2 substrates.

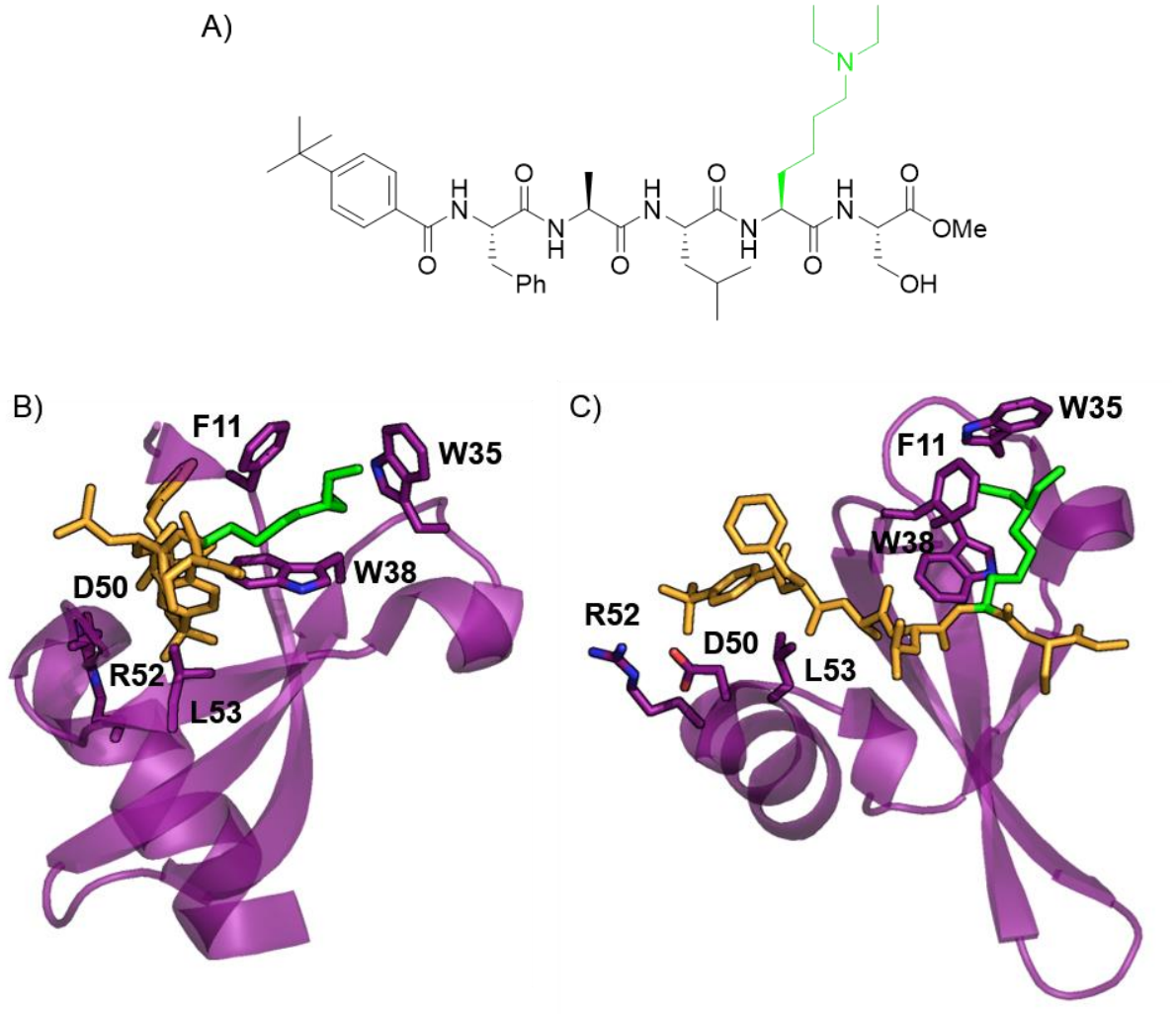


Figure 3.5 Structure of UNC3866 in complex with CBX7. A) Structure of UNC3866. B) Structure of CBX7 complexed with UNC3866. C) Complex in (B) rotated to better show binding of UNC3866 to the surface groove.

3.8 CBX7 as a Candidate for UAA Mutagenesis

CBX7 was a natural choice for UAA mutagenesis due to its well-characterized structure and comparable size to HP1 from *Drosophila*. The aromatic cage of CBX7 contains

two tryptophan residues (one located in the position analogous to Y48 in HP1) and a phenylalanine at position 11, which is analogous to the Y24 position in HP1. Conveniently, by investigating UNC3866 Kme3 analogs (Figure 3.6), we may exploit the high affinity of the inhibitor (typically mid to high nM) for CBX7 to reduce the amount of UAA-protein required for binding experiments such as ITC or anisotropy. CBX7 binding studies with UNC3866 and its analogs have been extensively studied by ITC,[7,9] allowing for facile transition from our HP1 system to the new CBX7 system.

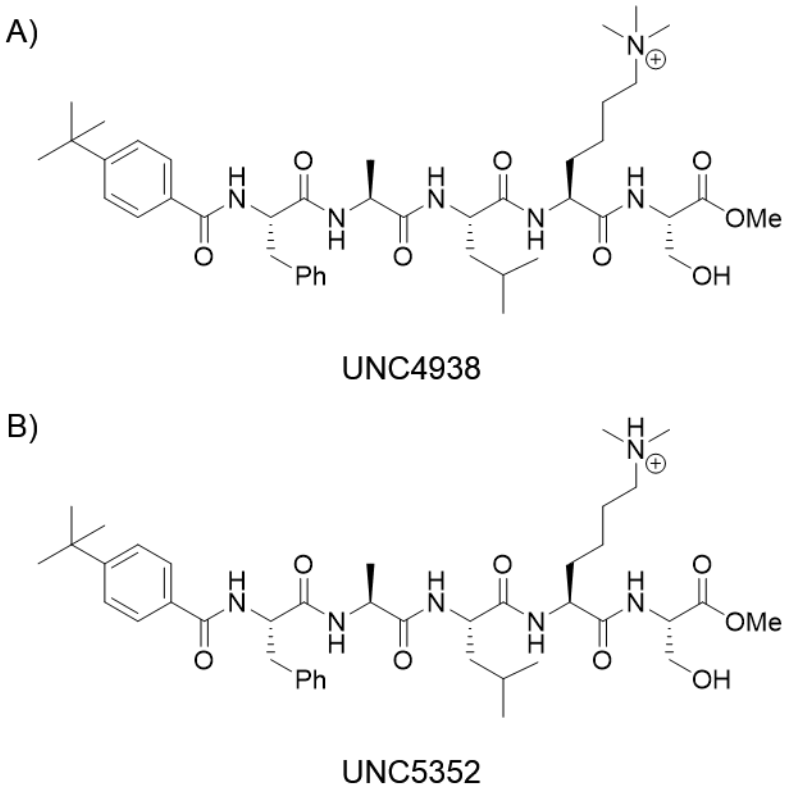


Figure 3.6 Structure of UNC3866 derivatives for CBX7-UAA binding.

3.9 Engineering of a CBX7 Expression System

CBX7 expression and purification conditions have been previously reported using BL21 Rosetta™ 2 (DE3) pLysS cells.[7] The Rosetta™ 2 (DE3) pLysS cell line is an optimized expression vector for expressing mammalian proteins in *E. coli*. These Rosetta™

cells contain an auxiliary chloramphenicol-resistant plasmid, pRARE2-pLysS. pRARE2 has been used to overcome codon bias of *E. coli* of mammalian proteins by supplying seven common mammalian codons that are used more scarcely in *E. coli* cells.[16] pLysS encodes the T7 lysozyme gene, which inhibits T7 RNA polymerase and prevents background protein expression until IPTG induction.[16] Unfortunately, this cell line appeared to be incompatible with the pUltra-*p*CNPheRS auxiliary vector required for high yields of UAA-proteins as evidenced by cell death, slow doubling times, and failed transformations of pUltra into Rosetta™ 2 (DE3) pLysS competent cells.

Initially, we theorized that cytotoxicity of pUltra-*p*CNPheRS in Rosetta™ 2 cells was due to a cross-reactivity of the *p*CNPheRS-tRNA and the tRNAs encoded in pRARE2. To alleviate the need for these codons, CBX7 was codon optimized for *E. coli* expression using Integrated DNA Technologies (IDT)'s codon optimization tool. Codon-optimized CBX7 mutants were transformed into BL21 (DE3) pLysS cells. These competent cells lack the exogenous tRNAs from pRARE2, but still contain the pLysS gene, which helps fine tune expression. Codon-optimized wild type CBX7 and F11Y mutants expressed well in BL21 (DE3) pLysS, however mutants co-transformed with pUltra still grew slowly or not at all. To determine if low expression and slow growth were due to the CBX7 mutant cytotoxicity, pET11a-HP1 wild type and pUC19, a control vector, were co-transformed into BL21 (DE3) pLysS cells with or without pUltra-*p*CNPheRS. This experiment would serve to determine whether expression issues are caused in a CBX7-dependent manner or due to UAA-incorporation technology. Cell populations bearing co-transformations of HP1/pUltra or pUC19/pUltra showed slowed growth and wild type HP1/pUltra cells showed no HP1 expression, suggesting that pLysS and pUltra-*p*CNPheRS are not compatible for protein

expression. The nature of this plasmid conflict remains unclear and of future investigation. This work does highlight the difficulty in translating UAA-based technologies from model systems to human proteins, where heterologous expressions are often highly tailored in a manner that is incompatible with UAA-mutagenesis.

Luckily, pLysS is not the only way to minimize background protein expression. pET vectors are known for having "leaky" promoters that make background expression more likely in the absence of inductant.[17,18] Other vectors, such as the pBad vector, have more tightly regulated promoters, which reduce basal expression. Most pBad vectors are not compatible with streptomycin-resistant pUltra vector due to antibiotic resistance: pBad is an arabinose-inducible vector and requires expression in DH10B, which is naturally streptomycin resistant. However, pBad has previously been shown to work well with an alternative accessory plasmid for UAA mutagenesis (pDule vectors) which are also appropriate for autoinduction-based expressions.[19] Anecdotally, pDule-*p*CNPheRS compatibility with HP1-UAA incorporation has been previously confirmed by our HP1 expression screening. Therefore, this alternate UAA accessory plasmid was added to screens for CBX7-UAA incorporation.

Basal expression may also be minimized using catabolite repression.[17,18] Studies have shown that high concentrations of glucose inhibit lactose activation of the *lac* operon. In high concentrations of glucose, cyclic adenosine monophosphate (cAMP) levels are low. cAMP and its receptor protein, cAMP-receptor protein (CAP), bind upstream of the *lac* promoter and thereby activate transcription by RNA polymerase. cAMP is required for activation of the *lac* operon, and so in high glucose concentrations the *lac* operon is repressed. However, once the cell has metabolized the glucose, the cell switches to another carbon source and

cAMP levels rise, activating the *lac* operon and allowing for induction of protein expression. By adding 0.5-1.0% glucose, background expression can be easily minimized in BL21 (DE3) strains.

After running initial test expressions (Figure 3.7), glucose-spiked BL21 (DE3) expressions showed CBX7-UAA incorporation. The increased glucose concentration repressed background expression of CBX7 and also increased cell density before induction, thereby increasing yields. With a functioning CBX7 UAA-incorporation expression system, we moved forward with CBX7 characterization.

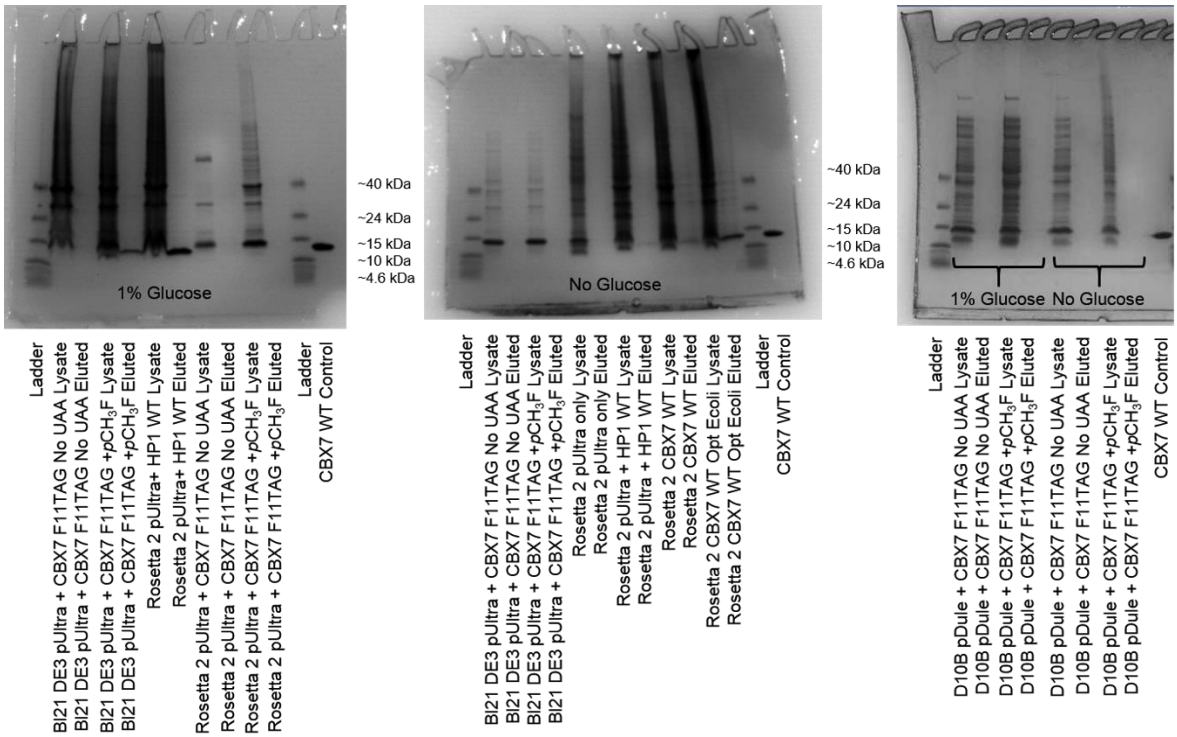


Figure 3.7 SDS-PAGE analysis of CBX7 expression screens. Due to cytotoxicity issues, cells containing the pLysS plasmid were not included.

3.10 Mutations to the F11 Residue Cause Higher-Order Structures

Due to their higher cation-π binding potential, the two tryptophan residues of CBX7 are expected to contribute more to binding than the F11 residue.[20] However, mutation of this

phenylalanine residue to alanine disrupts the H3K27me3/CBX7 interaction, suggesting that the F11 residue is important for binding.[1,21] Generally, mutations to solvent-exposed protein residues are not likely to cause structural abnormalities in the protein because the UAA side chain may protrude into the solvent while keeping most of the protein's intermolecular interactions intact. The F11 residue is solvent exposed in both the open and closed confirmations, making it a good candidate for UAA incorporation.

To start, we expressed the wild type, F11Y, and F11*p*NO_2F CBX7 chromodomain mutants. Size exclusion chromatography (SEC) revealed higher molecular radii species in the F11Y and F11*p*NO_2F samples. SEC chromatograms (Figure 3.8) showed varied elution profiles for the three CBX7 proteins. Preliminary ITC results suggested that the larger species in the F11Y peak does not bind its ligand ($N = 1.47 \times 10^{-5}$), however the F11*p*NO_2F sample does bind, albeit with lower affinity. To our knowledge, we are the first to express these F11 mutations and characterize their behavior. It is possible that the observed changes of the F11 mutations are due to the role of F11 in Kme3 recognition and dynamic formation of the aromatic cage. We are currently in the process of characterizing F11Y and F11*p*NO_2F species. Notably, higher molecular weight species occur even in the presence of the canonical F11Y mutation, and are therefore not due to UAA-mutagenesis, but fine-tuned alterations that occur upon mutation of F11. The fact that homologous proteins often feature a tyrosine at this location may indicate that CBX7 evolution selected for phenylalanine at this location for reasons of structural stability. However, due to these observed differences in oligomerization of the CBX7 reader protein upon mutation, we were prompted to select other proteins for targeted investigation.

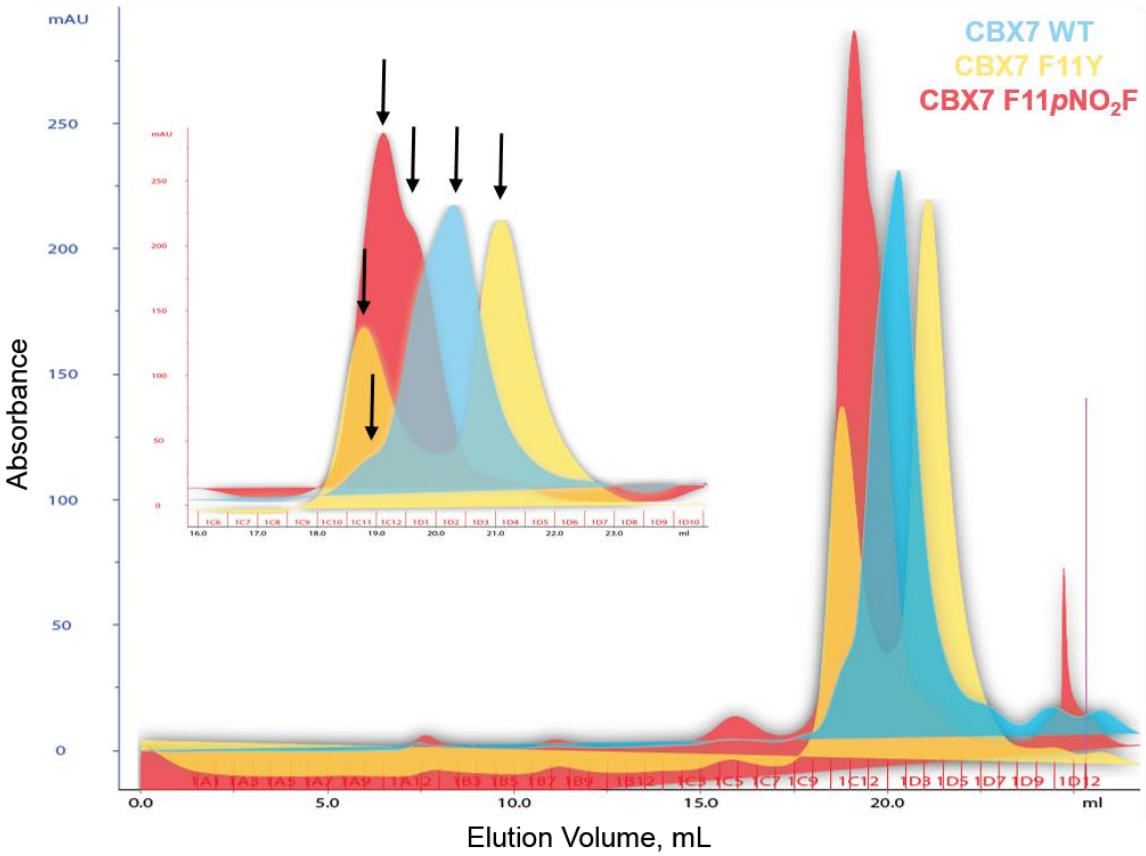


Figure 3.8 Size exclusion chromatograms of CBX7 chromodomains. Inset shows zoomed perspective of CBX7 peaks. CBX7 wild type is shown in blue, F11Y is shown in yellow, and F11*p*NO_2F is shown in red. Possible peaks are marked with arrows.

3.11 CBX5, a Mammalian HP1 Analog

CBX5 is another well-studied reader protein classified as a *Drosophila* HP1 homolog.[1,5] Not surprisingly, these proteins are involved in gene repression associated with heterochromatin.[1] Although not as well characterized, CBX5 has also been implicated in lung, colon, and breast cancer.[22] CBX5, unlike CBX1 or CBX3, is differentially expressed in cancerous and non-cancerous cells.[22] In breast and lung cancer, studies have shown that primary tumor cells show increased expression levels of CBX5.[22,23] CBX5 has also been implicated in formation of cancer stem cells, which generally bring worse prognoses.[22,23]

CBX5's aromatic cage contains one of each aromatic residue, the most diverse of the three cages studied. Interestingly, the 4 position of CBX5 (analogous to the Y24 position in HP1) also contains a tyrosine residue. The 28 position (Y48 analog), however, contains a phenylalanine. By studying CBX5's binding affinities, we may be able to determine any evolutionary relationships from these two homologs. A crystal structure of CBX5 has already been determined and based on NMR solution structures of the Apo-enzyme (PDB ID: 2RVL), the aromatic cage exhibits a less-drastic change upon ligand binding than CBX7. Inhibitors for CBX5 are also currently being developed by the Frye lab, which again allows us to probe cation-π binding with higher affinity, minimizing protein yield requirements.

Fortuitously, CBX5 expression required much less optimization than CBX7. Using the optimized glucose conditions for CBX7, CBX5 expression proved much less cytotoxic than the CBX7 counterparts. CBX5-UAA proteins were expressed in good yield (Figure 3.9) and are currently being characterized in preparation for ITC experiments. We are currently waiting on dimethyl and trimethyl derivatives of CBX5 inhibitors to probe cation-π binding.

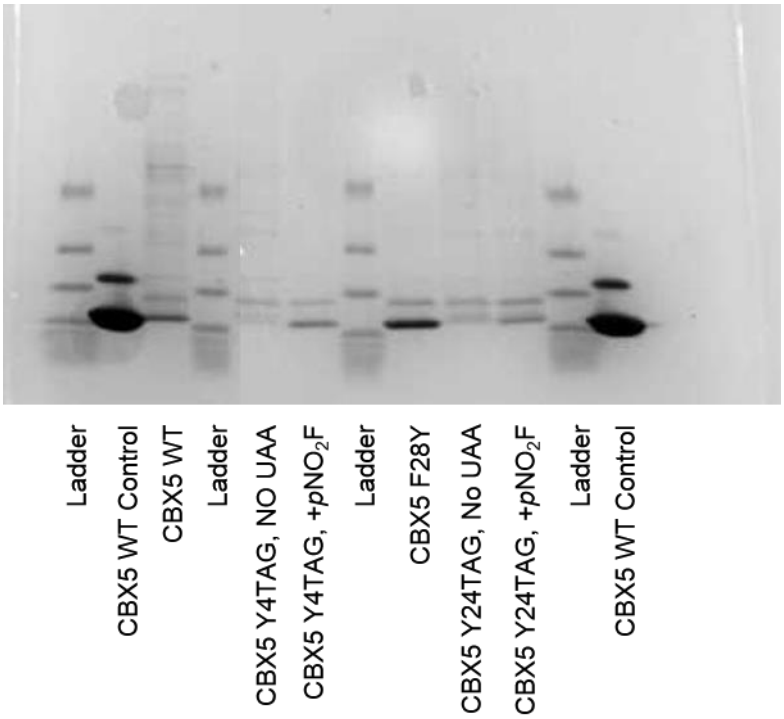


Figure 3.9 SDS-PAGE analysis of His-purified CBX5 mutants.

3.12 Discussion

While work is ongoing, this chapter sets the stage for *in vivo* unnatural amino acid mutagenesis in even more reader proteins of therapeutic interest. Rosetta™ 2 pLysS cells are used for expression of a variety of diverse reader protein chromodomains due to their cytotoxicity in *E. coli.* Prior to our work, UAA mutagenesis was impossible with the optimized pUltra vector due to incompatibility with the pLysS plasmid. Our method for expression of cytotoxic CBX proteins using catabolite repression allows for UAA incorporation into cytotoxic proteins. Unfortunately, the initial challenges associated with our mammalian reader studies have delayed our progress in characterizing the binding of UAA-variants with natural ligands and chemical probes. However, future directions will be focused on characterizing CBX-UAA variants that are now accessible as a result of the work

discussed earlier. Since CBX7 and CBX5 are both of significant therapeutic interest, these cation-π studies may glean information that facilitates the design of their inhibitors.

3.13 Experimental

3.13.1 Cloning, DNA Sequences, and Protein Sequences

pULTRA-*p*CNPheRS[24] was obtained from the lab of Dr. Peter Schultz and is also available from addgene (Plasmid # 48215). CBX7 and CBX5 genes were obtained from the lab of Dr. Stephen Frye. CBX7 and CBX5 were optimized for *E. coli* expression using the Codon Optimization Tool of Integrative DNA Technologies (IDT). The small size of the CBX's chromodomain (~70 residues plus additional for restriction enzyme sites and flanking sequences) allowed for the optimized gene to be purchased as a large oligonucleotide and cloned into the given vector using standard PCR techniques. Full genes were purchased as gBlocks® Gene Fragments from IDT and cloned into a pET11a vector using NdeI and BamHI restriction sites. Mutations to the gene were generated using standard overlap PCR. Oligonucleotides for PCR were obtained from IDT and enzymes and reagents used for cloning were obtained from New England BioLabs Inc. DNA sequences of cloned mutants from NdeI and BamHI restriction sites are shown below.

3.13.1.1 CBX7 Sequences

The underlined portion of the sequence is the CBX7 coding sequence, the 6X His tag is italicized, and the F11 position has been bolded for clarity. Mutations to the F11 position are shown in red.

CBX7 wild type codon optimized DNA sequence:

CATATGGAGCAGGTG**TTC**GCCGTAGAATCCATCCGCAAGAAACGTGTTCGCAAG
GGCAAAGTAGAATACCTTGTTAAGTGGAAGGGTTGGCCCCCGAAATACTCCACT
TGGGAACCAGAAGAACATATCTTAGACCCTCGTTTAGTAATGGCCTATGAGGAG
AAGGAGGAGCTTGAA*CATCACCACCACCACCAT*TGAGGATCC

CBX7 wild type protein sequence:

MEQV**F**AVESIRKKRVRKGKVEYLVKWKGWPPKYSTWEPEEHILDPRLVMAYEEKE
ELEHHHHHH

CBX7 F11Y codon optimized DNA sequence

CATATGGAGCAGGTG**TAC**GCCGTAGAATCCATCCGCAAGAAACGTGTTCGCAAG
GGCAAAGTAGAATACCTTGTTAAGTGGAAGGGTTGGCCCCCGAAATACTCCACT
TGGGAACCAGAAGAACATATCTTAGACCCTCGTTTAGTAATGGCCTATGAGGAG
AAGGAGGAGCTTGAA*CATCACCACCACCACCAT*TGAGGATCC

CBX7 F11Y protein sequence

MEQV**Y**AVESIRKKRVRKGKVEYLVKWKGWPPKYSTWEPEEHILDPRLVMAYEEKE
ELEHHHHHH

CBX7 F11TAG codon optimized DNA sequence:

CATATGGAGCAGGTG**TAG**GCCGTAGAATCCATCCGCAAGAAACGTGTTCGCAAG
GGCAAAGTAGAATACCTTGTTAAGTGGAAGGGTTGGCCCCCGAAATACTCCACT
TGGGAACCAGAAGAACATATCTTAGACCCTCGTTTAGTAATGGCCTATGAGGAG
AAGGAGGAGCTTGAA*CATCACCACCACCACCAT*TGAGGATCC

CBX7 F11TAG protein sequence (* denotes an UAA)

MEQV*****AVESIRKKRVRKGKVEYLVKWKGWPPKYSTWEPEEHILDPRLVMAYEEKE
ELEHHHHHH

3.13.1.2 CBX5 Sequences

The underlined portion of the sequence is the CBX5 coding sequence, the 6X His tag is italicized, and the Y4 and F28 position have been bolded for clarity. Mutations to the Y4 position are shown in red and mutations to the Y28 position are shown in blue.

CBX5 wild type codon optimized DNA sequence:

CATATG*CACCATCACCACCACCAC*TCGTCCGGGCGTGAAAACTTGTACTTCCAGGG
GGAGGAA**TAT**GTAGTCGAGAAAGTACTGGATCGCCGTGTTGTAAAGGGTCAAGT
GGAATACTTACTGAAGTGGAAGGGT**TTC**TCAGAGGAGCACAATACGTGGGAGCC
GGAAAAGAACCTTGACTGTCCTGAACTGATCTCAGAGTTTATGAAGAAGTATAA
GAAAATGAAAGAGTGAGGATCC

CBX5 wild type protein sequence:

MHHHHHHSSGRENLYFQGEE**Y**VVEKVLDRRVVKGQVEYLLKWKG**F**SEEHNTWEP

EKNLDC PELISEFMKKYKKMKE

CBX5 Y4F codon optimized DNA sequence:

CATATG*CACCATCACCACCACCAC*TCGTCCGGGCGTGAAAACTTGTACTTCCAGGG

GGAGGAA**TTT**GTAGTCGAGAAAGTACTGGATCGCCGTGTTGTAAAGGGTCAAGT

GGAATACTTACTGAAGTGGAAGGGT**TTC**TCAGAGGAGCACAATACGTGGGAGCC

GGAAAAGAACCTTGACTGTCCTGAACTGATCTCAGAGTTTATGAAGAAGTATAA

GAAAATGAAAGAGTGAGGATCC

CBX5 Y4F protein sequence:

MHHHHHHSSGRENLYFQGEE**F**VVEKVLDRRVVKGQVEYLLKWKG**F**SEEHNTWEP

EKNLDC PELISEFMKKYKKMKE

CBX5 Y4TAG codon optimized DNA sequence:

CATATG*CACCATCACCACCACCAC*TCGTCCGGGCGTGAAAACTTGTACTTCCAGGG

GGAGGAA**TAG**GTAGTCGAGAAAGTACTGGATCGCCGTGTTGTAAAGGGTCAAGT

GGAATACTTACTGAAGTGGAAGGGT**TTC**TCAGAGGAGCACAATACGTGGGAGCC

GGAAAAGAACCTTGACTGTCCTGAACTGATCTCAGAGTTTATGAAGAAGTATAA

GAAAATGAAAGAGTGAGGATCC

CBX5 Y4TAG protein sequence (* denotes an UAA):

MHHHHHHSSGRENLYFQGEE*VVEKVLDRRVVKGQVEYLLKWKG**F**SEEHNTWEPE

KNLDC PELISEFMKKYKKMKE

CBX5 F48Y codon optimized DNA sequence:

CATATG*CACCATCACCACCACCAC*TCGTCCGGGCGTGAAAACTTGTACTTCCAGGG

GGAGGAA**TAT**GTAGTCGAGAAAGTACTGGATCGCCGTGTTGTAAAGGGTCAAGT

GGAATACTTACTGAAGTGGAAGGGT**TAT**TCAGAGGAGCACAATACGTGGGAGCC

GGAAAAGAACCTTGACTGTCCTGAACTGATCTCAGAGTTTATGAAGAAGTATAA

GAAAATGAAAGAGTGAGGATCC

CBX5 F48Y protein sequence:

MHHHHHHSSGRENLYFQGEE**Y**VVEKVLDRRVVKGQVEYLLKWKG**Y**SEEHNTWEP

EKNLDC PELISEFMKKYKKMKE

CBX5 F28TAG codon optimized DNA sequence:

CATATG*CACCATCACCACCACCAC*TCGTCCGGGCGTGAAAACTTGTACTTCCAGGG

GGAGGAA**TAT**GTAGTCGAGAAAGTACTGGATCGCCGTGTTGTAAAGGGTCAAGT

GGAATACTTACTGAAGTGGAAGGGT**TAG**TCAGAGGAGCACAATACGTGGGAGC

CGGAAAAGAACCTTGACTGTCCTGAACTGATCTCAGAGTTTATGAAGAAGTATA

AGAAAATGAAAGAGTGAGGATCC

CBX5 F28TAG protein sequence (* denotes an UAA):

MHHHHHHSSGRENLYFQGEE**Y**VVEKVLDRRVVKGQVEYLLKWKG*SEEHNTWEP EKNLDC PELISEFMKKYKKMKE

3.13.2 CBX7 Protein Expression and Optimization

For UAA-CBX7 variants, pET11a-CBX7-F11TAG was co-transformed with pUltra-*p*CNPheRS into Bl21-Gold (DE3) competent cells (Agilent Technologies). For wild type and tyrosine mutations, pET11a vectors containing the desired gene were transformed into BL21-Gold (DE3) competent cells. Cells were rescued with 1 mL SOC broth and then incubated for 45 min at 37°C with shaking. 50 uL of each rescue was plated as follows: wild type and F11Y mutants on LB ampicillin (100 mg/L) agar plates; TAG mutants co-transformed with pUltra-*p*CNPheRS on LB ampicillin (100 mg/L) and spectinomycin (50 mg/L) agar plates. Plates were incubated overnight at 37°C. Single colonies from the transformation plates were used to inoculate LB with appropriate antibiotics in baffled flasks (flask volume <4X larger than LB volume). Cultures were grown to saturation overnight at 37°C with shaking at 225 RPM.

Initially, CBX7 proteins were expressed using Rosetta™ 2 (DE3) pLysS cells (Novagen), LB media and traditional IPTG induction as reported previously.[7] However, the Rosetta™ cells were incompatible with the pUltra-*p*CNPheRS vector. Large scale expressions required up to 24 hours to achieve $OD_{600} \approx 0.6$, whereas the wild type expressions took under 4 hours. Cells containing pUltra-*p*CNPheRS alone behaved abnormally, suggesting cytotoxicity was not due to CBX7 expression. Cells containing pUltra-*p*CNPheRS were unable to express either UAA-CBX7, CBX7 WT, or HP1 WT.

After transforming the appropriate plasmid into Rosetta™ 2 (DE3) pLysS, Rosetta™ 2 (DE3), BL21 (DE3) pLysS, and Bl21 (DE3), and DH10B cells, we tested to see which cell lines gave the best yields. A summary of cell line genotypes can be found below. The conditions selected for expression screening can be found in SI Table 3.1–3.3.

BL21(DE3)

F–ompT hsdSB(rB–mB–)gal dcm (DE3)

BL21(DE3) pLysS

F–ompT hsdSB(rB–mB–)gal dcm (DE3) pLysS (CamR)

Rosetta 2(DE3)

F–ompT hsdSB(rB–mB–)gal dcm (DE3) pRARE2(CamR)

Rosetta2(DE3) pLysS

F–ompT hsdSB(rB–mB–)gal dcm(DE3) pLysS pRARE2(CamR)

Each condition was expressed in 25 mL LB in 125 mL Ultra Yield Flasks™ (Thomson Instrument Company). Cells were grown to $OD_{600} \approx 0.6$ and induced with 500 uL of 0.5 M IPTG or 0.05% w/v (DH10B). Cultures were left to express overnight, pelleted, lysed, and clarified by centrifugation (as described in section 3.13.4). The clarified lysate was incubated on Ni Sepharose 6 Fast Flow Resin (GE) and left on a tube rocker overnight at 4°C. Ni-tagged proteins were purified off the resin according to the resin manual using the

HisTrap purification buffers described in section 3.13.4. The lysates and Ni-eluted samples were then visualized using SDS-PAGE (Figure 3.7).

To ensure glucose was present in all media prior to IPTG induction, 1% glucose was added to all LB agar plates and LB media for overnight growths. All proteins were expressed in 2.5 L Ultra Yield Flasks™ (Thomson Instrument Company) containing 500 mL of LB media supplemented with 5 mM $MgCl_2$, 5 mM $MgSO_4$, 1% glucose, and 1:5000 dilution of Antifoam 204 to increase oxygen uptake and prevent foaming over. Each flask also contained appropriate antibiotics (100 mg/L ampicillin (pET-CBX7s), 50 mg/L spectinomycin (pUltra-*p*CNPheRS)). For wild type and F11Y mutants, media was inoculated with 2.5 mL of saturated overnight culture. For F11TAG expressions, media was inoculated with 5 mL of saturated overnight culture to account for the slower initial growth in the presence of two antibiotics. After inoculation, cultures were incubated at 37°C with 310-350 RPM shaking until reaching an OD_{600} between 0.6-0.8. Dry UAA (Chem Impex International) was added to the appropriate TAG cultures (2.5 mmol UAA for 5 mM final concentrations. *p*NO_2F was increased to 10 mmol for 20 mM final concentration to compensate for lower affinity of the *p*CNPheRS for *p*NO_2Phe). Incubator temperature was then dropped to 16°C and the cultures were left to express for 16-24 hours. For expressions containing Y24*p*NO_2Phe, the incubator was covered with aluminum foil to prevent light degradation of *p*NO_2Phe.

After expression, cultures were pelleted at 4500 RPM for 10 min and the supernatant was decanted. Cell pellets were frozen overnight at -20°C and resuspended in 20 mL lysis buffer (50 mM sodium phosphate, pH 7.2, 500 mM NaCl, 30 mM imidazole, 0.25 mg/mL lysozyme, 1 mM phenylmethanesulfonyl fluoride, with cOmplete EDTA-Free Protease Inhibitor Cocktail Tablets (Roche)). The resuspended pellet was incubated at 37°C with 225

RPM shaking for 30 min and cooled on ice for 10 min. Pellets were sonicated on ice for 7.5 min (20% amplitude, 0.5 seconds on, 0.5 seconds off) until the lysate appeared homogenous. Lysate was clarified by centrifugation (19,000 RPM, Sorvall SS-34 rotor) for 45 min. Supernatant was decanted and filtered through a 0.45 um syringe filter.

3.13.3 CBX5 Protein Expression

CBX5 was found to express well in the optimized CBX7 conditions. CBX5 expression conditions can be found in section 3.13.2.

3.13.4 Protein Purification

Filtered lysate was purified on an ÄKTAPurifier UPC 10 (GE) equipped with a HisTrap-5mL HP column (GE). Proteins were 6XHis-tag purified using the buffers previously described and eluted using a linear gradient from 0–100% buffer B.[9] Eluted fractions were pooled and concentrated on a 3 kDa Amicon Ultra-15 Centrifugal filter. The concentrated His-tag-purified sample was purified by size exclusion chromatography using a Superdex 200 10/300 GL size exclusion column equilibrated in SEC buffer (50 mM Tris, pH 7.5, 250 mM NaCl, 2 mM DTT, 5% w/v glycerol). Eluted fractions (eluted at ~20 mL) were pooled, concentrated, and quantified using a Cary 100 UV/Vis Spectrophotometer (Agilent Technologies). Extinction coefficients for UAA proteins were calculated by measuring the extinction coefficient of each free amino acid in solution and adding the free UAA extinction coefficient to the extinction coefficient of the wild type CBX protein with the native residue removed. The extinction coefficient of wild type protein with native amino acid removed was calculated using the Scripps Protein Calculator (http://protcalc.sourceforge.net/). Extinction

coefficients are provided in SI Table 3.4. Protein purity and UAA incorporation was assessed using SDS-PAGE and ESI-LCMS (Figure 3.7, Figure 3.9, SI Figure 3.1–3.4).

3.13.5 ESI-LCMS Confirmation of UAA Incorporation

1 mL of a 10 μM solution of each protein was exchanged into HPLC-grade water using a 3 kDa Amicon Ultra-15 centrifugal filter and then filtered through glass wool. The samples were run on an Agilent 6520 Accurate-Mass Q-TOF ESI positive LCMS (SI Table 3.5) using the method described in SI Table 3.6. All LCMS chromatograms show evidence of the appropriate UAA incorporation with no detectible canonical amino acid contamination (SI Table 3.7). LCMS chromatograms from each CBX variant can be found in SI Figure 3.1–3.4. Although incorporation of tyrosine or phenylalanine can be detected in TAG mutants expressed in the absence of unnatural amino acid, no evidence of tyrosine or phenylalanine incorporation is detected in the presence of UAA (SI Figure 3.4).

3.13.6 CBX Compounds

All CBX inhibitors were synthesized by Kelsey Lamb of the Frye lab.

3.13.7 Isothermal Titration Calorimetry (ITC) Binding Measurements

Protein and compound were both prepared in buffer (25 mM Tris, pH 8, 150 mM NaCl, 2 mM β-mercaptoethanol) and then diluted down with additional buffer to 50 μM protein and 0.5 mM compound. ITC experiments were performed by titrating compound into CBX7 at 25°C using a Microcal AutoITC200.[7] Protein concentrations were determined by measuring absorbance at 280 nm on a Cary 100 UV/Vis Spectrophotometer (Agilent

Technologies). Heat of dilution was accounted for by subtracting the endpoint ΔH value from each prior injection. Data was analyzed using the One-Site binding model supplied in Origin software. While the binding stoichiometry is known to be 1:1, at the high concentrations used here active protein concentration may differ from measured concentration. Although we did not collect enough data for cation-π LFER plots, preliminary data is included in SI Table 3.8. The data are within error of each other, which prevents us from drawing any meaningful conclusions. ITC binding curves can be found in SI Figure 3.5–3.10.

3.14 Supplementary Information

3.14.1 Supplementary Tables

SI Table 3.1 CBX7 expression condition screening, part I

#	Plasmids	Cell Line	Notes	Antibiotics	Glucose
1	pUltra-*p*CNPheRS alone	Bl21(DE3)	Control	Strep	None
2	pET11a-HP1 WT	Bl21(DE3)	Positive control	Amp	None
3	pET11a-HP1 WT + pUltra-*p*CNPheRS	Bl21(DE3)	Cytotoxicity Control	Amp Strep	None
4	pET11a-CBX7-WT	Bl21(DE3)	Negative Control (requires Rosetta cells)	Amp	None
5	pET11a-CBX7-WT-Opt_Ecoli	Bl21(DE3)		Amp	None
6	pET11a-CBX7-F11TAG-opt + pUltra-*p*CNPheRS	Bl21(DE3)	UAA Added (*p*CH_3F)	Amp Strep	None
7	pET11a- CBX7-F11TAG-opt + pUltra-*p*CNPheRS	Bl21(DE3)	No UAA Added	Amp Strep	None
8	pUltra-*p*CNPheRS alone	Bl21(DE3)	Control	Strep	1%
9	pET11a-HP1 WT	Bl21(DE3)	Positive control	Amp	1%
10	pET11a-HP1 WT + pUltra-*p*CNPheRS	Bl21(DE3)	Cytotoxicity Control	Amp Strep	1%
11	pET11a-CBX7-WT	Bl21(DE3)	Negative Control (requires Rosetta cells)	Amp	1%
12	pET11a-CBX7-WT-Opt_Ecoli	Bl21(DE3)		Amp	1%
13	pET11a-CBX7-F11TAG-opt + pUltra-*p*CNPheRS	Bl21(DE3)	UAA Added (*p*CH_3F)	Amp	1%
14	pET11a- CBX7-F11TAG-opt + pUltra-*p*CNPheRS	Bl21(DE3)	No UAA Added	Amp Strep	1%
15	pUltra-*p*CNPheRS alone	Bl21(DE3) pLysS	Control	Cam Strep	None
16	pET11a-HP1 WT	Bl21(DE3) pLysS	Positive control	Amp Cam	None
17	pET11a-HP1 WT + pUltra-*p*CNPheRS	Bl21(DE3) pLysS	Cytotoxicity Control	Amp Cam Strep	None
18	pET11a-CBX7-WT	Bl21(DE3) pLysS	Negative Control (requires Rosetta cells)	Amp Cam	None
19	pET11a-CBX7-WT-Opt_Ecoli	Bl21(DE3) pLysS		Amp Cam	None
20	pET11a-CBX7-F11TAG-opt + pUltra-*p*CNPheRS	Bl21(DE3) pLysS	UAA Added (*p*CH_3F)	Amp Cam Strep	None
21	pET11a- CBX7-F11TAG-opt + pUltra-*p*CNPheRS	Bl21(DE3) pLysS	No UAA Added	Amp Cam Strep	None

SI Table 3.2 CBX7 expression condition screening, part II

#	Plasmids	Cell Lines	Notes	Antibiotics	Glucose
22	pUltra-*p*CNPheRS alone	Rosetta2(DE3)	Control	Strep	None
23	pET11a-HP1 WT	Rosetta2(DE3)	Positive control	Amp	None
24	pET11a-HP1 WT + pUltra-*p*CNPheRS	Rosetta2(DE3)	Cytotoxicity Control	Amp Strep	None
25	pET11a-CBX7-WT	Rosetta2(DE3)	Positive Control (requires Rosetta cells)	Amp	None
26	pET11a-CBX7-WT-Opt_Ecoli	Rosetta2(DE3)		Amp	None
27	pET11a-CBX7-F11TAG-opt + pUltra-*p*CNPheRS	Rosetta2(DE3)	UAA Added (*p*CH_3F)	Amp Strep	None
28	pET11a- CBX7-F11TAG-opt + pUltra-*p*CNPheRS	Rosetta2(DE3)	No UAA Added	Amp Strep	None
29	pUltra-*p*CNPheRS alone	Rosetta2(DE3)	Control	Strep	1%
30	pET11a-HP1 WT	Rosetta2(DE3)	Positive control	Amp	1%
31	pET11a-HP1 WT + pUltra-*p*CNPheRS	Rosetta2(DE3)	Cytotoxicity Control	Amp Strep	1%
32	pET11a-CBX7-WT	Rosetta2(DE3)	Positive Control (requires Rosetta cells)	Amp	1%
33	pET11a-CBX7-WT-Opt_Ecoli	Rosetta2(DE3)		Amp	1%
34	pET11a-CBX7-F11TAG-opt + pUltra-*p*CNPheRS	Rosetta2(DE3)	UAA Added (*p*CH_3F)	Amp Strep	1%
35	pET11a- CBX7-F11TAG-opt + pUltra-*p*CNPheRS	Rosetta2(DE3)	No UAA Added	Amp Strep	1%
36	pUltra-*p*CNPheRS alone	Rosetta2(DE3) pLysS	Control	Cam Strep	None
37	pET11a-HP1 WT	Rosetta2(DE3) pLysS	Positive control	Amp Cam	None
38	pET11a-HP1 WT + pUltra-*p*CNPheRS	Rosetta2(DE3) pLysS	Cytotoxicity Control	Amp Cam Strep	None
39	pET11a-CBX7-WT	Rosetta2(DE3) pLysS	Positive Control (requires Rosetta cells)	Amp Cam	None
40	pET11a-CBX7-WT-Opt_Ecoli	Rosetta2(DE3) pLysS		Amp Cam	None
41	pET11a-CBX7-F11TAG-opt + pUltra-*p*CNPheRS	Rosetta2(DE3) pLysS	UAA Added (*p*CH_3F)	Amp Cam Strep	None
42	pET11a- CBX7-F11TAG-opt + pUltra-*p*CNPheRS	Rosetta2 DE3 pLysS	No UAA Added	Amp Cam Strep	None

SI Table 3.3 CBX7 expression condition screening, part III

#	Plasmids	Cell Lines	Notes	Antibiotics	Glucose
43	pBad-CBX7-F11TAG + pDule-*p*CNPheRS	DH10B	UAA Added (*p*CH_3F)	Amp Tet	None
44	pBad-CBX7-F11TAG + pDule-*p*CNPheRS	DH10B	No UAA Added	Amp Tet	None
45	pBad-CBX7-F11TAG + pDule-*p*CNPheRS	DH10B	UAA Added (*p*CH_3F)	Amp Tet	1%
46	pBad-CBX7-F11TAG + pDule-*p*CNPheRS	DH10B	No UAA Added	Amp Tet	1%

SI Table 3.4 Extinction coefficients for UAAs and CBX7 variants

Mutant Name	Extinction Coefficient at 280 nm ($cm^{-1}M^{-1}$)	Molecular Weight (Da)
Wild type	20970	7943.08
F11Y	22190	7959.08
F11*p*CNF	21579	7986.11
F11*p*NO_2F	29227	8006.07
F11*p*CH_3F	21043	7975.14
F11*p*CF_3F	20914	8029.10
UAA	**Free UAA Extinction Coefficient at 280 nm ($cm^{-1}M^{-1}$)**	**Molecular Weight (Da)**
***p*CNF**	669.42	190.20
***p*NO_2F**	8317.26	210.16
***p*CH_3Phe**	132.76	179.23
***p*CF_3Phe**	4.39	233.19

SI Table 3.5 ESI-LCMS instrument information

Column	Restek Viva C4 5 µm 150 x 2.1 mm
Solvent A	0.1 % formic acid in water
Solvent B	0.1 % formic acid in acetonitrile
Temperature	35°C
Ion Source	Dual ESI
Ion Polarity	Positive
Abs. Threshold	200
Rel threshold (%)	0.01
Cycle Time	1 s
Gas Temp	350 °C
Drying gas	12 l/min
Nebulizer	50 psig
Fragmentor	200 V
Skimmer	65 V
OCT 1 RF VPP	750
Min Mass Range	100 m/z
Max Mass Range	3200 m/z
Acquisition Rate	1 spectra/s
Acquisition time	1000.2 ms/spectrum
Transients/spectrum	9898

SI Table 3.6 ESI-LCMS method information

Solvent A	Water
Solvent B	Acetonitrile
Flowrate	0.3 mL/min
Gradient	
Time (min)	%B
0	5
15	95
20	95
20.01	5
25	5

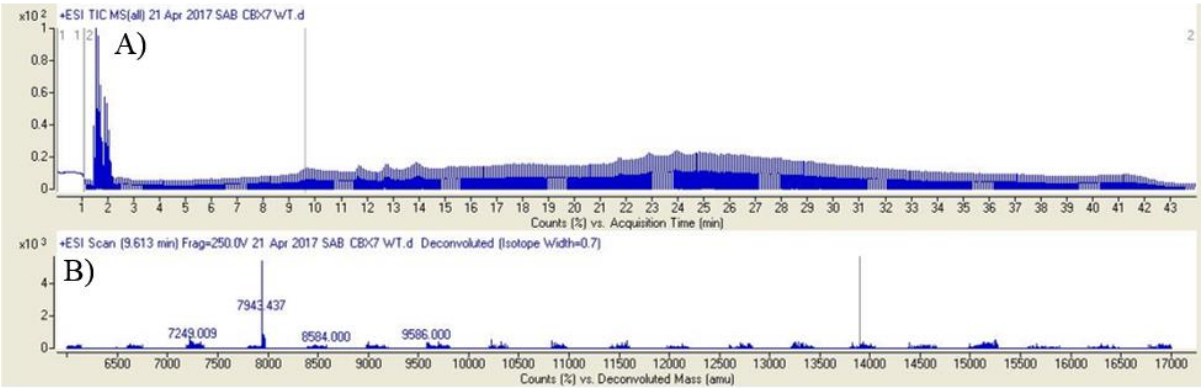
SI Table 3.7 ESI-LCMS data verifies UAA-incorporation

Sample	Expected Mass (Da)	Observed Mass (Da)	Difference (Da)	% Difference
Wild type	7943.08	7943.44	0.36	4.5×10^{-5}
F11Y	7959.07	7959.52	0.45	5.7×10^{-5}
F11*p*CH$_3$F	7957.23	7957.63	0.40	5.0×10^{-5}

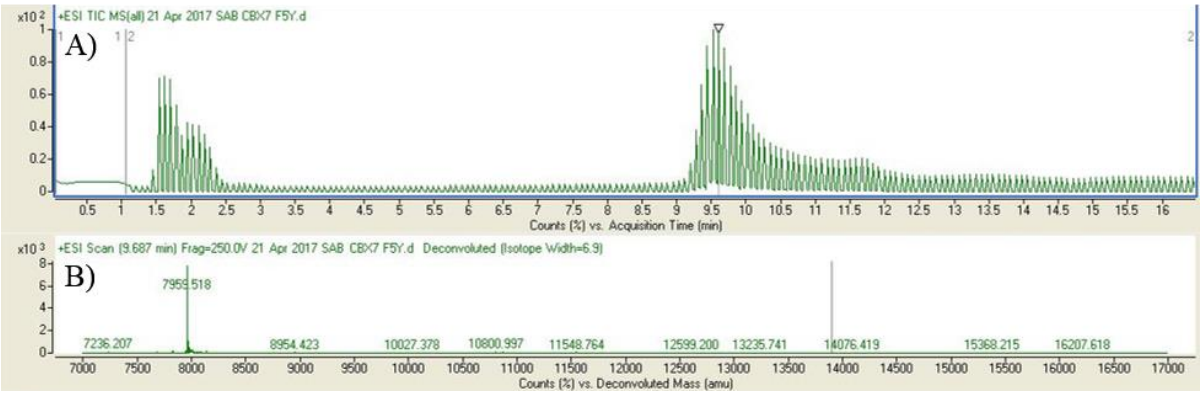
SI Table 3.8 ITC data for CBX7 mutants

Protein	Compound	Number of runs	Average K_d (nM)	St Dev K_d (nM)
Wild type	UNC4938	2	230	33
F11Y	UNC4938	3	231	41
F11Y Oligo	UNC4938	1	388	N/A
F11*p*NO$_2$F	UNC4938	1	274	N/A
WT	UNC5352	2	156	48
F11Y	UNC5352	4	182	34

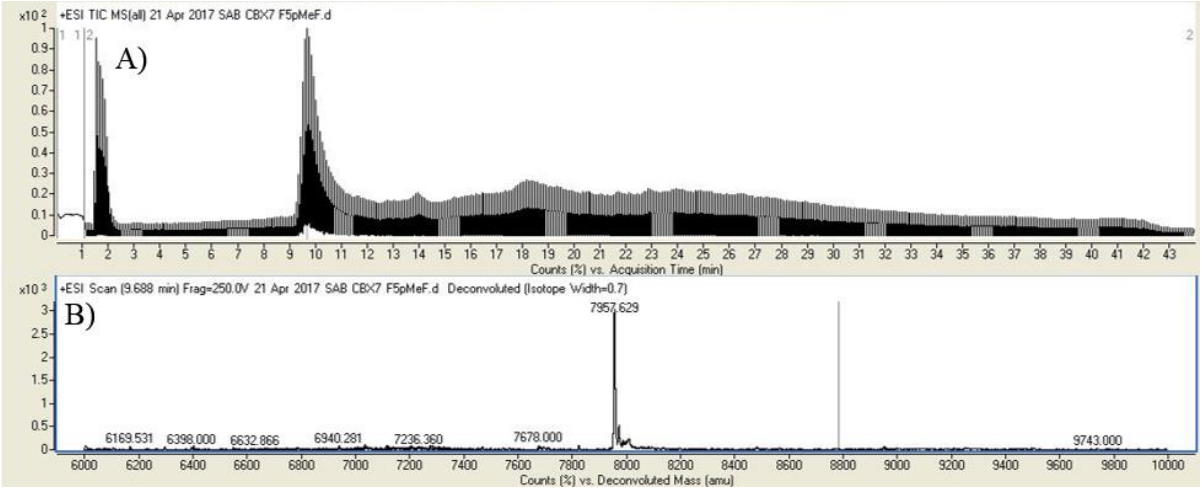
3.14.2 Supplementary Figures



SI Figure 3.1 LCMS of CBX7 wild type. TIC scan (A) and corresponding m/z deconvolution (B).TIC signal at 9.6 min lower than other spectra due to over-dilute solution.

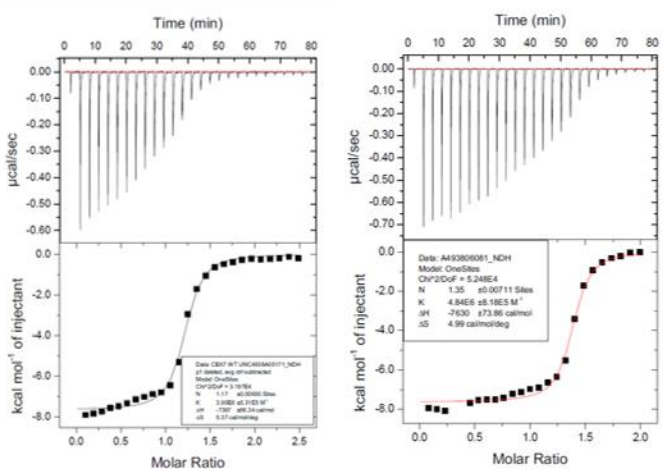


SI Figure 3.2 LCMS of CBX7 F11Y. TIC scan (A) and corresponding m/z deconvolution (B). Run was stopped prematurely, as evidencd by X-axis in (A).

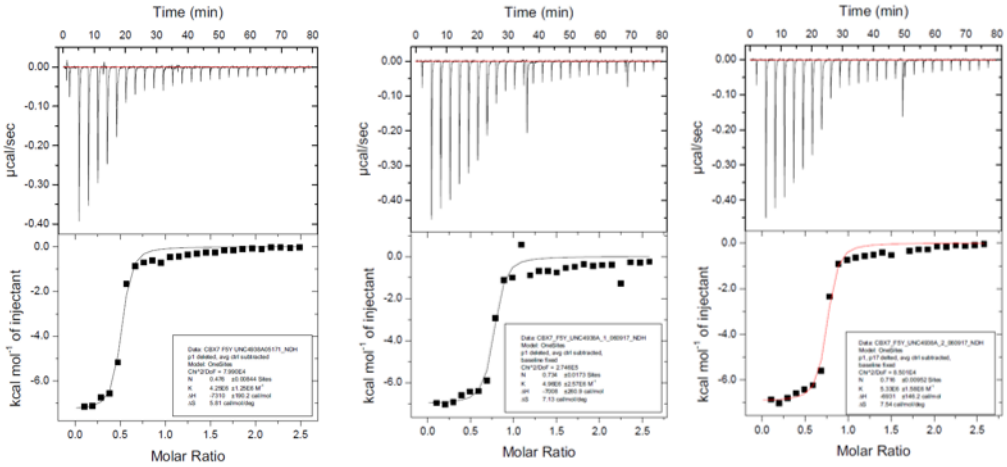


SI Figure 3.3 LCMS of CBX7 F11*p*CH_3F. TIC scan (A) and corresponding m/z deconvolution (B).

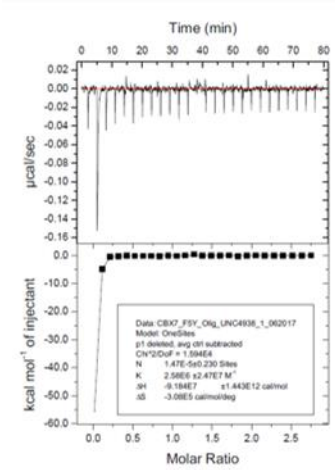
SI Figure 3.4 LCMS of CBX7 F11TAG with no UAA added. Only TIC scan included as deconvolution could not identify any protein in the given mass range.



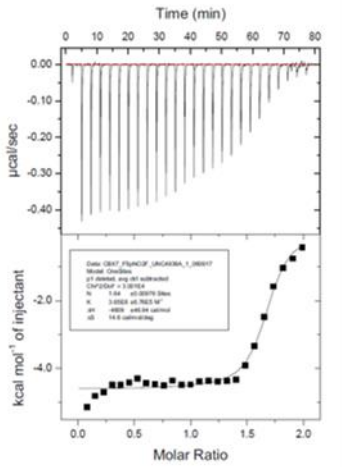
SI Figure 3.5 ITC curves of UNC4938 binding to wild type CBX7.



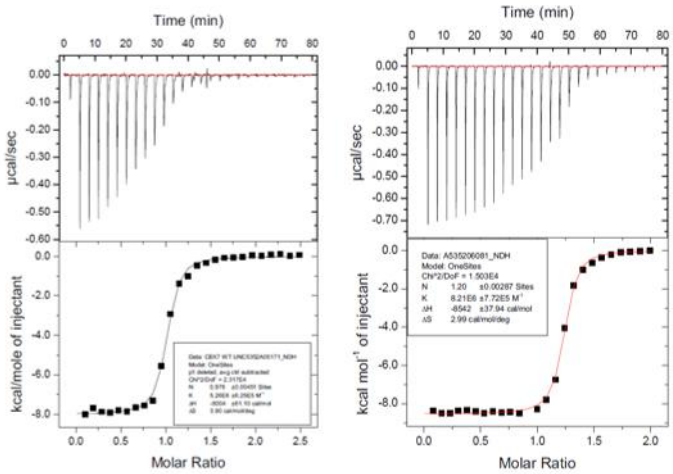
SI Figure 3.6 ITC curves of UNC4938 binding to CBX7 F11Y.



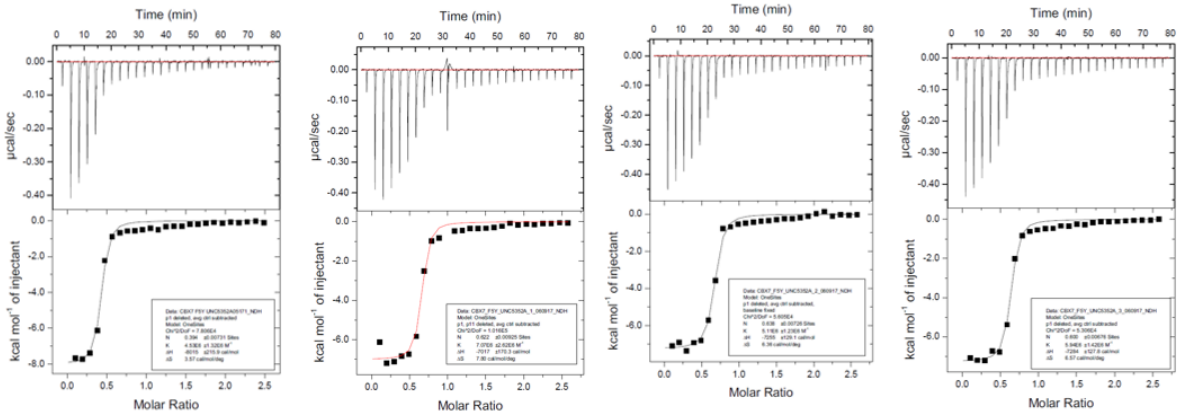
SI Figure 3.7 ITC curve of UNC4938 binding to CBX7 F11Y oligomer species.



SI Figure 3.8 ITC curve of UNC4938 binding to CBX7 F11*p*NO_2F.



SI Figure 3.9 ITC curves of UNC5352 binding to wild type CBX7.



SI Figure 3.10 ITC curves of UNC5352 binding to CBX7 F11Y.

REFERENCES

(1) Kaustov, L.; Ouyang, H.; Amaya, M.; Lemak, A.; Nady, N.; Duan, S.; Wasney, G. A.; Li, Z.; Vedadi, M.; Schapira, M.; Min, J.; Arrowsmith, C. H. *J. Biol. Chem.* **2011**, *286* (1), 521–529.

(2) Gil, J.; Bernard, D.; Martínez, D.; Beach, D. *Nat. Cell Biol.* **2004**, *6* (1), 67–72.

(3) Bernstein, E.; Duncan, E. M.; Masui, O.; Gil, J.; Heard, E.; Allis, C. D. *Mol. Cell. Biol.* **2006**, *26* (7), 2560–2569.

(4) Ma, R.; Zhang, Y.; Sun, T.; Cheng, B. *J. Zhejiang Univ. Sci. B* **2014**, *15* (5), 412–418.

(5) Kwon, S. H.; Workman, J. L. *BioEssays* **2011**, *33* (4), 280–289.

(6) Liu, L.; Zhen, X. T.; Denton, E.; Marsden, B. D.; Schapira, M. *Bioinformatics* **2012**, *28* (16), 2205–2206.

(7) Stuckey, J. I.; Dickson, B. M.; Cheng, N.; Liu, Y.; Norris, J. L.; Cholensky, S. H.; Tempel, W.; Qin, S.; Huber, K. G.; Sagum, C.; Black, K.; Li, F.; Huang, X.-P.; Roth, B. L.; Baughman, B. M.; Senisterra, G.; Pattenden, S. G.; Vedadi, M.; Brown, P. J.; Bedford, M. T.; Min, J.; Arrowsmith, C. H.; James, L. I.; Frye, S. V. *Nat. Chem. Biol.* **2016**, *12* (3), 180–187.

(8) Aranda, S.; Mas, G.; Di Croce, L. *Sci. Adv.* **2015**, *1* (11), e1500737.

(9) Stuckey, J. I.; Simpson, C.; Norris-Drouin, J. L.; Cholensky, S. H.; Lee, J.; Pasca, R.; Cheng, N.; Dickson, B. M.; Pearce, K. H.; Frye, S. V.; James, L. I. *J. Med. Chem.* **2016**, *59* (19), 8913–8923.

(10) Di Croce, L.; Helin, K. *Nat. Struct. Mol. Biol.* **2013**, *20* (10), 1147–1155.

(11) Kamps, J. J. A. G.; Huang, J.; Poater, J.; Xu, C.; Pieters, B. J. G. E.; Dong, A.; Min, J.; Sherman, W.; Beuming, T.; Matthias Bickelhaupt, F.; Li, H.; Mecinović, J. *Nat. Commun.* **2015**, *6*, 8911.

(12) Milosevich, N.; Hof, F. *Biochemistry* **2016**, *55* (11), 1570–1583.

(13) Konze, K. D.; Ma, A.; Li, F.; Barsyte-Lovejoy, D.; Parton, T.; MacNevin, C. J.; Liu, F.; Gao, C.; Huang, X. P.; Kuznetsova, E.; Rougie, M.; Jiang, A.; Pattenden, S. G.; Norris, J. L.; James, L. I.; Roth, B. L.; Brown, P. J.; Frye, S. V.; Arrowsmith, C. H.; Hahn, K. M.; Wang, G. G.; Vedadi, M.; Jin, J. *ACS Chem. Biol.* **2013**, *8* (6), 1324–1334.

(14) Grasso, C. S.; Tang, Y.; Truffaux, N.; Berlow, N. E.; Liu, L.; Debily, M.-A.; Quist, M. J.; Davis, L. E.; Huang, E. C.; Woo, P. J.; Ponnuswami, A.; Chen, S.; Johung, T. B.; Sun, W.; Kogiso, M.; Du, Y.; Qi, L.; Huang, Y.; Hütt-Cabezas, M.; Warren, K. E.; Le Dret, L.; Meltzer, P. S.; Mao, H.; Quezado, M.; van Vuurden, D. G.; Abraham, J.; Fouladi, M.; Svalina, M. N.; Wang, N.; Hawkins, C.; Nazarian, J.; Alonso, M. M.; Raabe, E. H.; Hulleman, E.; Spellman, P. T.; Li, X.-N.; Keller, C.; Pal, R.; Grill, J.; Monje, M. *Nat. Med.* **2015**, *21* (6), 555–559.

(15) Kersemans, K.; Mertens, J.; Caveliers, V. *J. Label. Compd. Radiopharm.* **2010**, *53*, 58–62.

(16) Novy, R.; Drott, D.; Yaeger, K.; Mierendorf, R. *Innovations* **2001**, *12*, 4–6.

(17) Novy, R.; Morris, B. *Innovations* **2001**, *13* (1), 13–15.

(18) Rosano, G. L.; Ceccarelli, E. A. *Front. Microbiol.* **2014**, *5* (APR), 1–17.

(19) Hammill, J. T.; Miyake-Stoner, S.; Hazen, J. L.; Jackson, J. C.; Mehl, R. A. *Nat. Protoc.* **2007**, *2* (10), 2601–2607.

(20) Davis, M. R.; Dougherty, D. A. *Phys. Chem. Chem. Phys.* **2015**, *17* (43), 29262–29270.

(21) Chen, C.; Jin, J.; James, D. A.; Adams-Cioaba, M. A.; Park, J. G.; Guo, Y.; Tenaglia, E.; Xu, C.; Gish, G.; Min, J.; Pawson, T. *Proc. Natl. Acad. Sci. U. S. A.* **2009**, *106* (48), 20336–20341.

(22) Vad-Nielsen, J.; Nielsen, A. L. *Cancer Biol. Ther.* **2015**, *16* (2), 189–200.

(23) Yu, Y.-H.; Chiou, G.-Y.; Huang, P.-I.; Lo, W.-L.; Wang, C.-Y.; Lu, K.-H.; Yu, C.-C.; Alterovitz, G.; Huang, W.-C.; Lo, J.-F.; Hsu, H.-S.; Chiou, S.-H. *Sci. Rep.* **2012**, *2* (584), 1–9.

(24) Young, D. D.; Young, T. S.; Jahnz, M.; Ahmad, I.; Spraggon, G.; Schultz, P. G. *Biochemistry* **2011**, *50* (11), 1894–1900.

CHAPTER 4: GENETIC ENCODING OF COFACTOR-LIKE UNNATURAL AMINO ACIDS

4.1 Enzymes are Efficient and Selective Biocatalysts

Enzymes are efficient and selective biocatalysts.[1] Having evolved over billions of years, enzymes are selective for their substrate and proficient for their given purpose. Enzymes greatly increase reaction rates while enforcing high regio-, diastereo- and enantioselectivity. In addition, enzymes are often capable of functioning under mild reaction conditions such as room temperature, atmospheric pressure, oxygenated environments, and in aqueous media. All of these factors make enzymes attractive complements to small-molecule catalysts.[2] However, in order to compete with small-molecule catalysis and be of industrial importance, enzymes often require modification. These alterations increase substrate scope, resistance to harsh conditions including high temperature, and performance in organic solvents. As a result of the diverse synthetic application of enzymes, there has been increased demands for enzymes with improved activity as well as novel catalytic functions. These enzymes are often produced using directed evolution.

Directed evolution is a two-step process that mimics natural evolution.[2] First, genetic diversity is introduced to the enzyme via *in vitro* recombination and/or random mutagenesis. Second, the desired function is selected or screened from the newly diversified gene pool.[1,2] Directed evolution has been used to increase solvent tolerance and thermal stability.[3,4] It has also been used to introduce specificity for new substrates or alter enzyme

enantioselectivity.[3,4] Additionally, directed evolution may be used to introduce new chemistries into the enzyme active site.[5]

4.2 Cofactors Expand Catalysis in Nature

Although scientists have expanded enzyme catalysis through directed evolution, nature has expanded its own enzymatic toolkit through the evolved use of cofactors. Cofactors are accessory small molecules, including heme, thiamine, pyridoxal phosphate, or metal ions, that are recruited to the active site for catalysis.[6] By recruiting a cofactor to the active site, enzymes may exploit additional chemistries not encoded by the 20 canonical amino acids. Cofactor-containing enzymes often play vital roles in cellular functions, and as a result cofactor deficiencies can be linked to disease.[7]

4.3 Thiamine is an Essential Cofactor

Thiamine diphosphate (ThDP), or vitamin B1, is a cofactor required for carbohydrate metabolism.[8] Thiamine contains three basic chemical moieties: a thiazolium ring, an aminopyrimidine ring, and a diphosphate-capped side chain (Figure 4.1).[9] The catalytic core of thiamine is the thiazole ring; deprotonation of the C_2 carbon forms a catalytically active ylide (Figure 4.2). Unlike many cofactors, thiamine is a true catalyst: the active ylide is regenerated at the end of the catalytic cycle (Figure 4.3).[10] Catalysts enhance reaction rates by stabilizing transition states, and for ThDP, this is due to resonance stabilization of the carbanion intermediate. The resonance-stabilized structure is known as the Breslow Intermediate, named for Ronald Breslow, who proposed the catalytic cycle in 1958.[8]

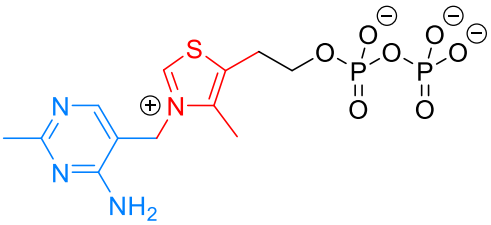


Figure 4.1 Structure of thiamine diphosphate. The aminopyrimidine ring is shown in blue, the thiazolium ring in red, and the diphosphate side chain in black.

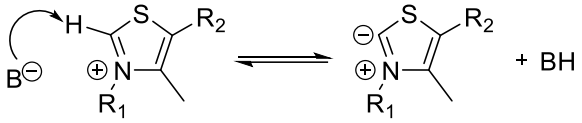


Figure 4.2 Formation of the ylide, the catalytically active form of the thiazole ring.

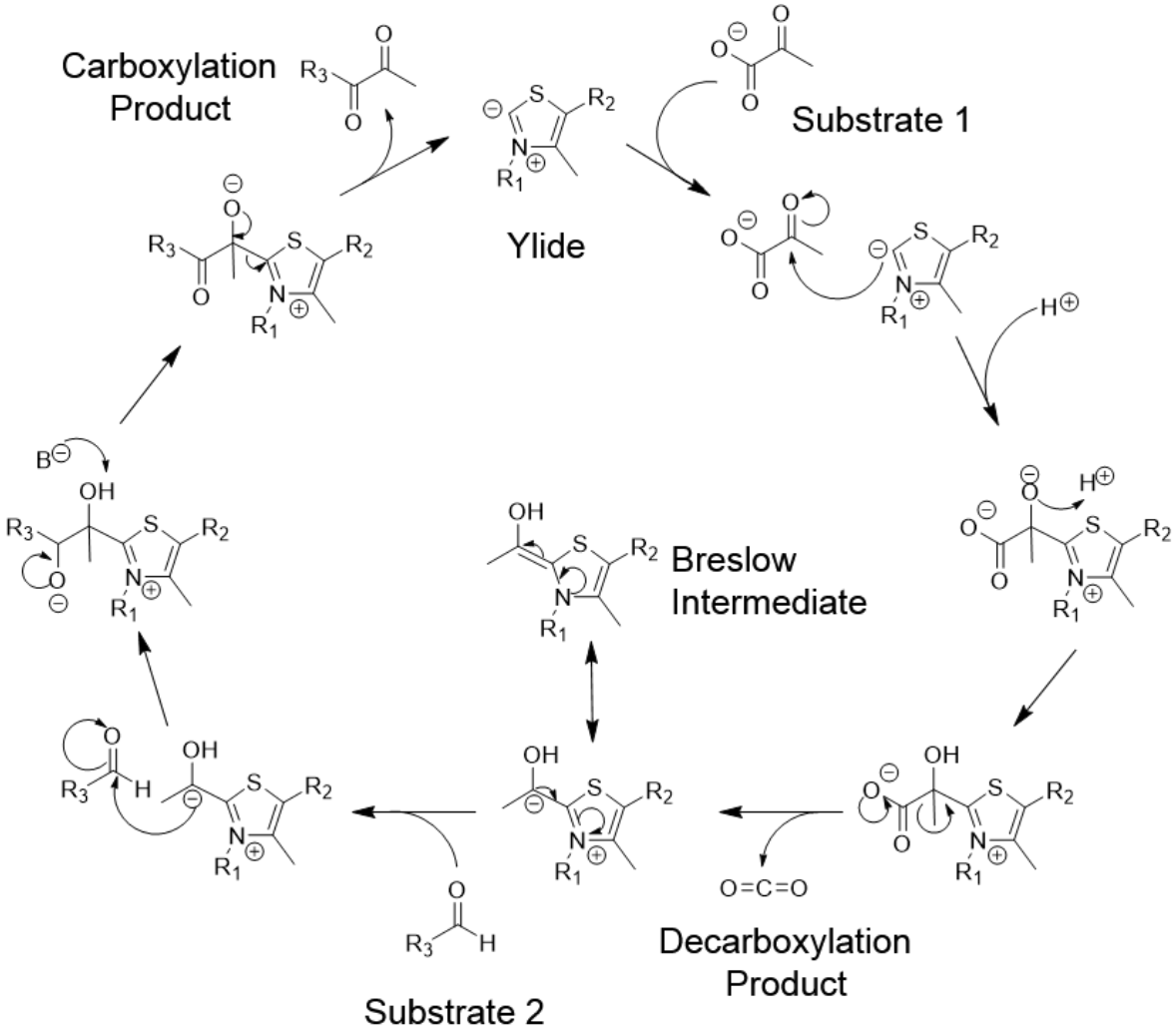


Figure 4.3 Catalytic cycle of thiamine diphosphate. Catalysis occurs through resonance stabilization of the carbanion, shown in middle.[10]

ThDP allows for normally electrophilic aldehydes or ketones to behave as nucleophiles, equivalent to acyl anions. Reactions that utilize this inversion of polarity are known as "umpolung" reactions, from the German word "umpolungskatalyse," or charge reverse catalysis.[11] Using this "umpolung" chemistry, ThDP-dependent enzymes are able to catalyze acyl addition reactions to electrophiles that lead to the formation of carbon-carbon, carbon-oxygen, carbon-nitrogen, and carbon-sulfur bonds, depending on the identity of both substrates (Figure 4.4).[12] For α-keto acid substrates, ThDP can also catalyze decarboxylation reactions. If a proton enters as substrate 2 in the second half of the cycle shown in Figure 4.3, a reduced decarboxylated aldehyde product is formed.

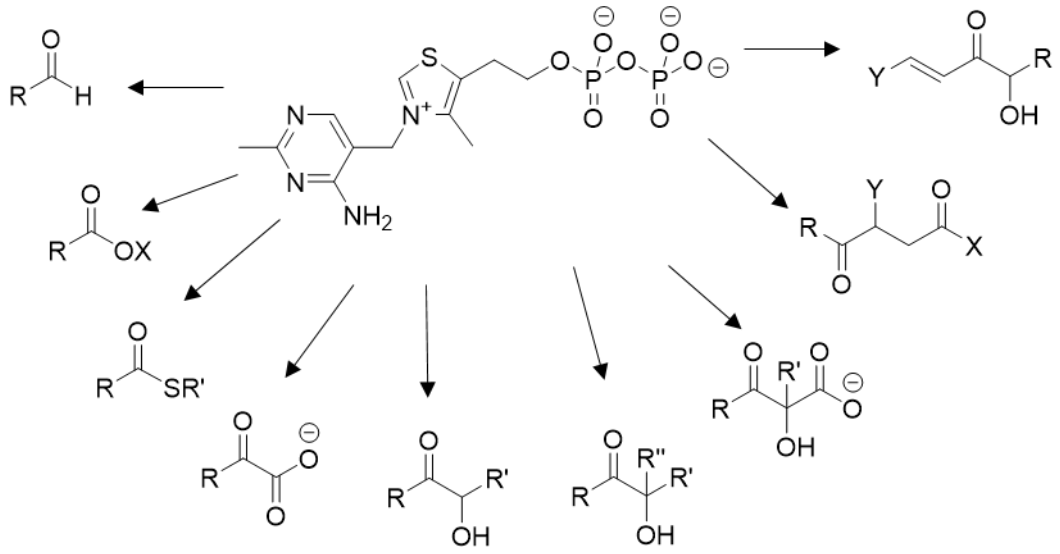


Figure 4.4 Products of thiamine catalysis.[13] Adapted and reprinted from Current Opinion in Biotechnology, Vol 15, Pohl, M.; Sprenger, G. A; Müller, M., A New Perspective on Thiamine Catalysis, 335-342, 2004, with permission from Elsevier.

4.4 ThDP-Dependent Enzymes Confer Stereospecificity to ThDP

ThDP-dependent enzymes have been shown to possess limited amino acid sequence similarity (usually less than 20% identity) with other ThDP-dependent enzymes of similar function.[9] Despite the disparate sequences, the folds are remarkably similar, a hallmark of

proteins that arose early in evolution. This observation suggests that ThDP-dependent enzymes divergently evolved from a common ancestor. ThDP enzymes have evolved catalysis for a wide scope of substrates due to the variability of the enzyme active sites.

As ThDP is regenerated at the end of the catalytic cycle, ThDP remains tethered to the active site of ThDP-dependent enzymes. ThDP's diphosphate sidechain, usually in concert with a divalent metal ion such as Mg^{2+}, docks the cofactor to the enzyme active site (Figure 4.5).[9] Interestingly, the cofactor adopts a "V" conformation where the $N_{4'}$ of the aminopyrimidine ring is brought into close proximity to the C_2 of the thiazole ring (Figure 4.6). This conformation is required for catalysis, and as a result, it has been suggested the $N_{4'}$ of the aminopyrimidine ring acts as the base responsible for deprotonation to form the catalytically active ylide. Many ThDP residues have an "invariant glutamate" residue that has supported the theory of the $N_{4'}$ nitrogen as the base in catalysis.[9] Intriguingly, this "V" conformation is not observed in free ThDP; it is a result of the active site. ThDP-dependent enzymes contain conserved residues around the three nitrogen atoms in the pyrimidine ring, presumably for hydrogen bonding the cofactor into the active "V" conformation.[9,10,14] Unexpectedly, the thiazole ring, the chemical moiety of catalysis, does not appear to make any direct contacts with the enzyme active site.

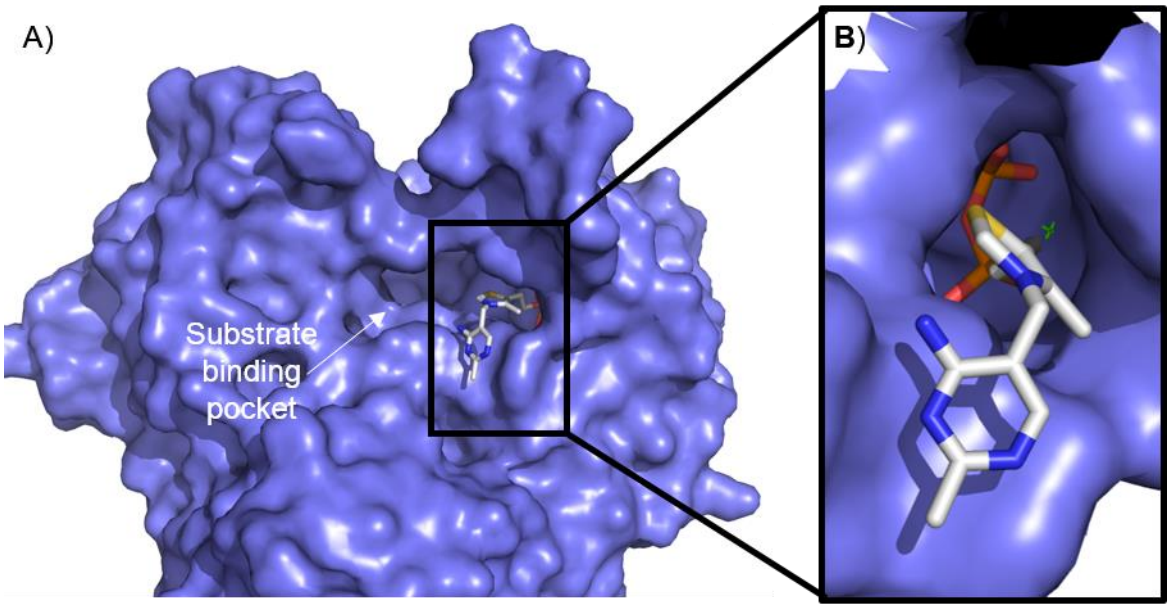


Figure 4.5 Structure of benzoylformate decarboxylase (PDB ID: 1BFD). A) Binding pocket of BDF contains distinct corridors for ThDP (gray) and substrate binding. B) Binding of ThDP using a Ca^{2+} ion, shown in green.

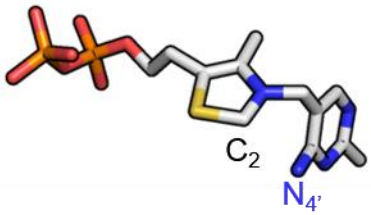


Figure 4.6 The "V" conformation of ThDP as extracted from benzyoylformate decarboxylase (PDB ID: 1BFD).

ThDP-dependent enzyme active sites serve functions both in activating ThDP and providing a suitable environment for catalysis. In aqueous conditions, ThDP is a poor catalyst due to C_2's high pK_a of 17–19. Binding pockets of ThDP enzymes are largely hydrophobic with an estimated "effective protein dielectric constant" of 13–15, which lies between those of 1-hexanol and 1-pentanol.[9,14] The hydrophobic pocket aids in catalysis by stabilizing zwitterionic intermediates and lowering the effective pK_a of the C_2 carbon by 9–10 units.[9,14] ThDP-dependent enzymes exclude most water and polar molecules from their

active sites, thereby protecting the nucleophile from being quenched before it meets its substrate.[9] As with any enzyme, the active site also contributes stereoselective control over catalysis.

4.5 Synthetic Chemists Drew Inspiration from ThDP

Nature, being the ideal engineer, has served as the inspiration for many synthetic chemists. This holds true for the thiazole ring of ThDP, which became the basis for many N-heterocyclic carbene (NHC) catalysts. NHCs such as imidazole, thiazole, and triazole (Figure 4.7) have been fashioned into diverse NHC catalysts. However, without the stereoselectivity imparted by an enzyme active site, many of these catalysts suffer from poor enantioselectivity. NHCs have been incorporated into chiral ligands and metal-binding complexes as a means of expanding stereoselective NHC-catalyzed chemistry.

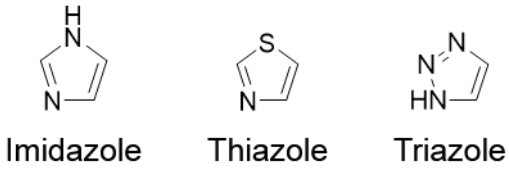


Figure 4.7 Structure of N-heterocyclic carbenes.

Cleverly, synthetic chemists conferred stereoselectivity onto thiazole compounds by exploiting the chirality of peptides. This pursuit was undertaken by Scott J. Miller and colleagues after discovering that histidine-derived peptides containing imidazole-like NHCs showed promise as enantioselective catalysts.[15] By harnessing the thiazole ring in the form of an unnatural amino acid (UAA), thiazolylalanine (TAZ, Figure 4.8), the NHC catalyst could be incorporated into a peptide using solid phase peptide synthesis.[15,16] By positioning the TAZ derivatives in the center of the peptide, stability, enantioselectivity, and yield were

increased relative to the free UAA. C-terminal and N-terminal TAZ peptides were not as stereoselective or productive as peptides with internally located TAZs.[15] Presumably, internal TAZ peptides worked in a manner similar to the active site: hydrophobic residues helped to stabilize the reactivity of the TAZ ylide, while conferring stereochemical restrictions on the products.[15,16]

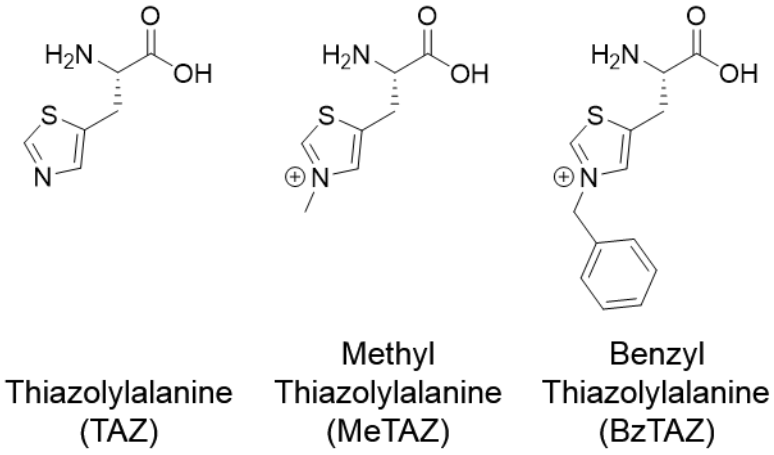


Figure 4.8 Structure of thiazolylalalnine and its derivatives.

4.6 TAZ Has Been Studied in Many Scaffolds

Miller and colleagues are not the first to have studied TAZ catalysis. Imperiali and coworkers also incorporated TAZ into even larger peptide scaffolds.[8,17] TAZ was incorporated into peptides or helical bundles modeled after DeGrado's $\alpha_1\beta$ helical bundles, which have previously been used successfully to design *de novo* polypeptides of defined structure.[17] These peptide-TAZ constructs, or "chimeras," were modeled to mimic the interior of pyruvate decarboxylase, a ThDP-dependent enzyme.[17] These constructs included TAZ derivatives such as MeTAZ and BzTAZ (Figure 4.8). The acidity of the C_2 carbon was assessed by hydrogen/deuterium (H/D) due to thiazole reactivity relying on deprotonation to form the active ylide species. The optimally performing BzTAZ chimeras featured BzTAZ

positioned at the interface of the helical bundle, the configuration most analogous with an enzyme active site.[17] Benzyl derivatives consistently outperformed their methyl counterparts, consistent with findings that H/D rates increase due to inductive effects of the N_3 substituent.[17] However, all chimeras were found to be catalytically active, suggesting that the enzyme active site helps increase catalytic activity and specificity, but is not solely responsible for it.

Suckling and colleagues expanded on Imperiali's approach by creating "thiazolopapain" constructs.[17,18] These thiazolopapains were produced by ligating methyl- and benzyl-thiazoles to the active site cysteine of papain, a cysteine protease. Reactivity was quantitated by product turnover of an acetoin condensation reaction, a well-characterized ThDP-catalyzed reaction. Free thiazolium compounds were found to catalyze turnover by 23%, while thiazolopapains increased the same reaction by up to 88%, suggesting thiazolium salts are more proficient when attached to a competent hydrophobic scaffold.[8]

4.7 Incorporating TAZ through *In Vivo* UAA Mutagenesis

Nature has utilized cofactors to expand functionality without expanding the genetic code. We aim to do what nature has not yet done: to expand functionality by expanding the genetic code through *in vivo* UAA mutagenesis. We theorize that a catalytically-active NHC-catalyst can be incorporated *in vivo* to expand biocatalysis. By incorporating a TAZ-like UAA *in vivo*, we may expand upon previous approaches to NHC catalysis by encoding a NHC into the active site of a protein. A genetically-encoded cofactor could make any binding pocket into a competent enzyme. Accordingly, this technique would allow for potent small

molecule NHCs with poor enantioselectivity to be incorporated into an enzyme for increased stereoselectivity.

For simplicity, we chose to begin with the simplest thiazolium UAA, MeTAZ, and imidazolium-UAA, dimethyl histidine (DMH, Figure 4.9). We believe these UAAs will be amenable to UAA mutagenesis due to their similarity to histidine, a canonical amino acid which also contains a 5-membered heterocycle (Figure 4.9). From the perspective of the cell, MeTAZ and histidine are likely similar enough for tolerance by the ribosome, elongation factors, and mechanisms of amino acid transport within the cell.[6]

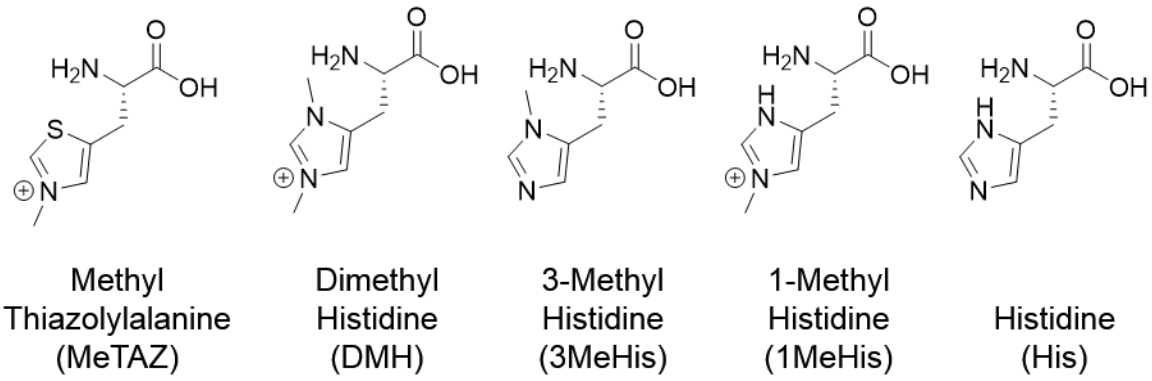


Figure 4.9 Methyl thiazolylalanine is structurally similar to histidine derivatives.

Thiazolium-containing UAAs have not been incorporated into proteins through *in vivo* UAA mutagenesis. However, Peter Schultz and colleagues have successfully incorporated TAZ and 3-methylhistidine, an even closer mimic of MeTAZ, by evolving the synthetase of *Methanosarcina barkeri* which encodes pyrrolysine (Figure 4.10).[19] Preliminary screening for TAZ incorporation by this synthetase (pylHRS) showed that it was not amenable to thiazolium UAA incorporation, which is likely due to the cationic nature of the MeTAZ derivative. We therefore chose to evolve our own MeTAZ synthetase (MeTAZRS).

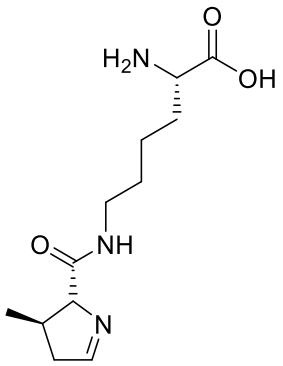


Figure 4.10 Structure of pyrrolysine (Pyl).

4.8 MeTAZRS Randomized Library Design

Randomized libraries were generated using "NNK" codons at chosen sites, where N represents any of the four nitrogenous bases and K represents G or T. There are 32 combinations that arise from the NNK codon, but all 20 amino acids may be encoded by these combinations. The limit of transformation efficiency for commercially available competent cells is $1x10^9$, thereby limiting the NNK library to six sites. We chose to base the design of our randomized synthetase library on two synthetases: the histidine synthetase reported by Schultz and coworkers and a promiscuous tyrosine synthetase introduced by Wenshe Liu and collaborators.[19,20] Both synthetases are actually derived from naturally-occurring pyrrolysine synthetases, from *M. barkeri* and *Methanosarcina mazei*, respectively. Both synthetases contain mutations at homologous cysteines (C313 and C348, respectively) lining the binding pocket that help to shorten the cavity for binding smaller UAAs. We chose to randomize these two residues in concert to every possible amino acid combination. In addition, a conserved Tyr → Phe mutation (Y384F in *M. mazei*) has been reported to increase aminoacylation rates of UAAs; therefore, the randomization of this site was also included in the library.[19–21] Additionally, V401 and W417 (*M. mazei* numbering) residues

were randomized in Shultz's PylHRS in order to accommodate a smaller NHC-UAA. Although the specific mutations introduced in PylHRS did not incorporate MeTAZ, other mutations at these sites may better accommodate the alkylated thiazole ring. The N346 residue participates in an important hydrogen bond with pyrrolysine, and mutation of this residue has proven useful in tuning selectivity from pyrrolysine to aromatic amino acids. The A302 residue is nestled at the mouth of the binding pocket near the base of pyrrolysine. These two residues, N346 and A302, were included in the library to help close off the long pyrrolysine binding pocket while accommodating an NHC. Although multiple rounds of randomization are often necessary, the A302, N346, C348, Y384, V401, and W417 residues are rational starting points for MeTAZ-incorporating synthetase design (Figure 4.11).

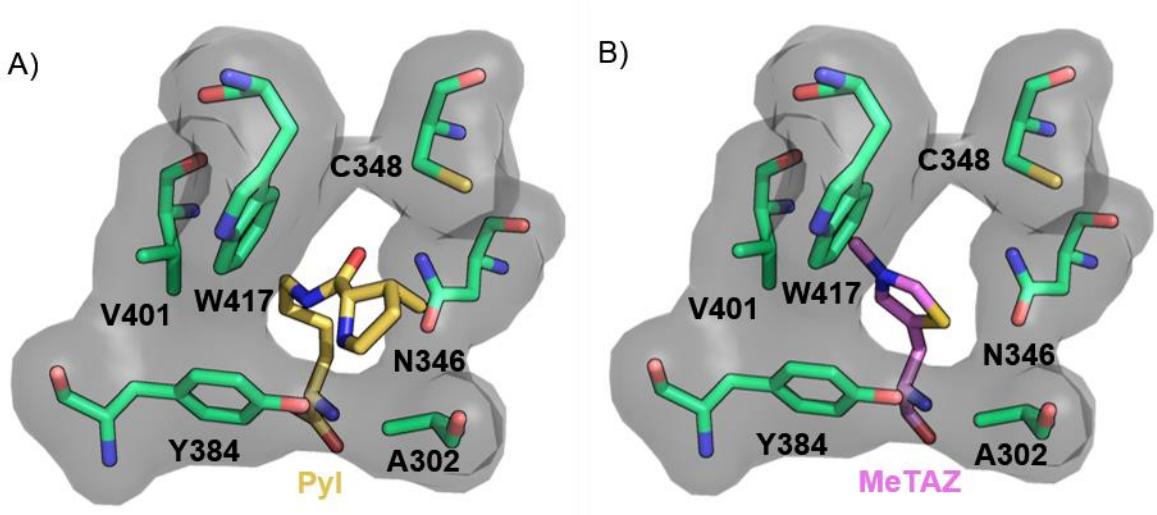


Figure 4.11 Residues for the randomized MeTAZRS library. A) Structure of pyrrolysine bound to pyrrolysine synthetase of *M. Mazei* (PDB ID: 2ZIM). Residues to be randomized are shown as sticks. B) Structure of pyrrolysine synthetase with MeTAZ modeled into the active site.

4.9 Library Selection and Screening

Generation of a randomized library can be readily accomplished via PCR: primers with "NNK" codons are commercially available from common suppliers such as IDT. For

our library, four rounds of PCR were required: the 346 and 348 residues can be mutated in a single primer, as can the 401 and 417 residues, while the 302 and 384 residues must be mutated individually. We chose to begin randomization with the 346 and 348 sites due to their positions in the pocket. From there, we continued to PCR additional NNK codons onto the gene until the six site library was generated. Once the two site 346NNK and 348NNK library was generated, we began screening as we worked to create the six site library.

Multiple selection and screening systems have been developed for UAA incorporation.[22] Selection systems are most efficient for evolving orthogonal synthetases: they amplify mutants with desired UAA-activity while reducing mutants with non-specific or canonical amino acid activity. The selection system we chose, known as pREP-pylT, contains a chloramphenicol resistance gene (chloramphenicol acyltransferase, *cat*) with an amber (TAG) codon (Figure 4.12A). When the TAG codon behaves as a sense codon and an amino acid (unnatural or canonical) is charged, the full gene is translated, allowing for cells to grow on chloramphenicol-containing media (Figure 4.12B and C).[23] When the TAG codon behaves as a stop codon, the full gene is not translated and cells will be sensitive to chloramphenicol (Figure 4.12D). The 346/348 NNK library was transformed into DH10B cells containing the pREP-pylT plasmid and then plated on agar plates with appropriate antibiotics, but no chloramphenicol. Transformant colonies were picked and inoculated in 96-well blocks. For a two-site library with 32 possible codons, at least 1024 (32^2) colonies were required to ensure library coverage, and therefore 12 blocks were inoculated.

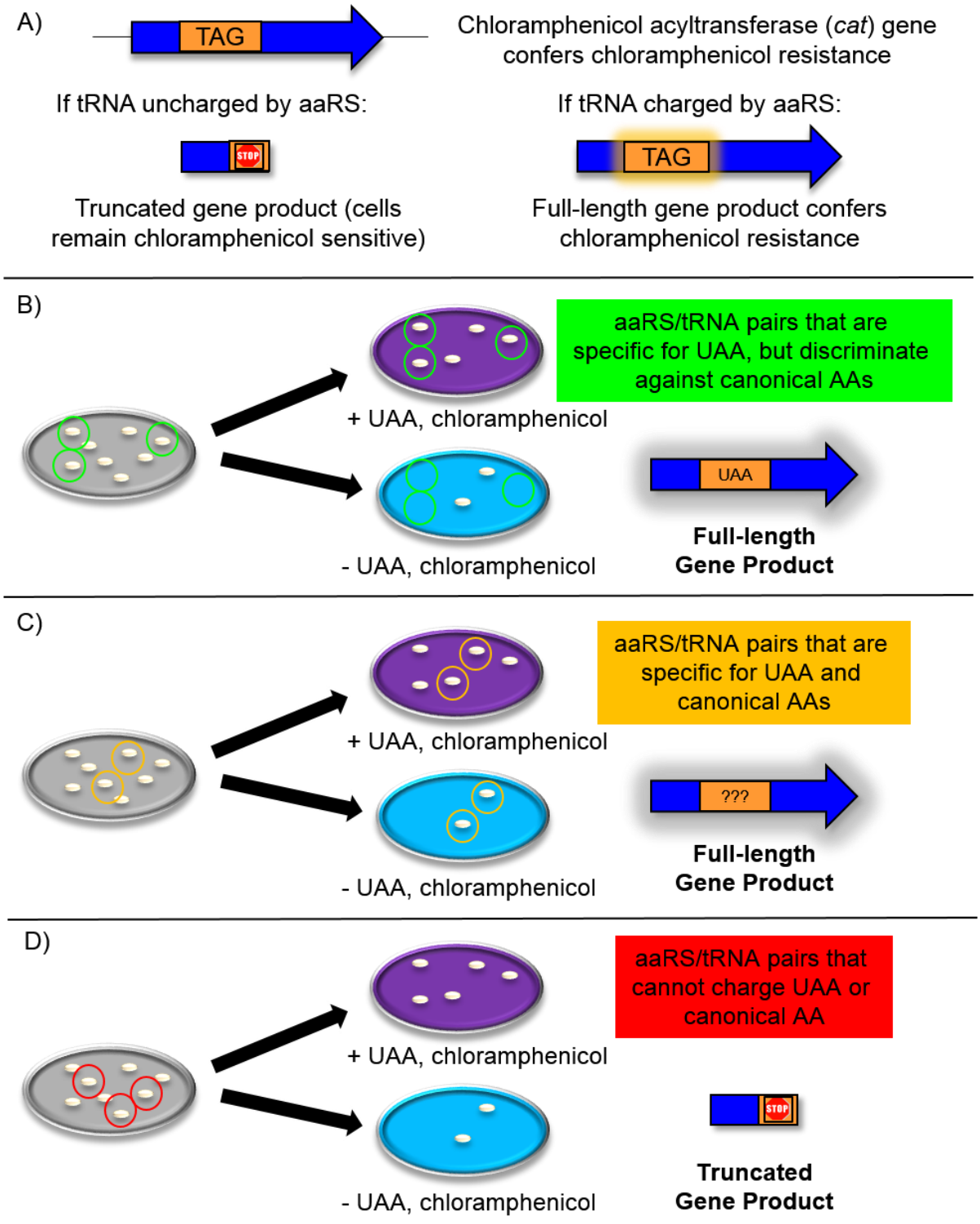


Figure 4.12 Positive and negative selections for UAA incorporation using the *cat*-TAG system.[23]

Chloramphenicol is a bacteriostatic antibiotic; as a result, chloramphenicol-sensitive cells may grow on chloramphenicol after sufficient time. pREP-pylT selections work best on solid media to minimize *E.coli*'s ability to overgrow their chloramphenicol sensitivity. Cultures in the 96-well blocks were grown to saturation and then stamped onto 24 cm agar plates containing chloramphenicol in the presence or absence of the following UAAs: Boc-lysine, a positive control for wild type pylRS,[24] 1-methyhistidine (1MeHis), 3-methylhistidine (3MeHis), MeTAZ, and DMH. Selective orthogonal pairs, or selection hits, were found on plates that grew in the presence of UAA, but not in the absence of UAA. Selections were then repeated with higher concentrations of chloramphenicol, thereby increasing the stringency of the selection.

Plasmid DNA for PylRS variants identified as hits was isolated from the pREP-pylT reporter plasmid and transformed into cells containing a superfolder green fluorescent protein (sfGFP)-TAG construct. These constructs allow for fluorescence screening to detect amino acid incorporation (Figure 4.13).[20,25] Similar to the previous selection, selective orthogonal synthetases show high fluorescence in the presence of UAA, but not in the absence of UAA. Once fluorescence was measured, the five best hits were found in wells 3E10, 3F3, 5A11, 5C1, and 5C3 (Figure 4.14). One mutant, from well 2B5, gave the lowest incorporation fluorescence of all the mutants. This mutant was selected as a negative control to see if an ill-functioning binding pocket differed in composition from the other synthetases. Mutant identities have been included with corresponding fluorescence data in Figure 4.14.

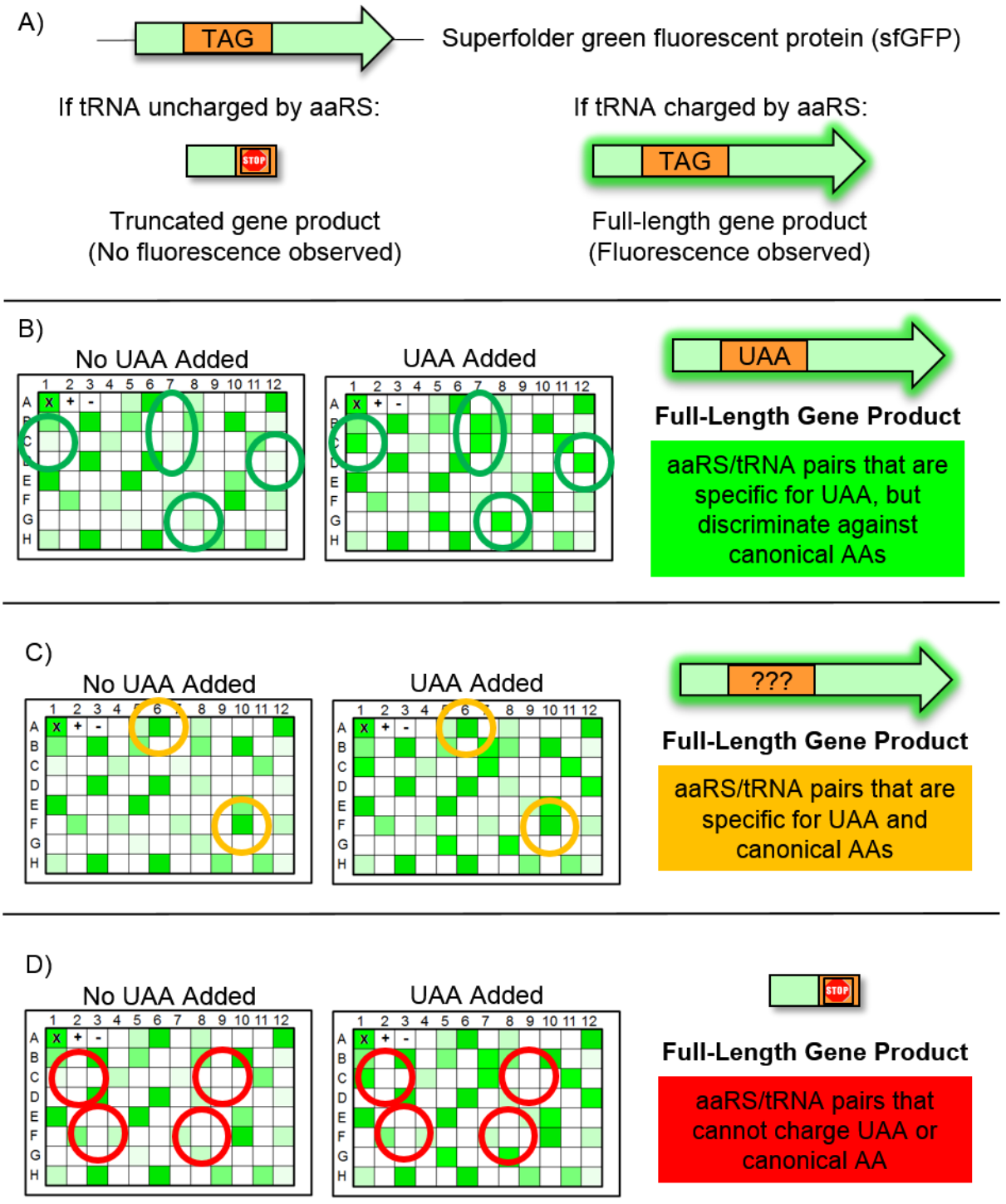


Figure 4.13 Fluorescence screening for UAA incorporation using the sfGFP(2TAG) system.[20,25]

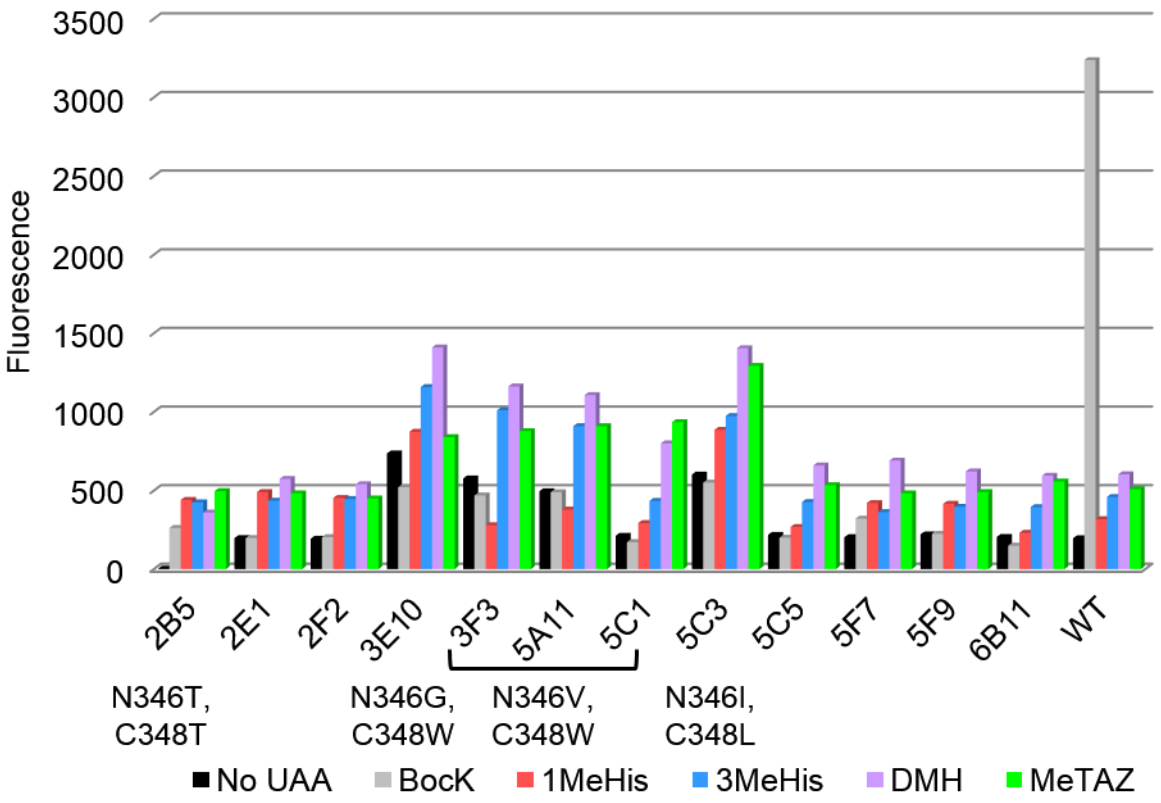


Figure 4.14 Fluorescence of mutants from 346/348 NNK library. BocK was used as a positive control for wild type mmPylRS.

4.10 Mutations to the mmPylRS Binding Site Are Largely Hydrophobic

Three of the five synthetases with the highest fluorescence levels contain a N346V/C348W mutation (Figure 4.15A). Even more intriguingly, these mutants have more than one sequence due to the degeneracy of the valine codon. Other mutations included N346G/C348W (3E10) and N346I/C348L (5C3) (Figure 4.15B and C). These mutations, in conjunction with the N346V/C348W mutation, change an acidic and polar residue (asparagine and cysteine, respectively) to two hydrophobic residues, reminiscent of the binding pocket of ThDP-dependent enzymes. Four of the five synthetases contain a tryptophan at the C348 position, which likely limits the depth of the binding pocket to better interact with the NHC-UAA. The remaining mutant contains a leucine at the C348 position,

which appears to be more confining than the tryptophan mutation (Figure 4.15C) Schultz and coworkers saw similar cavity-closing mutations from cysteine to phenylalanine in their pylHRS.[19] Intriguingly, the 2B5 mutant contains a N346T/C348T mutation (Figure 4.15D), which introduces two polar residues into the binding pocket. These mutations are distinct from those of the "active" MeTAZ synthetases, suggesting that our screens and selections are working to amplify sequences of selective, orthogonal synthetases.

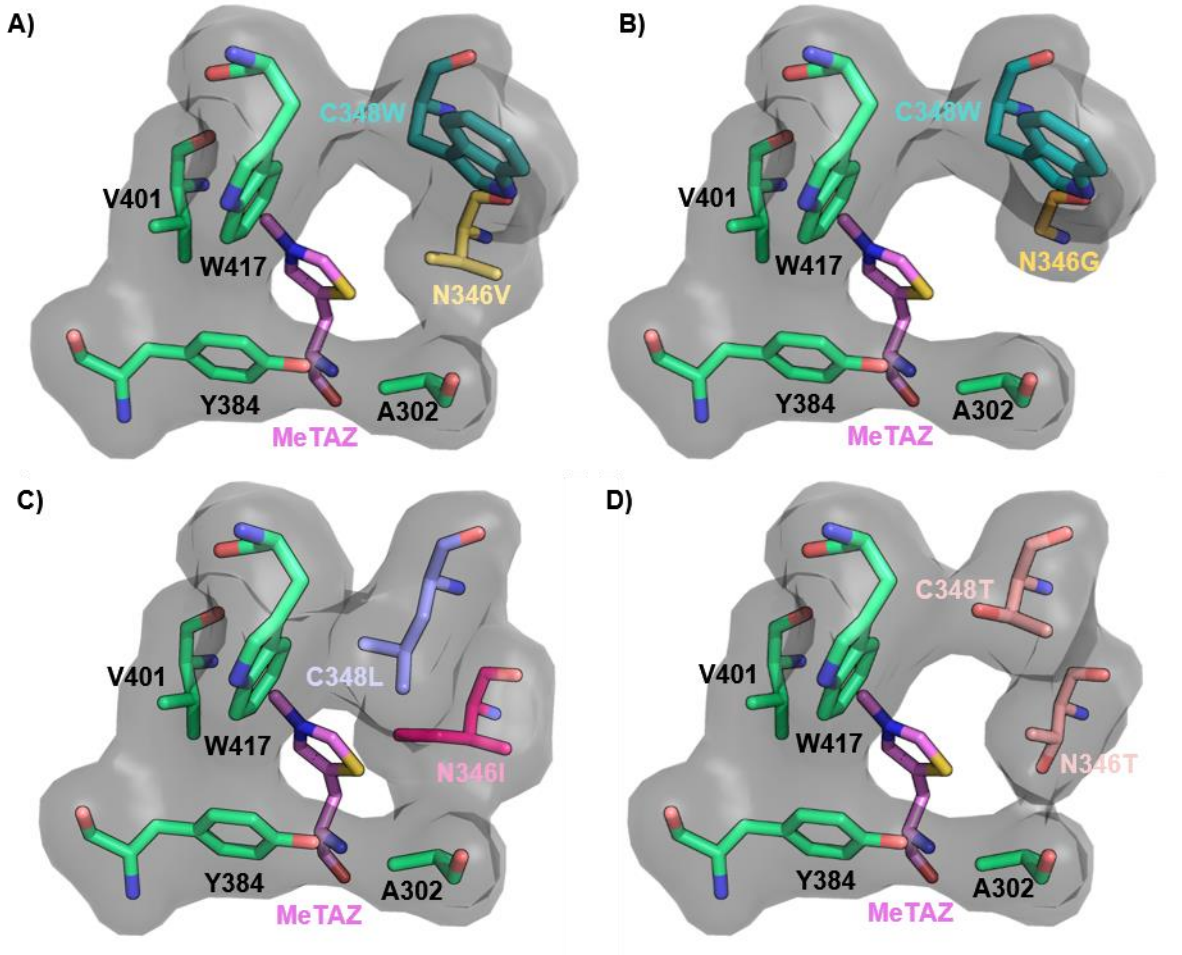


Figure 4.15 Mutations identified in sfGFP screen modeled onto the binding pocket of mmPylRS. A) N346V/C348W mutation, B) N346G/C348W mutation, C) N346I/C348L mutation, D) N346T/C348T mutation.

4.11 Discussion

Unfortunately, an orthogonal, selective synthetase for MeTAZ and DMH was not identified from only two NNK-sites. Results from the two-site library mimic a ThDP-dependent enzyme pocket and suggest our modifications are steps on the way to orthogonality. However, our most selective and orthogonal synthetases incorporate our UAAs at a level that is just 4-fold over background and exhibit under half the fluorescence of our positive control. The high background is likely due to histidine, which is similar in size and shape to our amino acids. Our results indicate that additional rounds of selection will be required, and introduction of additional NNK sites has proven challenging.

Despite difficulty with library expansion, the rate limiting step of this process is the synthesis of MeTAZ. As we have discussed, the thiazole ring is an incredibly powerful nucleophile. The thiazole-bromide species used in TAZ synthesis is notoriously sensitive to degradation, such that even a quick trip to take an NMR spectrum could spell disaster. These side effects are observed both in high concentrations and dilute solutions. Currently we are contemplating new scaffolds or methods that may avoid these problems. If we can identify a better thiazole scaffold, we can expand our thiazolium UAAs beyond MeTAZ. The N_3 nitrogen can be easily alkylated with an alkyl halide, allowing for easy production of BzTAZ or other aromatic derivatives. It is ironic that our greatest challenge in incorporating a catalytically active UAA is overcoming the UAA's catalytic power. Yet, we are hopeful that a new amino acid scaffold will allow for genetic incorporation of an NHC catalyst.

4.12 Experimental

4.12.1 Cloning, DNA Sequences, and Protein Sequences

pBK-mmPylRS, pREP-TyrT, and pEVOL-sfGFP(2TAG) were obtained from the lab of Dr. Peter Schultz. pREP-pylT was created by standard overlap PCR to substitute the mmPyl-tRNA for the Tyr-tRNA. Pyrrolysine synthetase libraries were constructed using a BsaI cloning strategy. Oligonucleotides for PCR were obtained from Integrated DNA Technologies and enzymes and reagents used for cloning were obtained from New England BioLabs Inc. DNA sequences of the NNK primers and mmPylRS are shown below. Mutation sites have been bolded for clarity and BsaI cut sites have been underlined.

PylRS_A302NNK_BsaI_F:

GAACTCGGTCTCAGCTT**NNK**CCAAACCTTTACAACTACC

PylRS_A302_BsaI_R:

GAACTCGGTCTCAAAGCATGGGTCTCAGGCAGAAGTTC

PylRS_N346NNK_C348NNK_BsaI_F

GAATTCGGTCTCAGCTGNNKTTC**NNK**CAGATGGGATCGGG

PylRS_N346_C348_BsaI_R

GAATTCGGTCTCACAGCATGGTAAACTCTTCGAGGTGTTC

PylRS_Y384_V401NNK_BsaI_F

GAACTCGGTCTCAACGGAGACCTGGAACTTTCCTCTGCA**NNK**GTCGG

PylRS_Y384NNK_V401_BsaI_R:

GAACTCGGTCTCACCGTGCATTACATCAAGGGTATCCCC**MNN**GACCATG

PylRS_W417NNK_BsaI_F:

GAACTCGGTCTCAACCC**NNK**ATAGGGGCAGGTTTCGGGCTCGAAC

PylRS_W417_BsaI_R:

GAACTCGGTCTCAGGGTTTATCAATACCCCATTCCCGGTCAAG

mmpylRS_Library_Sequencing_Primer_F:

GCAGATCTACGCGGAAGAAAGG

Methanosarcina mazei PylRS DNA sequence:

ATGGATAAAAAACCACTAAACACTCTGATATCTGCAACCGGGCTCTGGATGTCC

AGGACCGGAACAATTCATAAAATAAAACACCACGAAGTCTCTCGAAGCAAAATC

TATATTGAAATGGCATGCGGTGACCACCTTGTTGTAAACAACTCCAGGAGCAGC

AGGACTGCAAGAGCGCTCAGGCACCACAAATACAGGAAGACCTGCAAACGCTG

CAGGGTTTCGGATGAGGATCTCAATAAGTTCCTCACAAAGGCAAACGAAGACCA

GACAAGCGTAAAAGTCAAGGTCGTTTCTGCCCCTACCAGAACGAAAAAGGCAAT

GCCAAAATCCGTTGCGAGAGCCCCGAAACCTCTTGAGAATACAGAAGCGGCACA

GGCTCAACCTTCTGGATCTAAATTTTCACCTGCGATACCGGTTTCCACCCAAGAG

TCAGTTTCTGTCCCGGCATCTGTTTCAACATCAATATCAAGCATTTCTACAGGAG

CAACTGCATCCGCACTGGTAAAAGGGAATACGAACCCCATTACATCCATGTCTG

CCCCTGTTCAGGCAAGTGCCCCCGCACTTACGAAGAGCCAGACTGACAGGCTTA

AAGTCCTGTTAAACCCAAAAGATGAGATTTCCCTGAATTCCGGCAAGCCTTTCAG

GGAGCTTGAGTCCGAATTGCTCTCTCGCAGAAAAAAAGACCTGCAGCAGATCTA

CGCGGAAGAAAGGGAGAATTATCTGGGGAAACTCGAGCGTGAAATTACCAGGTT

CTTTGTGGACAGGGGTTTTCTGGAAATAAAATCCCCGATCCTGATCCCTCTTGAG

TATATCGAAAGGATGGGCATTGATAATGATACCGAACTTTCAAAACAGATCTTC

AGGGTTGACAAGAACTTCTGCCTGAGACCCATGCTT**GCT**CCAAACCTTTACAACT

ACCTGCGCAAGCTTGACAGGGCCCTGCCTGATCCAATAAAAATTTTTGAAATAG

GCCCATGCTACAGAAAAGAGTCCGACGGCAAAGAACACCTCGAAGAGTTTACCA

TGCTG**AAC**TTC**TGC**CAGATGGGATCGGGATGCACACGGGAAAATCTTGAAAGCA

TAATTACGGACTTCCTGAACCACCTGGGAATTGATTTCAAGATCGTAGGCGATTC

CTGCATGGTC**TAT**GGGGATACCCTTGATGTAATGCACGGAGACCTGGAACTTTCC

TCTGCA**GTA**GTCGGACCCATACCGCTTGACCGGGAATGGGGTATTGATAAACCC

TGGATAGGGGCAGGTTTCGGGCTCGAACGCCTTCTAAAGGTTAAACACGACTTT

AAAAATATCAAGAGAGCTGCAAGGTCCGAGTCTTACTATAACGGGATTTCTACC

AACCTGTAA

Methanosarcina mazei PylRS protein sequence:

MDKKPLNTLISATGLWMSRTGTIHKIKHHEVSRSKIYIEMACGDHLVVNNSRSSRTA

RALRHHKYRKTCKRCRVSDEDLNKFLTKANEDQTSVKVKVVSAPTRTKKAMPKSV

ARAPKPLENTEAAQAQPSGSKFSPAIPVSTQESVSVPASVSTSISSISTGATASALVKG
NTNPITSMSAPVQASAPALTKSQTDRLEVLLNPKDEISLNSGKPFRELESELLSRRKK
DLQQIYAEERENYLGKLEREITRFFVDRGFLEIKSPILIPLEYIERMGIDNDTELSKQIFR
VDKNFCLRPMLAPNLYNYLRKLDRALPDPIKIFEIGPCYRKESDGKEHLEEFTMLNFC
QMGSGCTRENLESIITDFLNHLGIDFKIVGDSCMVYGDTLDVMHGDLELSSAVVGPIP
LDREWGIDKPWIGAGFGLERLLKVKHDFKNIKRAARSESYYNGISTNL

The DNA and protein sequences of sfGFP(2TAG) of pEVOL-sfGFP(2TAG)-pylT are included below. The 2 position is shown in red for clarity.

sfGFP(2TAG) DNA sequence:

ATG**TAG**AAAGGAGAAGAACTTTTCACTGGAGTTGTCCCAATTCTTGTTGAATTAG
ATGGTGATGTTAATGGGCACAAATTTTCTGTCCGTGGAGAGGGTGAAGGTGATG
CTACAAACGGAAAACTCACCCTTAAATTTATTTGCACTACTGGAAAACTACCTGT
TCCGTGGCCAACACTTGTCACTACTCTGACCTATGGTGTTCAATGCTTTTCCCGTT
ATCCGGATCACATGAAACGGCATGACTTTTTCAAGAGTGCCATGCCCGAAGGTT
ATGTACAGGAACGCACTATATCTTTCAAAGATGACGGGACCTACAAGACGCGTG
CTGAAGTCAAGTTTGAAGGTGATACCCTTGTTAATCGTATCGAGTTAAAGGGTAT
TGATTTTAAAGAAGATGGAAACATTCTTGGACACAAACTCGAGTACAACTTTAA
CTCACACAATGTATACATCACGGCAGACAAACAAAAGAATGGAATCAAAGCTAA
CTTCAAAATTCGCCACAACGTTGAAGATGGTTCCGTTCAACTAGCAGACCATTAT
CAACAAAATACTCCAATTGGCGATGGCCCTGTCCTTTTACCAGACAACCATTACC
TGTCGACACAATCTGTCCTTTCGAAAGATCCCAACGAAAAGCGTGACCACATGG

TCCTTCTTGAGTTTGTAACTGCTGCTGGGATTACACATGGCATGGATGAGCTCTA
CAAAGGATCCCATCACCATCACCATCACTAA

sfGFP(2TAG) protein sequence (* denotes an UAA)

M*KGEELFTGVVPILVELDGDVNGHKFSVRGEGEGDATNGKLTLKFICTTGKLPVPW
PTLVTTLTYGVQCFSRYPDHMKRHDFFKSAMPEGYVQERTISFKDDGTYKTRAEVK
FEGDTLVNRIELKGIDFKEDGNILGHKLEYNFNSHNVYITADKQKNGIKANFKIRHNV
EDGSVQLADHYQQNTPIGDGPVLLPDNHYLSTQSVLSKDPNEKRDHMVLLEFVTAA
GITHGMDELYKGSHHHHHH

4.12.2 pREP-pylT NHC-UAA Selections

pREP-pylT was cotransformed with the pBK-mmpylRS-346/348NNK library into electrocompetent DH10B cells. The cells were rescued with 1 mL SOC and the resulting rescue solution was used to inoculate 9 mL of LB (10 mL total) in a 14 mL falcon tube. 10 uL of 50 mg/mL kanamycin and 10 uL of 25 mg/mL tetracycline were added to each tube and the cultures were left to grow overnight at 37°C with shaking at 225 RPM. The following day, OD_{600} of each culture was measured and cells were diluted down with LB based on A_{600} of 1 = 1 x 10^8 colonies to afford ~15,000 colonies plated in 200 uL of solution. 200 uL of diluted cells was plated onto each of two 150 mm diameter LB agar plates with 50 mg/L kanamycin and 25 mg/L tetracycline. Plates were incubated overnight at 37°C.

100 mM UAA stocks were prepared by dissolving dry UAA into the appropriate diluent (SI Table 4.1). For each UAA, 5 mL of 100 mM stock was added to a 500 mL flask of LB agar supplemented with 50 mg/L kanamycin, 25 mg/L tetracycline, and 40 mg/L

chloramphenicol (LB Kan Tet Cam agar). For a negative control, 5 mL of 0.1M NaOH was added to a 500 mL flask of LB Kan Tet Cam agar. For a growth control (no selection pressure), 500 mL of LB Kan Tet was prepared. For each flask, ~480 mL was split amongst three 24 cm x 24 cm square petri dishes and left to solidify at room temperature. Plates were covered with aluminum foil or left in the dark to prevent photodegradation of chloramphenicol.

Once single colonies were visible on the 150 mm round plates, twelve 96-well (2 mL well) blocks were filled with 1 mL of LB Kan Tet. In each block, A1 was inoculated with pBK-mmPylRS (wild type) + pREP-pylT in DH10B. Wild type mmPylRS can incorporate BocK with high affinity, therefore A1 was inoculated as a positive control.[24] A2 was inoculated with pBK-mmPylRS (wild type) + pREP-TyrT, a negative control, which contains a Tyr-tRNA that cannot be charged by the mmPylRS. The presence of growth from these colonies on chloramphenicol-containing plates would suggest either overgrowth or contamination. Using sterile toothpicks, each of the remaining wells were inoculated with colonies from the 150 mm mmPylRS library plates. Sterile toothpicks were used to inoculate each well and the toothpicks were left in the well until all wells in the block had been inoculated. The reasoning behind this was twofold: to ensure transfer of cells from the toothpick to the media and to serve as a placeholder to keep track of inoculated wells. Each block was covered with parafilm and then a sterile P10 micropipette tip was used to pierce holes over each well to allow for gas exchange. The plates were then incubated overnight at 37°C in a plate shaker at 990 RPM.

Each of the 12 blocks, now containing saturated cultures in each well, were then stamped onto each of the 24 cm square plates using a 96-prong pin tool. For each condition,

four different blocks were stamped side by side onto each of the three plates. The plates were left, cover off, next to a flame to dry and then incubated at 37°C for ~48 hours until colonies were visible on the A1 position of the BocK positive control plates. Each colony was inventoried and then "hits" were identified based on those with growth on UAA containing plates, but not in the absence of UAA. The most promising 94 colonies (along with the two controls previously discussed) were then inoculated into a fresh 96-well block with 1 mL of LB Kan Tet in each well and incubated overnight in a plate shaker at 37°C with shaking at 990 RPM.

The 94 hit conditions were then subjected to increased selection stringency of 100 mg/L chloramphenicol. For each UAA, four conical vials containing 40 mL of LB Kan Tet with 100 mg/L Cam were spiked with 400 uL of 100 mM UAA stock to afford a 1 mM final concentration. The contents of each conical vial were poured into a 150 mm diameter petri dish and left to solidify at room temp covered in aluminum foil. Plates were then stored overnight at 4°C in the dark.

Once the block containing the selection "hits" had grown to saturation, the contents of the block were stamped onto each of the plates using a 96-prong pin tool. Plates were left with lid slightly ajar near an open flame to dry. After drying, plates were incubated at 37°C for 24 hours. At this time point, plates were removed from the incubator and any colonies were marked. Plates were placed back in the incubator for 6 hours and then were imaged and inventoried. Of the 94 mutants in the block, 26 were chosen to undergo sfGFP(2TAG) screening.

4.12.3 sfGFP NHC-UAA Incorporations

From the pREP-pylT screening, 26 mutants were chosen for sfGFP(2TAG) screening. 10 mL of LB Kan Tet was inoculated from freezer strains of each of the chosen mutants. The cultures were grown to saturation at 37°C with shaking at 225 RPM. The cultures were then miniprepped to isolate plasmid DNA and the pBK plasmid was gel purified away from the pREP-plasmid. The pBK vectors were transformed into DH10B competent cells containing the pEVOL-sfGFP(2TAG)-pylT plasmid. The 1 mL rescues were transferred to a 96-well block and kanamycin and tetracycline were added. A1 of the block contained an empty well of LB Kan Tet which served as a contamination control. Positive controls (pBK-mmPylRS (wild type) + pEVOL-sfGFP (2TAG)-pylT) and negative controls (pBK-mmPylRS (wild type) + pEVOL-sfGFP (2TAG)-TyrT) were also inoculated and then the block was incubated at 37°C overnight with shaking at 990 RPM.

Once the cultures in the block had saturated, fresh 96-well blocks with LB Kan Cam (for pEVOL resistance, not a selection pressure) with 1 mM UAA were prepared as previously described. A negative control was similarly prepared using 0.1 M NaOH and no UAA. The blocks were incubated at 37°C with 990 RPM for two hours and then sfGFP expression was induced by adding arabinose to give a final concentration of 0.02% w/v. The cultures were left to express for 24 hours at 37°C with shaking and 70% humidity.

After 24 hours, the cultures were pelleted for 15 min at 4,500 RPM and the supernatant was carefully decanted. Pellets were resuspended in 500 uL of 0.1 M sodium phosphate, pH 8 with 0.25 mg/mL lysozyme. The blocks were incubated at 37°C with shaking for one hour and then the lysed cultures were clarified by centrifugation at 4,500 RPM (this speed was limited by the maximum speed on our centrifuge's plate attachment).

200 uL of the resulting supernatant was pipetted off the top (to not disturb the pellet) into a 200 uL black-bottomed and black-welled 96-well plate. The contamination control in position A1 had no growth, and instead a 1:1000 dilution of an 89 uM sfGFP control solution was included in the plate. The fluorescence gain was set based on A1 and samples were excited at 480 nm and fluorescence emission was collected at 510 nm (GFP's emission) and 650 nm (light-scattering control). Plates were centrifuged at 4,500 RPM to remove bubbles and then fluorescence was measured using a plate reader. Initial results showed high background fluorescence due to canonical amino acid incorporation, so the experiment was repeated with GMML minimal media for mutants with the highest UAA-sfGFP expression.[23] The mutants chosen were 2B5, 3E10, 3F3, 5A11, 5C1, 5C3, 7A1, and the wild type positive control. Resulting fluorescence measurements can be found in Figure 4.14.

Mutants were inoculated in 5 mL LB Kan Cam and grown to saturation overnight at 37°C with shaking at 225 RPM. The saturated solutions were then miniprepped and the plasmid DNA was sequenced by Genewiz, LLC, using the mmpylRS_Library_Sequencing_Primer_F primer.

4.12.4 NHC-UAA synthesis

All reagents were purchased from Chem Impex, Sigma-Aldrich, or Fisher Scientific. All reactions were performed in oven-dried glassware with a nitrogen atmosphere and anhydrous solvents.

4.12.4.3 Methyl Thiazolylalanine Synthesis

5-hydroxymethyl thiazole (SI Figure 4.1, (1), 1.0 g, 1 eq.) was dissolved in 25 mL of dry dichloromethane and placed on a dry ice and acetone bath. The solution was chilled for 20 minutes before adding phosphorous tribromide (1.1 eq.) dropwise to the solution. The solution was left to react until the acetone bath had warmed to room temperature. The solution was first quenched with water and then with sodium bicarbonate. The product was extracted with dichloromethane and the combined organic layers were washed with brine, dried with sodium sulfate, and concentrated to half volume on a rotary evaporator. 5-bromomethyl thiazole (2) is prone to polymerization at high concentrations, and so drying down the compound is nearly impossible. Instead, we exchanged (2) directly into ethanol by concentrating down the sample to half volume and adding additional ethanol until ~15 mL of (2) in ethanol remained.

Alkylation of diethylacetoamido malonate (3) with bromoalkanes has been previously reported as an amino acid synthetic strategy.[26] Since we were unable to calculate yields or dry masses on our previous steps, all reagents are calculated assuming 100% yield. Diethylacetoamido malonate (8.77 mmol, 1 eq.) was dissolved in 18 mL of ethanol and added dropwise added to a solution of NaOEt (1 eq.) in EtOH (21% solution). The solution was stirred at 50°C for 1 hour and then refluxed for 10 min. The solution of (2) was then dropwise added over 30 min and left to reflux overnight at 80°C. The solution was quenched with sodium bicarbonate and then extracted into diethyl ether. The organic layers were combined and washed with brine and dried over sodium sulfate. Organic layers were concentrated to produce a yellow oil, which was then purified on a silica gel column with a

3% methanol/DCM mobile phase. The product (4) was concentrated down from pooled chromatography fractions and evaporated to dryness.

Synthesis of (5) was based on previously reported alkylations of protected TAZ-derivatives.[15] Methyl iodide (112.4 mmol, ~50 eq.) was dissolved in ~7 mL MeCN and added dropwise to a solution of protected-TAZ derivative (4) (2.145 mmol, 1 eq., dissolved in ~8 mL MeCN). The solution was refluxed to 80°C with stirring overnight. The reaction mixture was quenched with bicarbonate and the product was extracted into ethyl acetate. The combined organic layers were washed over brine and dried over sodium sulfate. The resulting organic layers were concentrated on a rotary evaporator and purified on a silica gel column with a 20% ethyl acetate/hexanes mobile phase. Fractions containing the desired product (5) were pooled and concentrated to dryness to give a yellow solid.

The protected Me-TAZ product (5) was resuspended in a minimal amount of dichloromethane and transferred to a round bottom flask. The dichloromethane was evaporated off using a rotary evaporator and (5) was resuspended in 15 mL of 5 M HCl. The solution was refluxed at 110°C overnight and then cooled on ice. The product was then lyophilized to form a dark brown solid. After being redissolved in acetone and evaporated to dryness, (6) was found to be a yellow powder.

4.12.4.4 Dimethyl Histidine Synthesis

Dimethyl histidine was synthesized as previously described.[27] The synthetic scheme can be found in SI Figure 4.2.

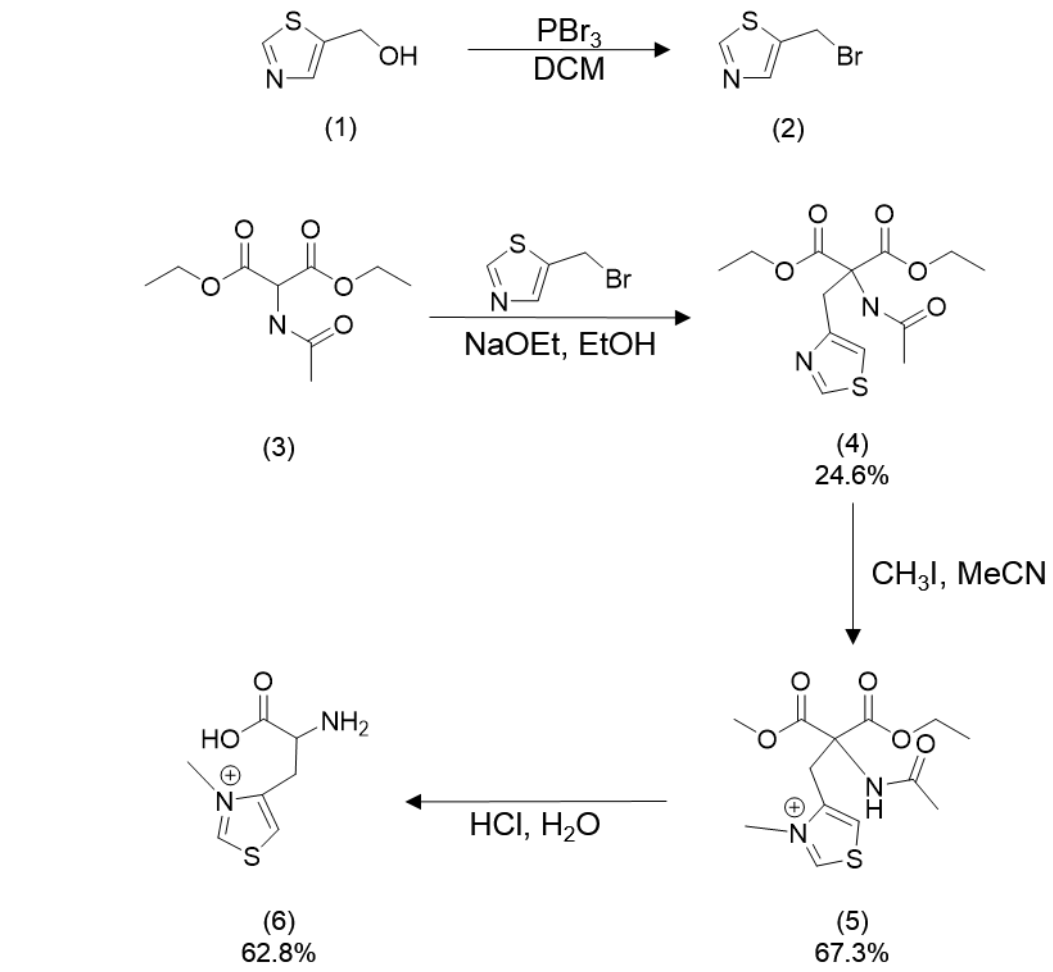
4.13 Supplementary Information

4.13.1 Supplementary Tables

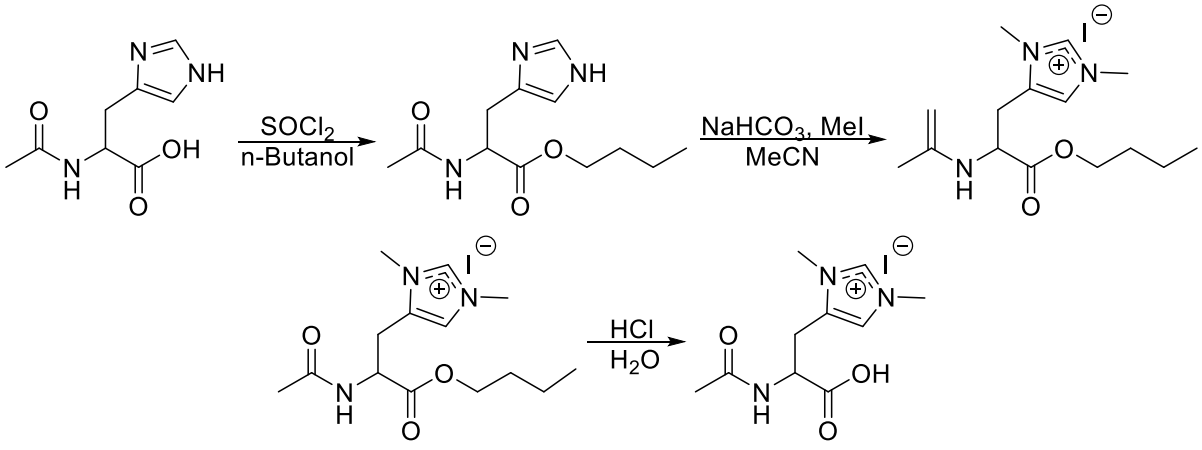
SI Table 4.1 Diluents of UAAs

UAA	Diluent
1MeHis	0.1 M NaOH
3MeHis	0.1 M NaOH
DMH	MilliQ Water
MeTAZ	MilliQ Water
BocK	0.1 M NaOH

4.13.2 Supplementary Figures



SI Figure 4.1 Synthetic scheme of MeTAZ. Yields of (2) could not be calculated due to high reactivity of the compound.



SI Figure 4.2 Synthetic scheme of dimethyl histidine.

REFERENCES

(1) Lutz, S. *Curr. Opin. Biotechnol.* **2010**, *21* (6), 734–743.

(2) Tao, H.; Cornish, V. W. *Curr. Opin. Chem. Biol.* **2002**, *6* (6), 858–864.

(3) Müller, M.; Gocke, D.; Pohl, M. *FEBS J.* **2009**, *276* (11), 2894–2904.

(4) Lingen, B.; Kolter-Jung, D.; Dünkelmann, P.; Feldmann, R.; Grötzinger, J.; Pohl, M.; Müller, M. *Chembiochem* **2003**, *4* (8), 721–726.

(5) Coelho, P. S.; Brustad, E. M.; Kannan, A.; Arnold, F. H. *Science* **2013**, *339* (6117), 307–310.

(6) Liu, C. C.; Schultz, P. G. *Annu. Rev. Biochem.* **2010**, *79*, 413–444.

(7) Kluger, R. *Chem. Rev.* **198**7 *87* (5), 863–876.

(8) Imperiali, B.; McDonnell, K. A.; Shogren-Knaak, M. In *Topics in Current Chemistry*; **1999** *202*, 1–38.

(9) Frank, R. A. W.; Leeper, F. J.; Luisi, B. F. *Cell. Mol. Life Sci.* **2007**, *64* (7–8), 892–905.

(10) Stamatis, A.; Malandrinos, G.; Louloudi, M.; Hadjiliadis, N. *Bioinorg. Chem. Appl.* **2007**, *2007*, 1-7.

(11) Schellenberger, A. *Biochim. Biophys. Acta* **1998**, *1385* (2), 177–186.

(12) Richter, M. *Nat. Prod. Rep.* **2013**, *30* (10), 1324–1345.

(13) Pohl, M.; Sprenger, G. A.; Müller, M. *Curr. Opin. Biotechnol.* **2004**, *15* (4), 335–342.

(14) Jordan, F. *Nat. Prod. Rep.* **2003**, *20* (2), 184–201.

(15) Mennen, S. M.; Blank, J. T.; Tran-Dubé, M. B.; Imbriglio, J. E.; Miller, S. J.; Tran-dube, M. B.; Imbriglio, J. E.; Miller, S. J. *Chem. Commun. (Camb).* **2005**, *2*, 195–197.

(16) Mennen, S. M.; Gipson, J. D.; Kim, Y. R.; Miller, S. J. *J. Am. Chem. Soc.* **2005**, *127* (6), 1654–1655.

(17) Imperiali, B.; Sinha Roy, R.; Walkup, G. K.; Wang, L. In *Molecular Design and Bioorganic Catalysis*; Wilcox, C., Hamilton, A., Eds.; Kluwer Academic Publishers: Netherlands, 1996; pp 35–52.

(18) Suckling, C. J.; Zhu, L.-M. *Bioorg. Med. Chem. Lett.* **1993**, *3* (4), 531–534.

(19) Xiao, H.; Peters, F. B.; Yang, P.-Y.; Reed, S.; Chittuluru, J. R.; Schultz, P. G. *ACS Chem. Biol.* **2014**, *9* (5), 1092–1096.

(20) Wang, Y.-S.; Fang, X.; Wallace, A. L.; Wu, B.; Liu, W. R. *J. Am. Chem. Soc.* **2012**, *134* (6), 2950–2953.

(21) Yanagisawa, T.; Ishii, R.; Fukunaga, R.; Kobayashi, T.; Sakamoto, K.; Yokoyama, S. *Chem. Biol.* **2008**, *15* (11), 1187–1197.

(22) Wang, L.; Brock, A.; Herberich, B.; Schultz, P. G. *Science* **2001**, *292* (5516), 498–500.

(23) Rackham, O.; Chin, J. W. *Nat. Chem. Biol.* **2005**, *1* (3), 159–166.

(24) Neumann, H.; Peak-Chew, S. Y.; Chin, J. W. *Nat. Chem. Biol.* **2008**, *4* (4), 232–234.

(25) Wang, Y.-S.; Fang, X.; Chen, H.-Y.; Wu, B.; Wang, Z. U.; Hilty, C.; Liu, W. R. *ACS Chem. Biol.* **2013**, *8* (2), 405–415.

(26) González-Bulnes, P.; González-Roura, A.; Canals, D.; Delgado, A.; Casas, J.; Llebaria, A. *Bioorg. Med. Chem.* **2010**, *18* (24), 8549–8555.

(27) Monney, A.; Venkatachalam, G.; Albrecht, M. *Dalton Trans.* **2011**, *40* (12), 2716–2719.