Title

Optimization of optogenetic proteins and protein-focused deep learning algorithms

Author

Halloran, Marianne Catanho

Publication Date

2018

UNIVERSITY OF CALIFORNIA, SAN DIEGO

Optimization of optogenetic proteins and protein-focused deep learning algorithms

A Dissertation submitted in partial satisfaction of the
requirements for the degree Doctor of Philosophy

in

Bioengineering

by

Marianne Catanho Halloran

Committee in charge:

- Professor Todd P. Coleman, Chair
- Professor Brian Head
- Professor Prashant Mali
- Professor Gentry Patrick
- Professor Arvind Ramanathan
- Professor Shankar Subramaniam

2018

The Dissertation of Marianne Catanho Halloran is approved, and it is acceptable in quality and form for publication on microfilm and electronically:

Chair

University of California, San Diego

2018

DEDICATION

To my mother, Teresa, my brother, Gabriel, and sister, Elise, whose words of encouragement I held dear throughout the years.

To my husband, Doug, whose caring, support and reassurance helped me stay focused, sane, and strong throughout some of the most trying years of my doctoral career.

To the Shumaker family, for their encouragement, support and generosity I wouldn’t have made it this far without them. I am forever grateful, awestruck, and humbled by their trust and investment in my life. Because of them, I have been able to live the greatest adventure, beyond my wildest hopes.

To my friend, Paul, who taught me humility and generosity, gratitude and compassion; and showed me how to live a life worth not only remembering but celebrating: Thank you.

TABLE OF CONTENTS

LIST OF ABBREVIATIONS

PCB	Phycocyanobilin
PΦB	Phytochromobilin
Fd	Ferredoxin
FNR	Ferredoxin-NADP+-Reductase
Fd+FNR	Endogenous oxidation-reduction system containing Fd and FNR
PcyA	Phycocyanobilin:ferredoxin oxidoreductase
Hy2	Phytochromobilin:ferredoxin oxidoreductase
BV	Biliverdin Ixα
HO1	Heme oxygenase
NIR	Near-infrared
PhyB	Phytochrome B
PIF	Phytochrome-interacting factor
PhyB·PCB	PCB-bound PhyB
m^2	meter squared
s	seconds
h	hours
nm	nanometer
HAN	Hierarchical Attention network
MSA	Multiple sequence alignment
BLSTM	Bidirectional Long-short term memory
LSTM	Long-short term memory
RNN	Recurrent neural network
CNN	Convolutional neural network

SVM	Support vector machine
SCA	Statistical coupling analysis
DCA	Direct coupling analysis
NCBI	National Center for Biotechnology Information

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

I would like to express my sincere gratitude to my adviser Dr. Todd P. Coleman, whose vision and support allowed me to expand my knowledge and pursue research in the most diverse topics since 2010. I would also like to thank my committee members for their support and guidance to help me fulfill the requirements of this degree. My gratitude to Dr. Phillip Kyriakakis, for his enthusiasm, drive, and unwavering encouragement. His ideas and spirit are at the core of the first chapters of this work. Thank you also to Dr. Arvind Ramanathan, for his dedication, selflessness and generosity in working with me in the deep learning aspects of this work. I sincerely appreciate the collaboration and support of my fellow labmates and colleagues: Nicole Hoffner, Walter Thavarajah, Vincent Jian-Yu Hu, Syh-Shiuan Chao, Athena Hsu, Vivian Pham, Ladan Naghavian, and Lara E. Dozier.

Chapter 1 and 2, in full, are a reprint of the material as it appears in ACS Synthetic Biology, 2018. “Biosynthesis of Orthogonal Molecules Using Ferredoxin and Ferredoxin-NADP+ Reductase Systems Enables Genetically Encoded PhyB Optogenetics" Phillip Kyriakakis, Marianne Catanho, Nicole Hoffner, Walter Thavarajah, Vincent Jian-Yu Hu, Syh-Shiuan Chao, Athena Hsu, Vivian Pham, Ladan Naghavian, Lara E. Dozier, Gentry Patrick and Todd P. Coleman. DOI: 10.1021/acssynbio.7b00413.

Chapter 3, in part, is currently being prepared for submission for publication of the material by Catanho, Marianne; Gao, Shang; Kyriakakis, Phillip; Coleman, Todd P.; Ramanathan, Arvind. “Discovering sequence coevolution signatures with hierarchical attention networks”. The dissertation author is the primary investigator and author of this material.

VITA

2009 Bachelor of Science in Electrical Engineering, University of Missouri - Columbia

2012 Master of Science in Electrical and Computer Engineering, University of Illinois at Urbana Champaign

2013-2014 Teaching Assistant, University of California, San Diego

2012-2018 Research Assistant, University of California, San Diego

2018 Doctor of Philosophy, University of California, San Diego

PUBLICATIONS

Kyriakakis, P*., Catanho, M.*, Hoffner, N., Thavarajah, W., Jian-Yu, V., Chao, S., Hsu, A., Pham, V., Naghavian, L., Dozier, L., Patrick, G., Coleman, T. (2018). *Biosynthesis of Orthogonal Molecules Using Ferredoxin and Ferredoxin-NADP+ Reductase Systems Enables Genetically Encoded PhyB Optogenetics*. ACS Synthetic Biology.

* These authors contributed equally to the work.

Brown, J., Caetano-Anolles, D., Catanho, M., Ekaterina, G., Ryckman, N., Tian, K., Voloshin, M., Gillette, R. (accepted December 2017). *Implementing goal-directed foraging decisions of a simpler nervous system in simulation*. eNeuro.

Catanho, M., Sinha, M., Mack, H., Coleman, T.P., Fraley, S. I. (2017) *Deep Learning framework for identification of bacteria with Universal-digital High-Resolution Melt and anomaly detection.* NIPS 2017 Women in Machine Learning Workshop (WiML 2017).

Catanho, M. (2012). *Electrophysiology, agent-based modeling and inverse optimal control applications in neuroethology* (Master dissertation, University of Illinois at Urbana

Champaign). Retrieved from https://www.ideals.illinois.edu/bitstream/handle/2142/31205/Catanho_Marianne.pdf

Hirayama, K., Catanho, M., Brown, J.W., and Gillette, R. (2012) *A core circuit module for cost/benefit decision*. Frontiers in Neuroscience 6:123.

Brown, J. W., Catanho, M., & Gillette, R. (2011). *Chemotactile integration in the peripheral nervous system of a predatory sea slug.* Integrative and Comparative Biology 51:E169-E169.

ABSTRACT OF THE DISSERTATION

Optimization of optogenetic proteins and protein-focused deep learning algorithms

by

Marianne Catanho Halloran

Doctor of Philosophy in Bioengineering

University of California, San Diego, 2018

Professor Todd Prentice Coleman, Chair

Light-responsive proteins enable control of biological processes with unprecedented precision, holding great promise for clinical and industrial applications. Introducing these proteins into cultured cells or tissues of live animals allows investigation and control of various cellular and organism functions, from neuronal activity, to intracellular signaling, gene expression and cell proliferation, for example. In this work, we take different approaches to the optimization of phytochromes and phytochrome-based optobiology tools. We also present a deep learning framework for protein biology with direct implications on identification of functionally relevant residues in phytochromes.

First, by co-expressing cyanobacterial enzymes, we show that it is possible to increase endogenous chromophore production. Chromophores are bilin molecules that covalently bind to phytochromes, enabling photoconversion. Endogenous production of chromophores is a key development for phytochrome its use in mammalian cells. We demonstrate the limiting factors in chromophore production are two of the required enzymes in the chromophore’s pathway, and not solely heme as previously reported. We show how stoichiometry and species-matching affect chromophore production, and how chromophore levels can impact the performance of phytochrome-based optogenetic systems. Next, we demonstrate the utility of coupling the endogenous chromophore pathway and a light-responsive module composed of cyanobacterial Phytochrome B (PhyB) and its interacting factor (PIF3) to control expression of reporter genes.

Finally, we present a deep learning framework to identify complex relationships inherent in multiple sequence alignments. We develop a Hierarchical Attention network (HAN) for protein sequence families (HANprot) and demonstrate its performance in terms of relevant residue matching. We also demonstrate its utility in finding relevant residues for PhyB, towards potential optimization of its photolabile properties. The residues identified by HANprot can be used as a starting point for further protein investigations when structural or database annotations are lacking.

INTRODUCTION

Optical control of biology holds great promise as a tool for studying gene function, developmental biology, gene therapies and tissue engineering (Müller, Engesser, Metzger, *et al.*, 2013). The exquisite temporal and spatial precision achieved through optics has been used to develop an assortment of tools to control biological functions such as gene expression, neural activity (Levskaya, Weiner, Lim, & Voigt, 2009), cell signaling (Zhang & Cui, 2015), secretion (Müller, Engesser, Metzger, *et al.*, 2013), and protein activity (Beyer *et al.*, 2015). Rapidly reversible, space- and time-resolved light-inducible expression systems are poised to become an important tool in the areas mentioned above, as well as translational medicine and biomedical applications (Shimizu-Sato, Huq, Tepperman, & Quail, 2002). Similar controllable gene expression systems, like chemically induced systems, suffer from poor spatiotemporal control and activity.

As a whole, the Phytochrome photoreceptor family of proteins has been utilized in several gene expression systems, in which light is used to induce conformational changes and make possible to direct photoregulation of gene expression and protein production in plant and animal cells (Auldridge & Forest, 2011; Beyer *et al.*, 2015; M. Chen, Tao, Lim, Shaw, & Chory, 2005; Hughes, Bolger, Tapadia, & Tucker, 2012; Li, Li, Wang, & Wang Deng, 2011; Quail *et al.*, 1995; Rockwell, Su, & Lagarias, 2006; Sakamoto & Nagatani, 1996; Shimizu-Sato *et al.*, 2002; von Horsten *et al.*, 2016). Within the phytochrome family, phytochrome B (PhyB) has the optical characteristics long sought after in optobiology: it requires minimal light for activation and absorbs light in the near-infrared (NIR) window (Li, Li, Wang, & Wang Deng, 2011; Shimizu-Sato, Huq, Tepperman, & Quail, 2002). PhyB-based switches have been shown to be very robust compared to other switches, but required external addition of a chromophore, limiting them to *in vitro* applications (Beyer *et al.*, 2015; Müller, Engesser, Metzger, *et al.*, 2013).

In this work, we approach optimization and mammalian *in vitro* application of PhyB as an optogenetic tool to control biology in a multifaced approach: genetic engineering and deep learning. In Chapter 1, we show that genetically encoding mammalian cells to produce the chromophores required of PhyB activity can be achieved by co-expressing Ferredoxin and Ferredoxin-oxyreductase (Fd and FNR, respectively) in mammalian cells. These results were confirmed both for cytoplasmic and mitochondrial production of chromophores in mammalian cells. This effectively removed the barriers for multiple uses *in vivo* and *in vitro* studies. Combined with the endogenous production of chromophores enabled by our results, a robust NIR gene switch was developed that is fully genetically encoded, as shown in Chapter 2. This optimized switch can control genes with low background, high dynamic range, and orders of magnitude less light than any other optogenetic system. This finding creates many new opportunities for engineering synthetic systems to produce these molecules, along with many others. The principles presented can be applied industrially to cost effective production of plant molecules in microbes or for drug delivery by genetically encoding the pathway to make therapeutic molecules.

Finally, in Chapter 3, we explore the complex dependency between a protein's sequence and its function. Functional divergence is often reflected in changes in evolutionary rate of a particular protein family (Chakrabarti, Bryant, & Panchenko, 2007). Those changes are hard to detect, being determined by small changes in a residue's stereochemistry. Physio-chemical mechanisms are often sought as answer to this problem, since several functional activities are based on a same region or fold within a protein family (Chakrabarti et al., 2007). Machine and deep learning, rapidly developing fields, have been shown to bring new perspectives to problems centered around complex relationships, such as the one between protein sequence, structure and function. Deep learning methods have been shown to achieve the good performance in residue-residue contact prediction and disorder prediction (Chakrabarti et al., 2007; Marsella, Sirocco, Trovato, Seno, & Tosatto, 2009; Walsh, Martin, Di Domenico, & Tosatto, 2012).

We propose a biologically-centered Hierarchical Attention Network (HAN) with two hierarchies, each composed of bidirectional long-short term memory layers with a softmax attention mechanism, to predict residues relevant to a protein's function, based solely on a familial sequence alignment. As such, this attention-based network for relation classification enables identification of important features and residues, and prediction of relevant residues for the protein's function. In the case of PhyB, identifying more key residues associated with the dynamics of how light triggers photoisomerization of the protein between its different conformational states could lead to more control and engineering of this system. These changes could lead to shifted wavelengths, length of stay in a conformational state, speed of reversal, etc., which can be used in several different applications, from gene therapy to understanding neurological disease with spatial, temporal, and cell type precision.

Our results show that HANprot is sensible to functional sites and its measure is significantly different from methods based on amino acid distribution and coevolution. The key difference to previous work is that HANprot discovers single residues or short sequences of relevant residues based on context and relies entirely on non-annotated inputs from multiple sequence alignments, rather than three-dimensional distances or position-specific scoring matrices. By focusing on the complex relationships within protein sequences through the application of deep learning algorithms, this work could be a gateway to explore perturbation analyses, drug response, and enzyme kinetics, among other possibilities.

Overall, this work aims to optimize and better understand of the fundamental mechanisms of photolabile proteins' interaction and resulting dynamics in response to light. Application of the methodologies and architectures proposed can amplify PhyB's usability and applicability in biological and biomedical studies, translational medicine, and potentially in biofuel engineering. Future work could enable modification and customization of the protein's photoswitchable properties towards shifted wavelengths, length of stay in a conformational state, speed of reversal, etc., enabling further optimization of our proposed light switch for different applications.

CHAPTER 1 CHROMOPHORE PATHWAY

Reproduced in part with permission from:

Kyriakakis, P., Catanho, M., Hoffner, N., Thavarajah, W., Jian-Yu, V., Chao, S.-S., … Coleman, T. (2018). Biosynthesis of Orthogonal Molecules Using Ferredoxin and Ferredoxin-NADP+ Reductase Systems Enables Genetically Encoded PhyB Optogenetics. ACS Synthetic Biology. doi:10.1021/acssynbio.7b00413

1.1 Abstract

The chromophores phycocyanobilin (PCB) and phytochromobilin (PΦB) are pigments used for photoreception in cyanobacteria and plants. Those chromophores operate in the near-infrared (NIR) range (600-900nm), a range ideal for use in optobiological manipulation since these wavelengths allow for maximal tissue penetration with minimal light phototoxicity. However, the genetic switches reliant on those chromophores are often used in mammalian cell lines, whose chromophore production is limited by the endogenous oxidation-reduction system containing Ferredoxin (Fd) and Ferredoxin-NADP+-Reductase (FNR) (Fd+FNR).

We show that by co-expressing the cyanobacterial Fd+FNR along with their interacting biosynthetic enzymes, endogenous chromophore production increases by over 20-fold. We delineated the rate limiting factors and found that the main metabolic precursor, heme, was not the primary limiting factor for producing either the cyanobacterial or plant chromophores, but that in fact Fd is limiting, followed by Fd+FNR and finally heme. Boosting chromophore production by matching metabolic pathways with specific

ferredoxin systems enables unparalleled use of many optogenetic tools and has broader implications for optimizing synthetic metabolic pathways.

1.2 Introduction

The transplantation of metabolic reactions from one species to another is an established research practice used in synthetic biology. Wide-ranging potential applications of this methodology include metabolic gene therapy(Gaspar *et al.*, 2011; X. Y. Zhou *et al.*, 1995), production of crops without fertilizer (Burén *et al.*, 2017; Shintani & DellaPenna, 1998), and more fundamental applications in research, such as optogenetics. The temporal and spatial precision achieved through optogenetics has been used to develop an assortment of powerful analytical tools to control biological functions such as gene expression (Folcher *et al.*, 2014; Kaberniuk, Shemetov, & Verkhusha, 2016; Müller, Engesser, Metzger, *et al.*, 2013; Pathak, Strickland, Vrana, & Tucker, 2014; Shimizu-Sato *et al.*, 2002; X. Wang, Chen, & Yang, 2012), neural activity (Boyden, Zhang, Bamberg, Nagel, & Deisseroth, 2005; John Y Lin, Knutsen, Muller, Kleinfeld, & Tsien, 2013), cell signaling (Levskaya *et al.*, 2009), secretion (D. Chen, Gibson, & Kennedy, 2013), peroxisomal trafficking (Spiltoir, Strickland, Glotzer, & Tucker, 2016), and protein activity (X. X. Zhou, Chung, Lam, & Lin, 2012). Metabolically engineering cells to endogenously produce specific chromophores enables many of those optogenetic applications, including genetically encoded systems for optical control of genes (Müller, Engesser, Timmer, *et al.*, 2013). Many of the systems used and characterized for these applications utilize proteins that require red and far-red responsive phytobilin chromophores like phycocyanobilin (PCB) and phytochromobilin (PΦB).

These molecules originate from phytochrome systems in cyanobacteria, algae, and plants, but are not naturally made in many fungal species, bacteria, or animal cells

(Auldridge & Forest, 2011; Karniol, Wagner, Walker, & Vierstra, 2005; Rockwell *et al.*, 2006; Rodriguez-Romero, Hedtke, Kastner, Müller, & Fischer, 2010). They are produced by the enzymes phycocyanobilin:ferredoxin oxidoreductase (PcyA) and phytochromobilin:ferredoxin oxidoreductase (Hy2), respectively, from Biliverdin IXα (BV), a degradation product of heme (Figure 1) (Beale, 1993; N Frankenberg, Mukougawa, Kohchi, & Lagarias, 2001; Hübschmann, Börner, Hartmann, & Lamparter, 2001; Terry, McDowell, & Lagarias, 1995). Several groups have shown that it is possible to produce these chromophores in *E. coli* by expressing PcyA or Hy2 without adding the matching ferredoxin (Fd) and ferredoxin-NADP$^+$-reductase (FNR) reduction system (Landgraf, Forreiter, Hurtado Picó, Lamparter, & Hughes, 2001; Mukougawa, Kanamoto, Kobayashi, Yokota, & Kohchi, 2006; Tooley, Cai, & Glazer, 2001).

Likewise, Müller *et al.* tested PCB production in mammalian cells by expressing PcyA and HO1, but there was no direct measurement of chromophore production (Müller, Engesser, Timmer, *et al.*, 2013). Müller *et al.* reasoned mitochondrial placement of PcyA and HO1 in the same cellular compartment where the chromophore precursor (heme) is produced would enhance PCB production (Müller, Engesser, Timmer, *et al.*, 2013). However, because mammalian cells also express Fd and FNR (Fd+FNR) exclusively in the mitochondria, those experiments did not address the possibility that PCB production failed to occur in the cytoplasm because of the mitochondrial localization of Fd+FNR. However, in addition to heme, HO1, PcyA, and HY2 also depend on Fd activity, leaving open the possibility that Fd and not heme was limiting.

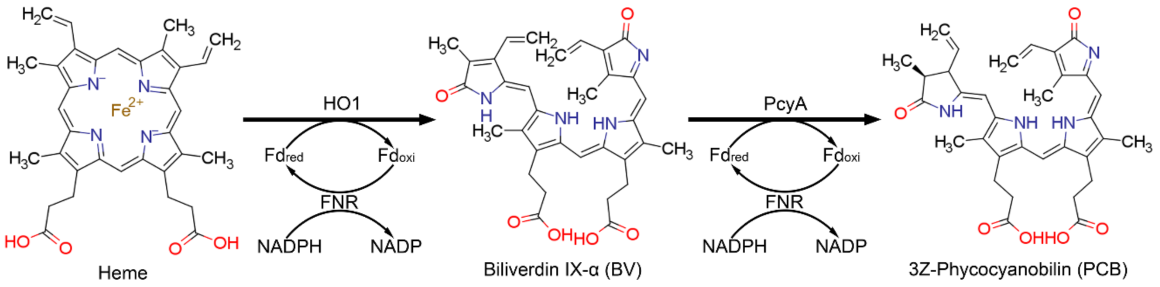


Figure 1: PCB metabolic production pathway. The metabolic pathway for PCB synthesis including the NADPH/FNR/Fd redox cascade (Heme: ChemSpider ID 4802, Bv: ChemSpider ID 10628548, PCB: ChemSpider ID 16736730).

Frankenberg *et al.* demonstrated *in vitro* that Fd activity on PcyA from *Anabaena* sp. PCC 7120 varies greatly depending on the Fd species (Nicole Frankenberg & Lagarias, 2003). Beale *et al.* and Frankenberg *et al.* demonstrated that Fd activity on PcyA from *Anabaena* sp. PCC 7120 varies greatly depending on the species Fd comes from (Beale, 1993; Nicole Frankenberg & Lagarias, 2003). Similarly, mammalian Fds have also been shown to be highly specific to their target enzymes, suggesting that Fd and/or FNR may be limiting for chromophore production in mammalian cells (Aliverti, Pandini, Pennati, de Rosa, & Zanetti, 2008; Sheftel *et al.*, 2010). Most cells already contain endogenous Fd; therefore, researchers have not typically considered it when transplanting enzymes from one species to another. Consequently, to increase production of molecules like PCB for optogenetic uses in animal cells, we investigated the limiting factors for the PCB and PΦB production in mammalian cells.

Because mammalian Fds have also been shown to be highly substrate- and tissue-specific, it was possible that mammalian Fds may not be efficient replacements for cyanobacterial or plant ferredoxins (Matsubara & Saeki, 1992; Sheftel *et al.*, 2010). This remained untested and may be important for the production of many plant and bacterial molecules in other cells, or generally when introducing metabolic pathways from one species to another. Moreover, since Fds are the some of the most electronegative proteins in metabolic pathways (Matsubara & Saeki, 1992), introducing the matching Fd for a orthogonal biosynthetic pathway could be key for efficiently producing a wide array of molecules including lipids, sterols, dolichols, luciferins, quinones, carotenoids, nitrates/nitrogen, and sulfites (Burén *et al.*, 2017; Curatti & Rubio, 2014; G. Hanke & Mulo, 2013; Pinto, Harrison, Hsu, Jacobs, & Leyh, 2007; Rekittke *et al.*, 2013; Yonekura-Sakakibara *et al.*, 2000).

Using PCB and PΦB as examples, we show that by species matching the Fd+FNR system, it is possible to produce over one order of magnitude higher levels of PCB compared to relying on endogenous Fd+FNR. This highlights the importance of our finding that the availibility of electrons in the biosynthetic pathway are important considerations in synthetic biology. Production of molecules from one species in another can be used to deliver plant molecules (*e.g.* steroids or lipids) through human gene therapy, produce bacterial molecules in plants or a number of molecules in bioreactor friendly species. We demonstrate the utility of coupling matching reduction systems, such as Fd+FNR, with the PCB metabolic pathway by developing a phytochrome based tool to control biological processes with NIR light in mammalian cells. In addition, to evaluate the rate-limiting reactants for endogenous chromophore production, we systematically tested each component of the biosynthetic pathway, including Fd and FNR. We showed that Fd+FNR is the primary rate-limiting component, followed by heme. The increased PCB production found with the addition of Fd+FNR was further improved by testing different stoichiometric expression levels of each enzyme. Endogenous PCB production was greatly increased compared to previous approaches (Müller, Engesser, Timmer, *et al.*, 2013) that did not consider metabolic engineering with Fd+FNR systems.

Using PySB(Lopez, Muhlich, Bachman, & Sorger, 2013), we generated an *in silico* model to describe the biochemical interactions among the enzymes that compose the hypothesized PCB-production pathway, as seen in Figure 1. The quantitative mathematical model was parametrized (Appendix C) by experimental data and uses ordinary differential equations to describe the changes in the concentration of the molecular components of the reaction. We probed the proposed model directly as proposed in the literature and similar pathways published (Gambetta & Lagarias, 2001; Müller, Engesser, Timmer, *et al.*, 2013; Okada, 2009). We complement this work showing the model’s agreement with the tested pathway, demonstrating how heme, Fd, and FNR are rate limiting factors for the production of PCB, as confirmed experimentally in Figures 2, 3, 4.and 5.

More generally than optogenetics, there are numerous biomolecules produced in bacteria and plants that are Fd-dependent. Matching the Fd species to a biosynthetic production pathway makes possible the metabolism of many other classes of molecules such as lipids, sterols, luciferins, quinones, carotenoids, nitrates/nitrogen, and sulfites not normally produced in those cells (Burén *et al.*, 2017; Cahoon & Shanklin, 2000; Curatti & Rubio, 2014; G. Hanke & Mulo, 2013; Pinto *et al.*, 2007; Rekittke *et al.*, 2013; Yonekura-Sakakibara *et al.*, 2000). Increasing product ion of these classes of molecules can improve agriculture, increase the production of pharmaceuticals, and enable other tools for synthetic biology.

1.3 Methodology

1.3.1 Marvin

Marvin was used for drawing and displaying chemical structures in Figure 1. Marvin 17.28.0, 2017, ChemAxon (http://www.chemaxon.com).

1.3.2 Zinc-PAGE-Immunoprecipitation Assays

Protein G PLUS-Agarose (ThermoFisher, 22851) beads were prepared by adding 200µg anti-HA (clone HA-7, Sigma H9658) into 2ml 25% agarose. After overnight binding at 4°C, unbound anti-HA was washed off four times with 1X Phosphate-buffered Saline (PBS, pH 7.4, ThermoFisher, 10010023). For each 6-well plate, 500,000 HEK293 cells (ATCC, CRL-1573) were transfected using 2.5µg DNA and 6µl of Lipofectamine 2000 per well (ThermoFisher Scientific, 11668019). For heme experiments, media or media containing 10µM heme (Frontier Scientific, H651-9), was exchanged 18 hours after transfection and again 43 hours after transfection. Heme was dissolved at 10 mM in 100 mM NaOH and sterile filtered with a 0.22µM filter (Millipore, SLGP033RS). Cells were then harvested with RIPA buffer (1% Triton X-100, 0.5% Sodium Deoxycholate, 25 mM Tris pH8.0, 150 mM NaCl, 0.10% SDS and 2.5 mM EDTA, and 2X protease inhibitors (Sigma, P8340-1ML), immediately placed on ice, sonicated briefly and then centrifuged for 30 minutes at 21,000g. BCA assays (ThermoFisher Scientific, 23225) were used to determine the protein concentration of resulting supernatant/lysates. Equal masses for each protein sample were diluted with two parts of cold PBS, then loaded onto Protein G PLUS-Agarose beads containing anti-HA (preparation above), for overnight binding while mixing at 4°C.

Next beads were washed and boiled in sample buffer (30% glycerol, 10% SDS, 300 mM Tris pH 6.8, 0.03% Bromophenol Blue, 179 mM 2-Mercaptoethanol). After loading and running the samples in a SDS-PAGE gel, the gels were incubated in SDS-

PAGE Running Buffer (25 mM Tris, 192 mM glycine, 0.1% SDS) containing 10 mM Zinc Acetate for 10 minutes prior to imaging in a Fluorochem E (Protein Simple). Gels were then transferred onto nitrocellulose and probed with the primary antibody anti-HA 1:5000 (Sigma, clone HA-7, H9658), and by Goat anti-Mouse secondary antibody 1:5000 (ThermoFisher, 32230). Western blots were imaged in a Fluorochem E (Protein Simple). Gel bands were quantified using the FIJI (ImageJ) gel analysis tool (Schindelin *et al.*, 2012).

1.3.3 Imaging PCB Production

HEK293 cells (ATCC, CRL-1573), plated at 100,000 cells per well in a 24-well plate, were transfected 24 hours after plating on polylysine (Sigma P6407-5mg) coated coverslips in each well. Forty-three hours later, the media was exchanged with fresh media or media+5µM PCB (Frontier Scientific, P14137) for the NE+PCB control. One hour later, cells were rinsed in 1X PBS and then fixed in 4% Paraformaldehyde in 1X PBS for 10 minutes. Cells were then washed with 1X PBS before incubating in permeabilization buffer (5% BSA + 0.3% TritonX-100 in PBS) for 30 minutes, followed by incubating with primary antibodies, anti-flag mouse monoclonal 1:1000 (Sigma, F3165) and polyclonal anti-HA rabbit 1:500 (Santa Cruz, Y-11) in antibody buffer (2% BSA + 0.2% TritonX-100 in PBS) at 4°C overnight. Next coverslips were rinsed twice and washed three times in 1X PBS and then incubated in antibody buffer containing goat anti-mouse AlexaFluor 488 1:1000 (ThermoFisher, A11001), and goat anti-rabbit AlexaFluor 568 1:1000 (ThermoFisher, A11011). Coverslips were rinsed and washed again, then mounted with Fluoromount-G (SouthernBiotech, 0100-20). Images were taken using a DeltaVision RT Deconvolution Microscope (Figure 9).

1.3.4 Cell Culture, Transfection, Light Induction and Reporter Gene Assays

Human Embryonic Kidney 293 cells (HEK293, ATCC CRL-1573) were cultivated in Dulbecco's Modified Eagle Medium (DMEM, Gibco, 11965-092) supplemented with 10% fetal bovine serum (FBS, Omega Scientific, FB-02) and 100 U/ml of penicillin and 0.1 mg/ml of streptomycin (Gibco, 11548876). All cells were cultured under 5% CO2 at 37°C. Cells were seeded at 100,000 HEK293 cells per well in 24-well plates, 24 hours before transfection. Transfection of plasmids was achieved through lipofection following the manufacturer's instructions and protocol (Lipofectamine 2000, ThermoFisher, 11668019). For each transfection reaction, a total of 0.5 μg of plasmid DNA was combined with specific plasmid ratios for each experiment as detailed in Appendix A and B. A construct with Renilla luciferase reporter plasmid DNA was included as an internal transfection control in all transfections. The culture medium was replaced with fresh medium 24 hours after transfection and the plates were placed inside black boxes (Hammond Manufacturing Company, 1591ESBK) for the remainder of the experimental procedure.

1.3.5 Luciferase Activity Assay

Luciferase assays were carried out using the Dual-Luciferase Assay system (Promega, PRE1960), and following the manufacturer's protocol. Cells were lysed immediately after removing from the incubator using the manufacturer's instructions. Firefly and Renilla Luciferase activities were measured from cell lysates using the luminometer module of the Infinite 200 PRO multimode reader (Tecan). Results of luciferase activity assays are expressed as a ratio of firefly luciferase (Fluc) activity to Renilla luciferase (Rluc) activity.

1.3.6 Illumination Circuits and Software

To obtain programmable control needed to drive the high-power LEDs used in our experiments, we designed the light control system shown in Figure 5. The light control system employs an Arduino Uno and a light intensity control circuit driven by a user interface developed in LabVIEW (National Instruments) to control each box's LED intensity (Figure 6). This system is ideal for precise timing and light-intensity control of each experimental box while allowing for user-determined experimental start delay, illumination frequencies, and control of the total duration of the experiment. Using this system, we have precise timing and light-intensity control for 8 experimental boxes that required red and/or far-red illumination. Each black box can house a standard 6-well, 12-well, 24-well, 96-well plate or can be fitted for a single dish with minimum modifications. The system can be replicated for experiments requiring a larger number of boxes or experimental conditions. Far-red and red lights can be controlled independently if placed in the same box. For our experimental setup, boxes contained either far-red 735nm LEDs or red 660nm LEDs. The light control system employs: (a) an Arduino Uno and voltage regulation circuits, managed through a (b) user interface developed in LabVIEW (National Instruments).

The voltage regulation circuit is shown in Figure 6. Coupled with the Arduino signals, this system delivers light pulses with precise timing and intensity control to the experiment boxes. The circuit is build using a LM317T linear voltage regulator (STMicroelectronics), a NPN general-purpose amplifier (2N2222, Fairchild Semiconductors), a resistor and a trimmer potentiometer (Helitrim, model 75PK10K). An external power supply was outfitted for the circuit (Safety Mark, 12V 1.5A Switch-mode power supply). The power supply allows the circuit to vary its current and voltage needs depending upon the intensity chosen by a user using the trimmer potentiometer.

The LabVIEW user interface, available for download at https://github.com/mcatanho/Kyriakakis_et_al_SupplementaryFiles (See Supplementary Note), controls the Arduino and connected circuits. It allows the user to connect to the Arduino effortlessly and to control experimental conditions such as time delay before illumination, a total duration of sample illumination, and pulse frequencies for each individual illumination box. It also contains digital displays of all relevant experimental times.

Figure 2: Plasmid used to test for PCB expression under different species of PcyA. HEK293 cells were analyzed for phytobilin production using the plasmids shown. Phytobilin production was measured by covalent linkage to PhyB followed by immunoprecipitation with anti-HA, Zn-PAGE and western blots. sPCYA and tPCYA produce PCB and aHY2 produces PΦB. Cells were either transfected with two ferredoxin-dependent enzymes (ho1 and pcyA or ho1 and HY2) alone (condition M2) or along with matching Fd+FNR (tpetF+tpetH) plasmids (condition M4*). ho1 = heme oxygenase, pcyA = phycocyanobilin:ferredoxin oxidoreductase, HY2 = phytochromobilin:ferredoxin oxidoreductase, petF = ferredoxin, petH = ferredoxin:oxidoreductase/FNR*, NE = No Enzymes, SYNP2= *Synechococcus* PCC7002 and THEEB= *Thermosynechococcus elongatus*, ARATH= *Arabidopsis thaliana*, MTS = Mitochondrial Targeting Sequence, P2A = 2A self-cleaving peptide, IRES = Internal Ribosome Entry Site, NLS = Nuclear Localization Sequence, DBD = DNA Binding Domain.

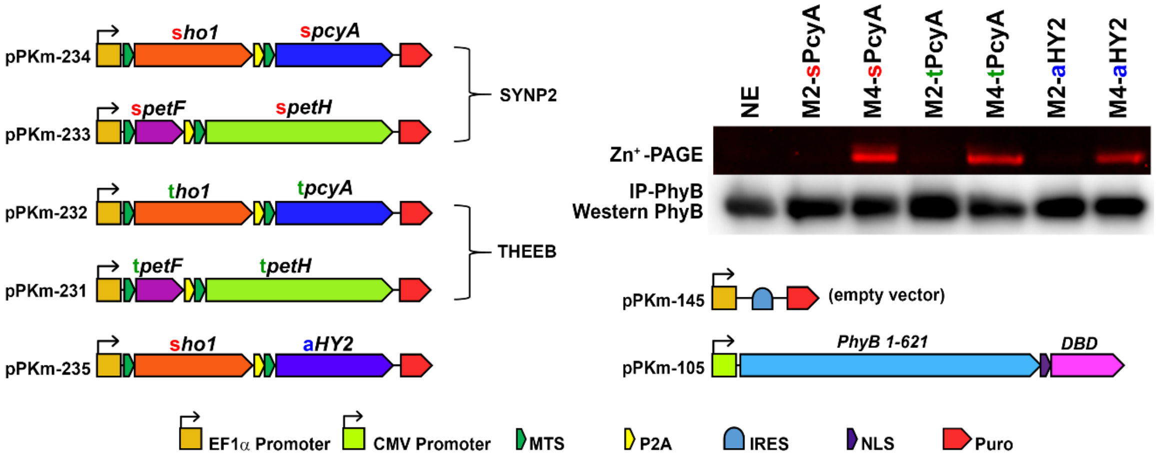
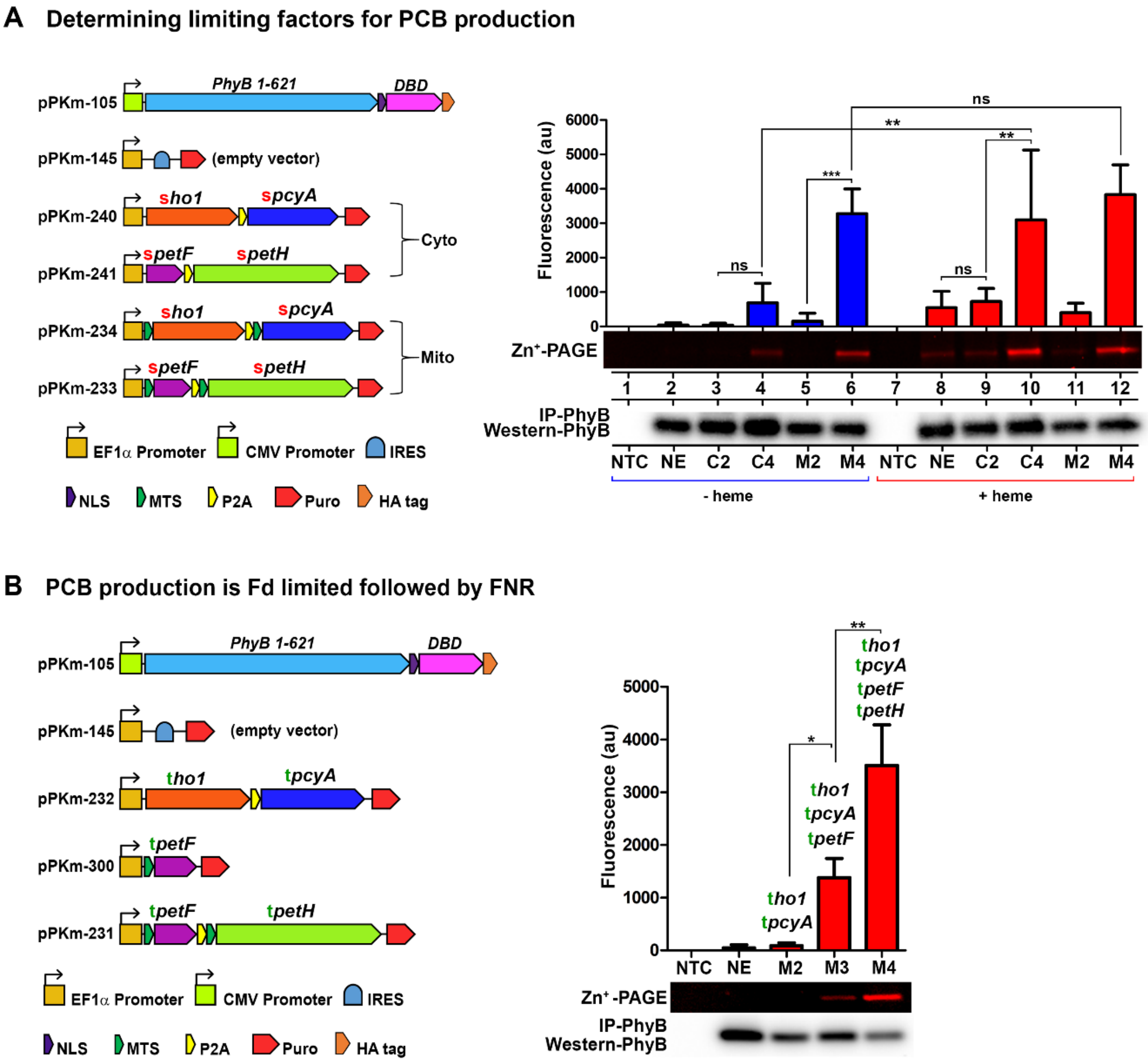

pPKm-234
sho1
spcyA
pPKm-233
spetF
spetH
SYNP2
pPKm-232
tho1
tpcyA
pPKm-231
tpetF
tpetH
THEEB
pPKm-235
sho1
aHY2
NE
M2-sPcyA
M4-sPcyA
M2-tPcyA
M4-tPcyA
M2-aHY2
M4-aHY2
Zn+-PAGE
IP-PhyB
Western PhyB
pPKm-145
(empty vector)
pPKm-105
PhyB 1-621
DBD
EF1α Promoter
CMV Promoter
MTS
P2A
IRES
NLS
Puro

Figure 3: Order of rate limiting factors of PCB production in mammalian cells.(A-B) HEK293 cells were analyzed for PCB production using the plasmids shown. PCB production was measured by covalent linkage to PhyB followed by immunoprecipitation with anti-HA, Zn-PAGE and western blots. (A) PCB production was compared with excess (+heme) and without (-heme), using the cytoplasmic expression of pcyA+ho1 alone (condition C2) or with cytoplasmic pcyA+ho1 +fd+fnr (condition C4); mitochondrial expression of pcyA+ho1 alone (condition M2) or with mitochondrial pcyA+ho1 +fd+fnr (condition M4). (n = 4) (B) Cells were either transfected with two ferredoxin-dependent enzymes alone, ho1 and pcyA (condition M2), or along with a matching fd:tpetF (condition M3) or along with matching fd+fnr:tpetF + tpetH (condition M4). (n = 4)
ho1 = heme oxygenase, *pcyA = phycocyanobilin:ferredoxin oxidoreductase*, *HY2 = phytochromobilin:ferredoxin oxidoreductase*, *petF = ferredoxin/fd*, *petH = ferredoxin:oxidoreductase/fnr*, NE = No Enzymes, SYNP2= Synechococcus PCC7002 and THEEB= Thermosynechococcus elongatus, ARATH= *Arabidopsis thaliana*, IRES = Internal Ribosome Entry Site, NLS = Nuclear Localization Sequence, MTS = Mitochondrial Targeting Sequence, P2A = 2A self-cleaving peptide, DBD = DNA Binding Domain. (One-way ANOVA with Bonferroni post-test was used to calculate p values using GraphPad Prism 5.01. (*) = p<0.05, (**) = p<0.01, (***) = p<0.001 Error bars = Standard Deviation, n = independent experiments).



A Determining limiting factors for PCB production

B PCB production is Fd limited followed by FNR

Figure 4: Stoichiometry of PCB production constructs. (A) PCB production assay comparing plasmid ratios of pcyA+ho1 to fd+fnr using the plasmids shown. Transfection ratios are indicated in boxes below the western blot. PCB production was measured by covalent linkage to PhyB followed by immunoprecipitation with anti-HA, Zn-PAGE and western blots. (B) Schematic of the PhyB/PIF33 light switch. PhyB is fused to a DNA Binding Domain (DBD) and bound to a light-sensitive chromophore (PCB). The PhyB-DBD fusion remains bound to the UAS promoter. PIF3 is fused to an Activation Domain (AD). Upon absorption of a red photon (660nm), PhyB changes conformation and recruits PIF3 to the promoter region. The AD fused to PIF3 then activates the gene downstream of the promoter. Upon absorption of a far-red photon (735nm), PhyB changes conformation that leads to PIF3 unbinding, removing the AD from the promoter, shutting the downstream gene off (C) Plasmid maps for endogenous PCB production and PhyB/PIF33 light switchable promoter. (D) Luciferase gene activation levels using endogenously produced PCB with several ratios of pcyA+ho1:petF+petH (n=3). ho1 = heme oxygenase, pcyA = Phycocyanobilin:ferredoxin oxidoreductase, petF = ferredoxin, petH = ferredoxin:oxidoreductase/FNR, MTS = Mitochondrial Targeting Sequence, P2A = 2A self-cleaving peptide, NLS = Nuclear Localization Sequence, IRES = Internal Ribosome Entry Site, AD = Activation Domain, DBD = DNA Binding Domain, R/FR = Red light/Far-red light. Error bars = Standard Deviation, (*) = $p<0.05$, (**) = $p<0.01$. Statistics were calculated using one-way ANOVA with Bonferroni post-test using GraphPad Prism 5.01. n = individual experiments.

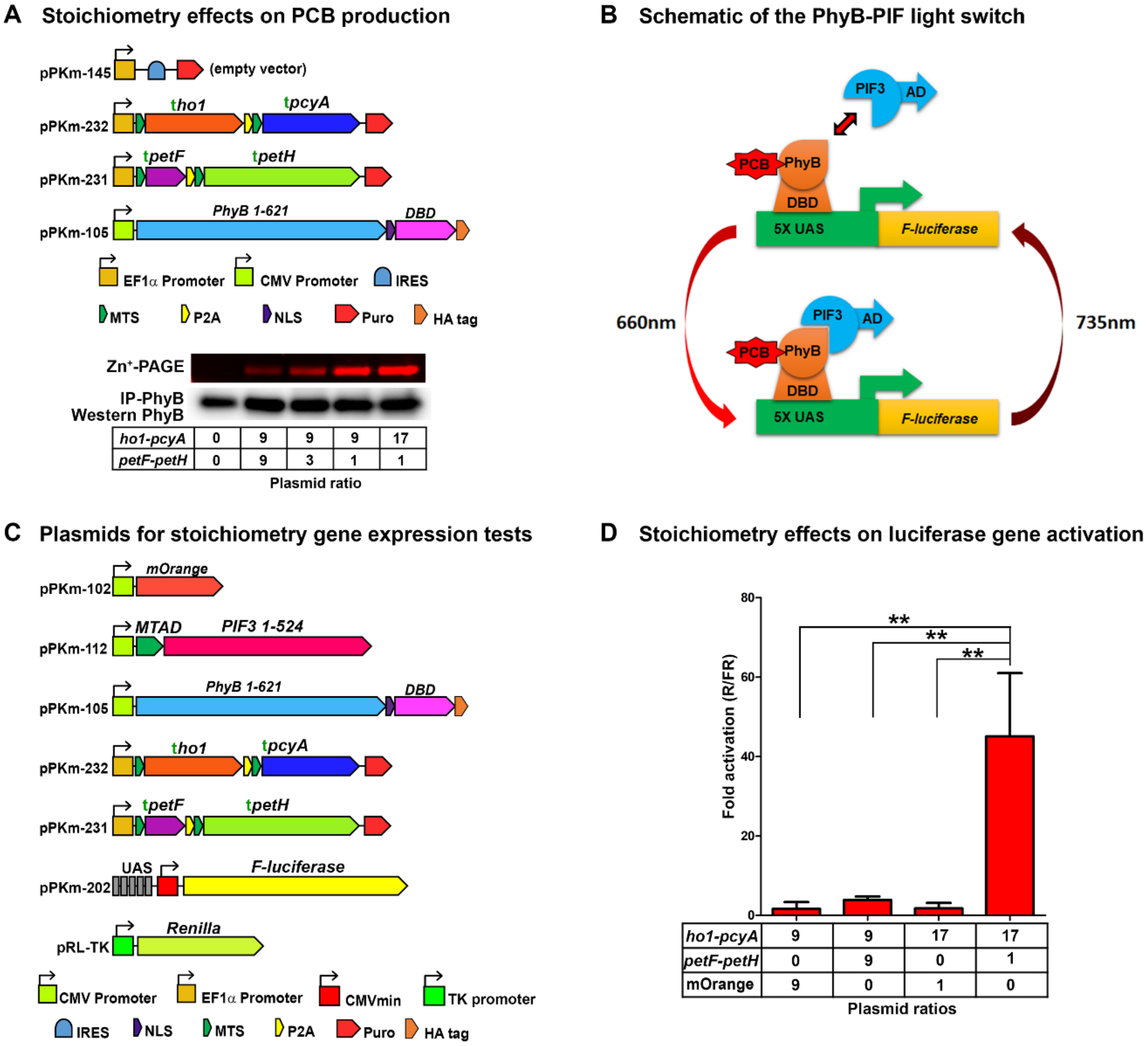
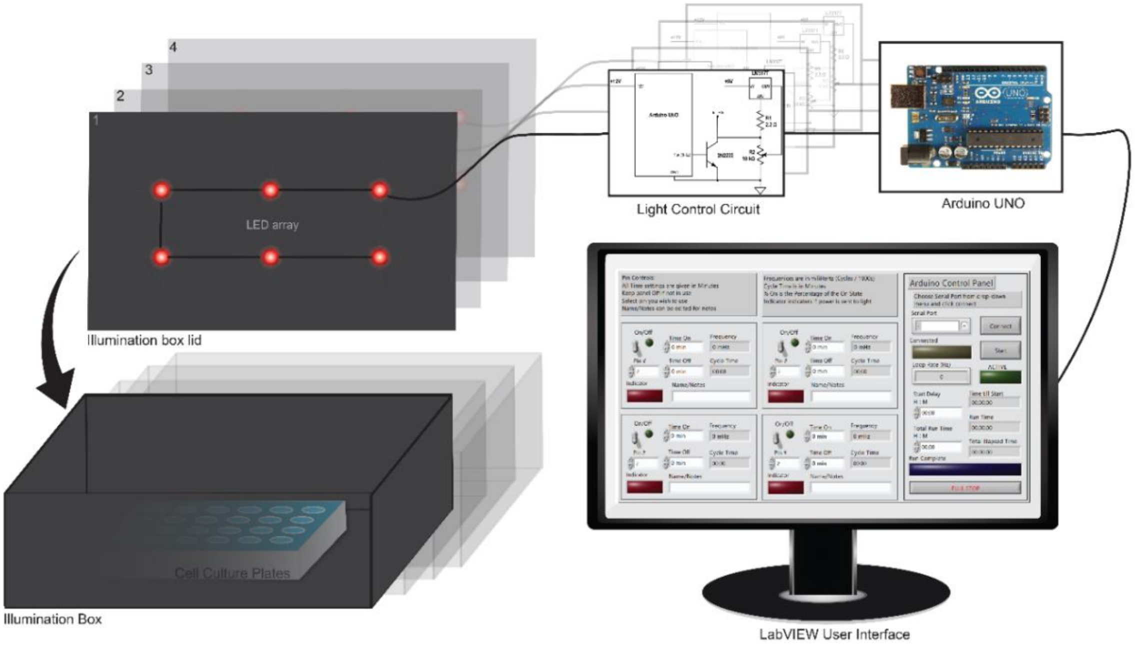

A Stoichiometry effects on PCB production
pPKm-145
(empty vector)
pPKm-232
tho1
tpcyA
pPKm-231
tpetF
tpetH
pPKm-105
PhyB 1-621
DBD
EF1α Promoter
CMV Promoter
IRES
MTS
P2A
NLS
Puro
HA tag
Zn+-PAGE
IP-PhyB
Western PhyB
ho1-pcyA 0 9 9 9 17
petF-petH 0 9 3 1 1
Plasmid ratio
B Schematic of the PhyB-PIF light switch
PIF3
AD
PCB
PhyB
DBD
5X UAS
F-luciferase
660nm
735nm
C Plasmids for stoichiometry gene expression tests
pPKm-102
mOrange
pPKm-112
MTAD
PIF3 1-524
pPKm-105
PhyB 1-621
DBD
pPKm-232
tho1
tpcyA
pPKm-231
tpetF
tpetH
pPKm-202
UAS
F-luciferase
pRL-TK
Renilla
CMV Promoter
EF1α Promoter
CMVmin
TK promoter
IRES
NLS
MTS
P2A
Puro
HA tag
D Stoichiometry effects on luciferase gene activation
Fold activation (R/FR)
80
60
40
20
0
**
**
**
ho1-pcyA 9 9 17 17
petF-petH 0 9 0 1
mOrange 9 0 1 0
Plasmid ratios

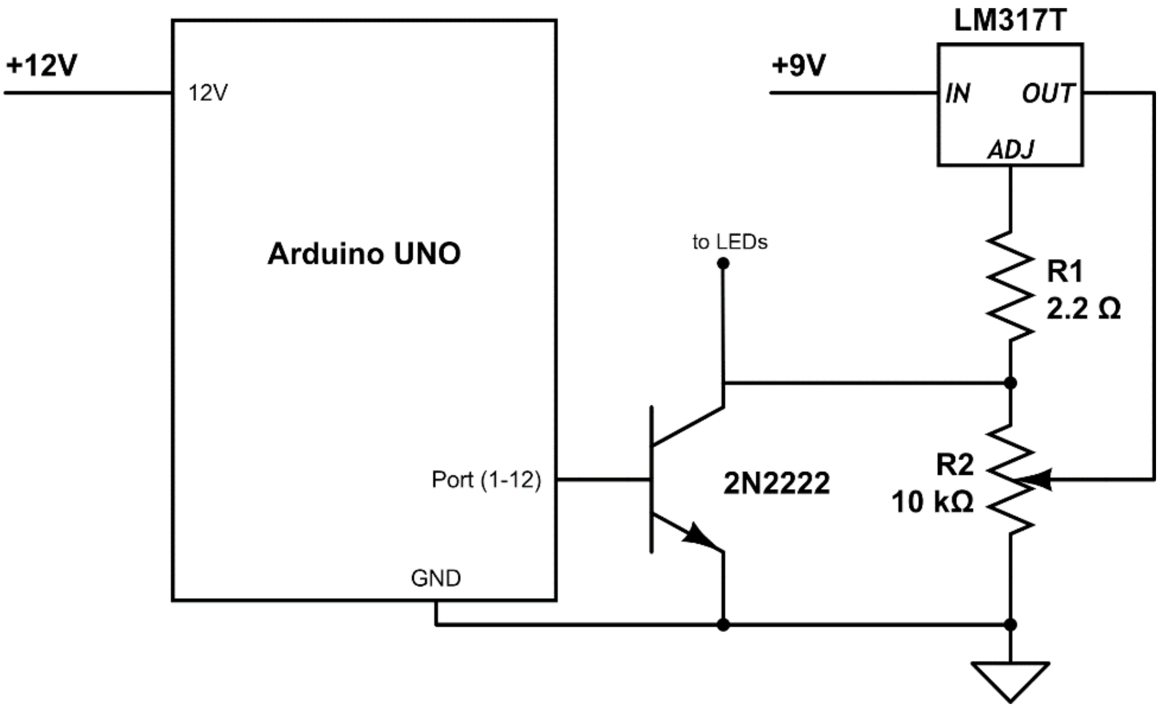
Figure 5: Illumination setup consists of black boxes with LED arrays controlled via an Arduino-driven circuitry and a LabVIEW user interface. The system is easily expandable to allow for the control of up to 12 boxes simultaneously. Each box can be activated at different frequencies.


Pin Controls:
All Time settings are given in Minutes
Keep panel Off if not in use
Select pin you wish to use
Name/Notes can be edited for notes

Frequencies are in milliHertz (Cycles / 1000s)
Cycle Time is in Minutes
% On is the Percentage of the On State
Indicator indicates if power is sent to light

Choose Serial Port from drop-down menu and click connect

Figure 6: Circuit Design for LED illumination. Electronic schematic of the circuit used to control the LEDs for each box, coupled with an Arduino UNO. The circuit requires a 9 Volt voltage source and uses simple components. A trimmer potential allows for intensity and brightness control of the LEDs. This circuit can control 6 high power LEDs in series.



+12V
12V
Arduino UNO
Port (1-12)
GND
to LEDs
2N2222
+9V
LM317T
IN
OUT
ADJ
R1
2.2 Ω
R2
10 kΩ

1.3.7 Kinetic Model

We demonstrate the biochemical interactions among the enzymes shown in Figure 1 in the production of PCB through a kinetic model developed with the PySB framework (Lopez *et al.*, 2013). The model's code, equations, and simulation files are available for download at https://github.com/mcatanho/Kyriakakis_et_al_SupplementaryFiles. The quantitative mathematical model was parametrized (Appendix C) by experimental data and uses ordinary differential equations to describe the changes in the concentration of the molecular components of the reaction.

For the model, we assume that the production of PCB can be described by the set of sequential steps detailed in Section 1.3.8, and depicted in Figure 1. This kinetic model builds upon Tu *et al.* description of the four electron reduction of biliverdin IX-alpha (BV) to phycocyanobilin (PCB), catalyzed by cyanobacterial phycocyanobilin:ferredoxin oxidoreductase (PcyA) (Tu, Gunn, Toney, Britt, & Lagarias, 2004). As demonstrated experimentally in this work, the ferredoxin (Fd) and ferredoxin:oxidoreductase (FNR) complex is of paramount importance to the redox metabolism in plants and cyanobacteria, working as an electron transfer complex to reduce or oxidize enzymes in different pathways, further acting to reduce or NADP+ to NADPH or the reverse of this reaction (Batie & Kamin, 1984; G. Hanke & Mulo, 2013; G. T. Hanke, Kurisu, Kusunoki, & Hase, 2004). As described in Figure 1, the first step in the PCB production pathway involves the formation of the HO1:Heme complex, which receives electron transfers from reduced ferredoxin (Fd_{red}), producing BV (Okada, 2009). Following a PcyA:BV complex is formed, which in turn also receives electron transfers from Fd_{red}, leading to the production of PCB. As the preferred electron donor for HO1 and PcyA, reduced Fd allows for continuous turnover of those enzymes in the PCB production pathway (Okada, 2009).

The reactions described above to produce PCB are shown in Section 1.3.8. The model assumes that those molecules are present *in vitro* at stoichiometry levels compatible with our transient transfection plasmid ratio. For simplicity, the model ignores differences in overall expression and degradation of each enzyme. Our model does not assume degradation of heme or BV, since we assumed there were saturating amounts in the cell medium. We also assume that the oxidized ferredoxin, a result of the electron transfer to the HO1:Heme and PcyA:BV complexes, is renewed in the NADP$^+$/NADPH pathway catalyzed by FNR. We probed the proposed model directly as proposed in literature (Nicole Frankenberg & Lagarias, 2003; Müller, Engesser, Timmer, *et al.*, 2013; Tu *et al.*, 2004), and similar pathways published. We complement this work showing the model's agreement with the hypothesized pathway, confirming that in the presence of heme, Fd and FNR are the rate limiting factors to produce PCB, confirmed experimentally in Figure 2. We also show in Figure 1, how PCB's production dependence on Heme and the NADP/NAPDH pathway, characterized by the presence of Fd and FNR, are interlinked.

1.3.8 Design and Parametrization of the Mathematical Model

Coupled, first order, ordinary differential equations (ODEs), parametrization of the model was performed using previously reported endogenous PCB production curves (Müller, Engesser, Timmer, *et al.*, 2013). The reaction schemes below were translated into the PySB rule-based language. Rates were calculated through a parametric sweep method utilizing maximum-likelihood minimization for model-fitting procedures. The rule-based model simulates PCB production, following the reactions described in below.

(1) Formation of the Heme and HO1 complex

(2) Formation of Fd_{red}:HO1:Heme complex, electron transfer from Fd_{red}, producing BV

(3) Formation of the BV:PcyA complex

(4) Fd_{red}:PcyA:BV complex formation, and electron transfer from Fd_{red}, producing PCB

(5) FNR-enabled Fd reduction

(6) Spontaneous degradation of PCB, as described by Mueller *et al* (Müller, Engesser, Timmer, *et al.*, 2013).

$$Heme + HO1 \overset{k_1}{\underset{k_2}{\rightleftarrows}} HO1{:}\,Heme \tag{1}$$

$$HO1{:}\,Heme + Fd_{red} \overset{k_3}{\underset{k_4}{\rightleftarrows}} Fd_{red}{:}\,HO1{:}\,Heme \tag{2}$$

$$Fd_{red}{:}\,HO1{:}\,Heme \overset{k_5}{\rightarrow} HO1{:}\,BV + Fd_{oxi} \tag{3}$$

$$HO1{:}\,BV \overset{k_6}{\rightarrow} BV + \; HO1 \tag{4}$$

$$BV + PcyA \overset{k_7}{\underset{k_8}{\rightleftarrows}} PcyA{:}\,BV \tag{5}$$

$$PcyA{:}\,BV + Fd_{red} \overset{k_9}{\underset{k_{10}}{\rightleftarrows}} Fd_{red}{:}\,PcyA{:}\,BV \tag{6}$$

$$Fd_{red}{:}\,PcyA{:}\,BV \xrightarrow{k_{11}} PcyA{:}\,PCB + Fd_{oxi} \tag{7}$$

$$PcyA{:}\,PCB \xrightarrow{k_{12}} PCB + PcyA \tag{8}$$

$$Fd_{oxi} \xrightarrow{k_{13}} Fd_{red} \tag{9}$$

$$PCB \xrightarrow{k_{deg,PCB}} \emptyset \tag{10}$$

The set of coupled ordinary differential equations obtained from those reactions, following mass-action kinetics (Chellaboina, Bhat, Haddad, & Bernstein, 2009), is shown below.

$$\frac{d[Heme](t)}{dt} = -k_1[Heme][HO1] + k_2[HO1\!:Heme] \tag{11}$$

$$\frac{d[HO1](t)}{dt} = -k_1[Heme][HO1] + k_2[HO1\!:Heme] + k_6[HO1\!:BV] \tag{12}$$

$$\begin{aligned}\frac{d[Fd_{red}](t)}{dt} &= -k_9[PcyA\!:BV][Fd_{red}] + k_{10}[Fd_{red}\!:PcyA\!:BV] \\ &\quad - k_3[Fd_{red}][HO1\!:Heme] + \ k_{13}[Fd_{oxi}] + k_4[Fd_{red}\!:HO1\!:Heme]\end{aligned} \tag{13}$$

$$\frac{d[Fd_{oxi}](t)}{dt} = k_{11}[Fd_{red}\!:PcyA\!:BV] - k_{13}[Fd_{red}] + k_5[Fd_{red}\!:HO1\!:Heme] \tag{14}$$

$$\frac{d[PcyA](t)}{dt} = k_8[PcyA\!:BV] + k_{12}[PcyA\!:PCB] - k_7[PcyA][BV] \tag{15}$$

$$\begin{aligned}&\frac{d[Heme\!:HO1](t)}{dt} \\ &\quad = k_1[Heme][HO1] - k_3[Fd_{red}][HO1\!:Heme] - k_2[HO1\!:Heme] \\ &\quad + k_4[Fd_{red}\!:HO1\!:Heme]\end{aligned} \tag{16}$$

$$\begin{aligned}&\frac{d[Fd_{red}\!:HO1\!:Heme](t)}{dt} \\ &\quad = k_3[Fd_{red}][HO1\!:Heme] - (k_4 + k_5)[Fd_{red}\!:HO1\!:Heme]\end{aligned} \tag{17}$$

$$\frac{d[HO1\!:BV](t)}{dt} = k_5[Fd_{red}\!:HO1\!:Heme] - k_6[HO1\!:BV] \tag{18}$$

$$\frac{d[BV](t)}{dt} = -k_7[Pcya][BV] + k_6[HO1: BV] + k_8[PcyA: BV] \tag{19}$$

$$\frac{d[PcyA: BV](t)}{dt} = -k_9[PCyA: BV][Fd_{red}] - k_8[PcyA: BV] + k_{10}[Fd_{red}: BV: PcyA] + k_7[BV][PcyA] \tag{20}$$

$$\frac{d[Fd_{red}: PcyA: BV](t)}{dt} = k_9[Fd_{red}][PcyA: BV] - (k_{10} + k_{11})\ [Fd_{red}: PcyA: BV] \tag{21}$$

$$\frac{d[PcyA: PCB](t)}{dt} = -k_{12}[PcyA: PCB] + k_{11}\ [Fd_{red}: PcyA: BV] \tag{22}$$

$$\frac{d[PCB](t)}{dt} = k_{12} * [PcyA: PCB] - k_{degPCB} * [PCB] \tag{23}$$

A. Fitting the Model to Experimental Data.

The model's unknown parameters were determined by a maximum likelihood approach fitted to the data shown in Muller *et al* (Müller, Engesser, Timmer, *et al.*, 2013). Units are defined in S.I. units with concentrations as the number of molecules for species ($\#molecules$, or c), and parameters as bimolecular rate constants in $\#molecules/s^{-1}$(or c/s^{-1}).

B. Sum-of-Squares and Parameter Estimation.

We assume that the system of ordinary differential equations (ODE) shown in above can be represented as a dynamical system given by an N-dimensional state variable $x(t) \in \mathbb{R}^N$, at time $t \in I = [t_0, t_f]$, which is the unique and differentiable solution for the initial value problem given by:

$$\dot{x}(t) = f(x(t), t, \theta) \quad x(t_0) = x_0 \tag{24}$$

As such, the ODE depends on certain parameters $\theta \in \mathbb{R}^{n_p}$ (Peifer & Timmer, 2007). Also, let Y_i denote the data of measurement $i = 1, \ldots, n$, where n represents the total amount of data. Moreover, the data Y_i satisfies $Y_i = g(t_i, \theta) + \sigma_i \epsilon_i$, for some function $g: \mathbb{R}^d \rightarrow \mathbb{R}^{obs}$, and $d \geq obs, \sigma_i > 0$ and ϵ_i are independent and standard Gaussian distributed random variables (Peifer & Timmer, 2007). The function $g(\cdot)$ is continuously differentiable. To estimate the parameters θ, given the initial conditions, utilizing the principle of maximum-likelihood to yield a cost function to be minimized gives us:

$$\mathcal{L}(\theta) = \sum_{i=1}^{n} \frac{\left(Y_i - g(x(t_i; \theta), \theta)\right)^2}{2\sigma_i^2} \tag{25}$$

We perform a direct minimization of $\mathcal{L}$ with respect to θ to obtain the parameters show in Appendix C, and used throughout the experiments described next.

C. Implementation of Experiments.

Our model was used to gain insight into the dependencies of this pathway and to further validate our experimental results. HO1 and PcyA were assumed to be at equimolar amounts and Fd at 1/10th of that molar concentration. Unless stated otherwise, the following initial conditions were used. If not listed, the initial concentrations were set to zero at $t = 0$.

$[Heme](0) = 100$

$[HO1](0) = 10$

$[Fd\,red, oxi](0) = 5$

$[PcyA](0) = 10$

Figure 7: Kinetic model results. By varying the initial Heme concentrations and the rate of renewal of Fdoxi to Fdred, we show the dependence on these parameters in the PCB pathway.

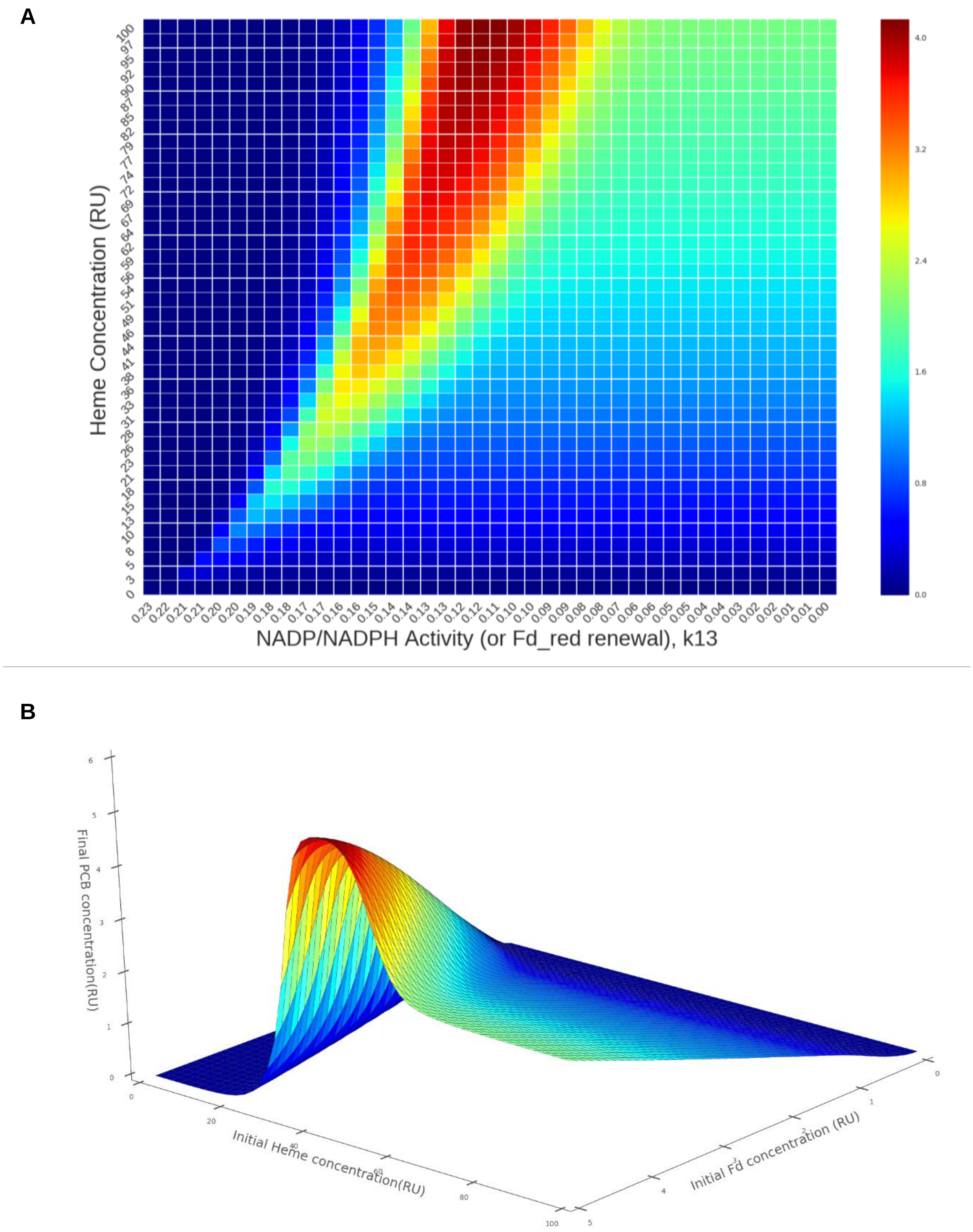

A
Heme Concentration (RU)
0 3 5 8 10 13 15 18 21 23 26 28 31 33 36 38 41 44 46 49 51 54 56 59 62 64 67 69 72 74 77 79 82 85 87 90 92 95 97 100
NADP/NADPH Activity (or Fd_red renewal), k13
0.23 0.22 0.21 0.21 0.20 0.20 0.19 0.18 0.18 0.17 0.17 0.16 0.16 0.15 0.14 0.14 0.13 0.13 0.12 0.12 0.11 0.10 0.10 0.09 0.09 0.08 0.08 0.07 0.06 0.06 0.05 0.05 0.04 0.04 0.03 0.02 0.02 0.01 0.01 0.00
4.0
3.2
2.4
1.6
0.8
0.0
B
Final PCB concentration(RU)
0 1 2 3 4 5 6
Initial Heme concentration(RU)
0 20 40 60 80 100
Initial Fd concentration (RU)
0 1 2 3 4 5

Figure 8: Kinetic model results. (A) We simulate the presence and absence of the FD: FNR complex, demonstrating more robust production of PCB with the 4 enzymes. (B) Decreasing sweep Figure 8: Kinetic model results, Continued: through the parameters k_3 and k_9, which control binding of HO1 and PcyA to Fd respectively. This graph shows that with decreasing species specificity, a decrease in PCB production is observed. (C) Varying initial concentrations of heme, demonstrating PCB dependence to Heme levels.

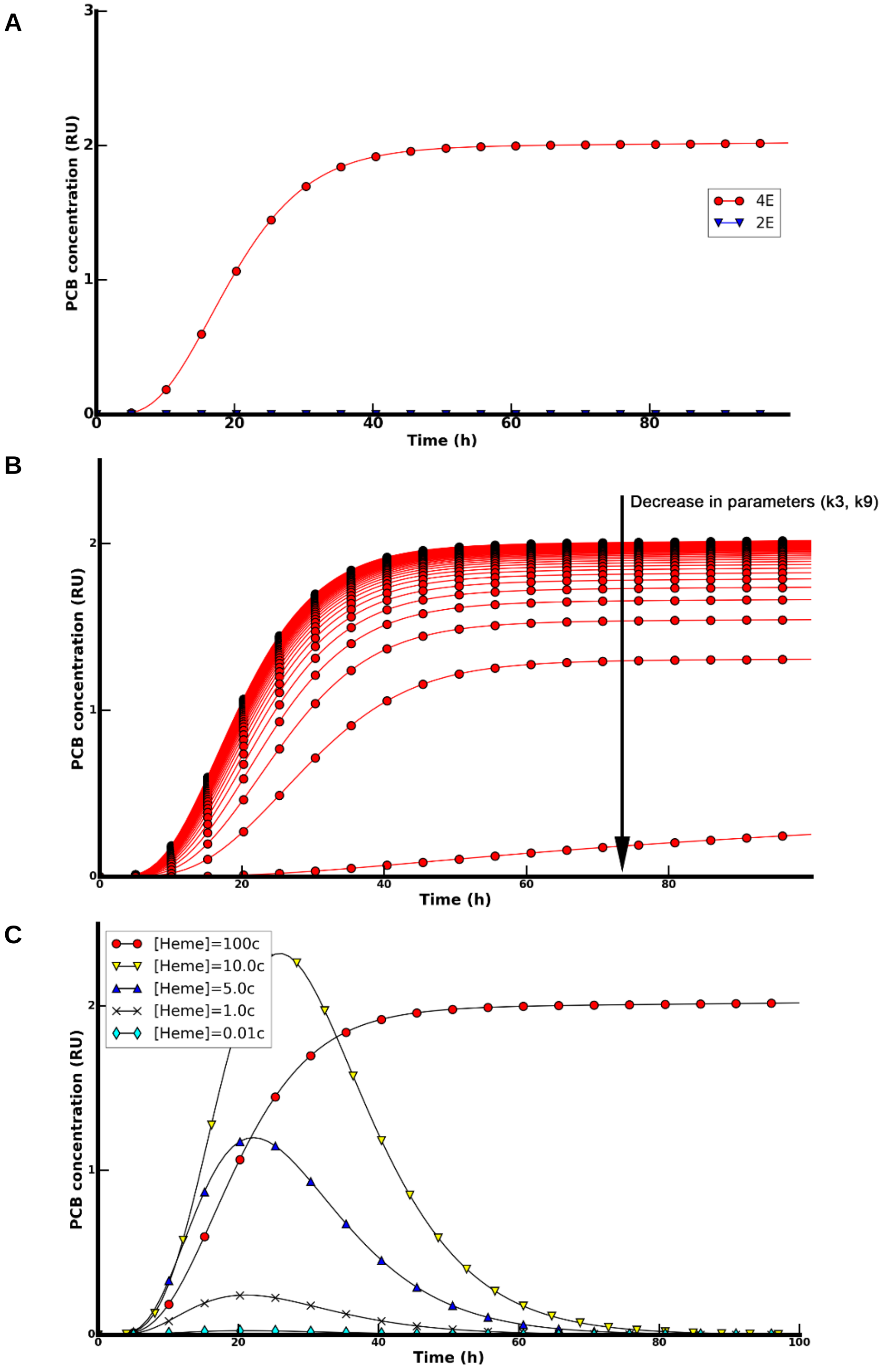
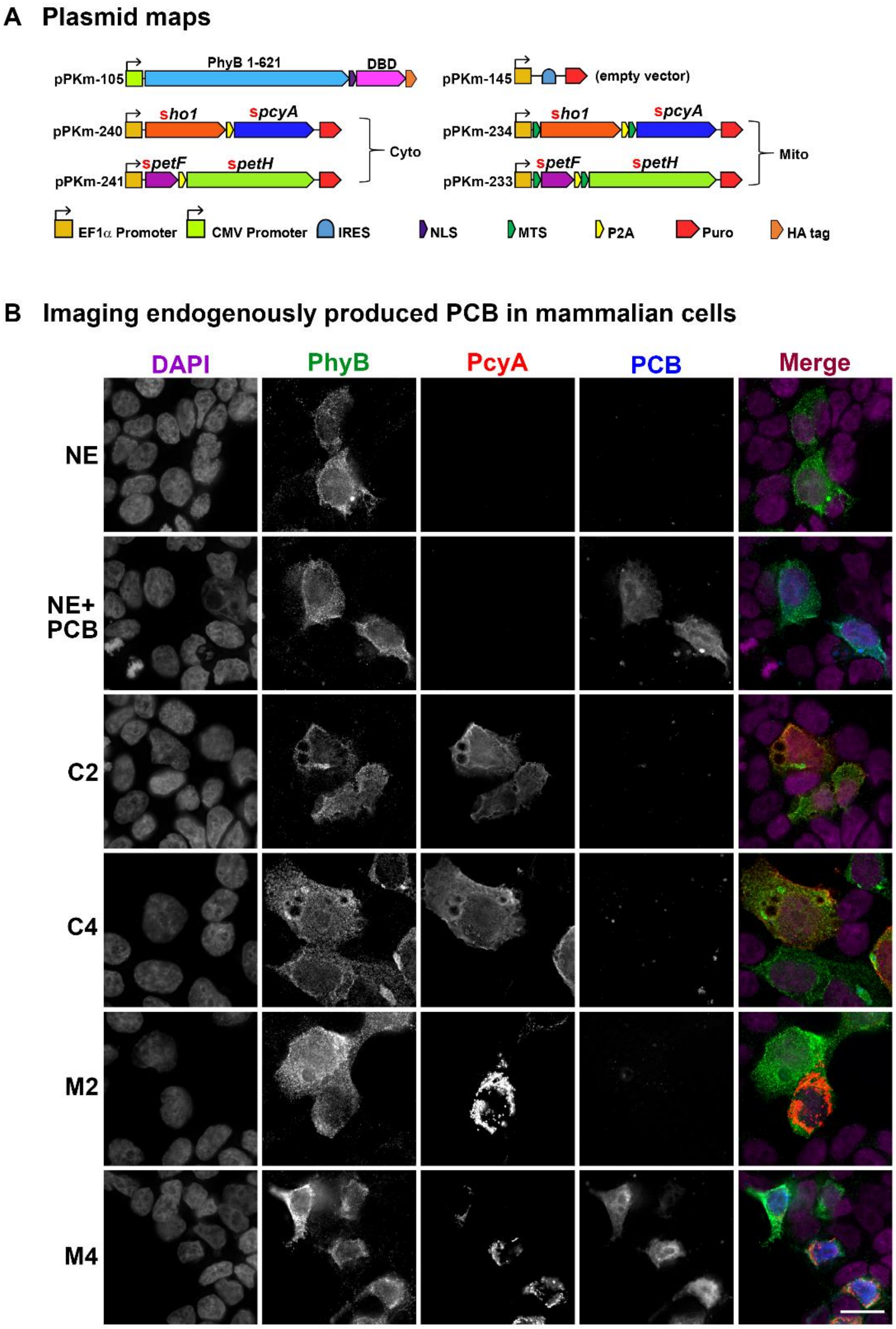

A
PCB concentration (RU)
3
2
1
0
0
20
40
60
80
Time (h)
4E
2E
B
Decrease in parameters (k3, k9)
PCB concentration (RU)
2
1
0
0
20
40
60
80
Time (h)
C
[Heme]=100c
[Heme]=10.0c
[Heme]=5.0c
[Heme]=1.0c
[Heme]=0.01c
PCB concentration (RU)
2
1
0
0
20
40
60
80
100
Time (h)

Figure 9: Imaging endogenously produced PCB in mammalian cells. HEK293 cells were transfected with PhyB alone (NE), PhyB+5µM PCB (NE+PCB), cytoplasmic *sho1+spcyA* (C2), cytoplasmic *sho1+spcyA+spetF+spetH* (C4), mitochondrial *sho1+spcyA* (M2), or mitochondrial *sho1+spcyA+spetF+spetH* (M4). DAPI DNA stain was imaged using the DAPI channel (purple). PhyB tagged with HA was imaged using anti-HA (green), PcyA tagged with FLAG was imaged using anti-FLAG (red). PCB was imaged using the Cy-5 channel (blue). All images were taken under the same exposure and contrast settings using a 60X (1.40NA) objective. IRES = Internal Ribosome Entry Site, NLS = Nuclear Localization Sequence, MTS = Mitochondria Targeting Sequence, P2A = 2A self-cleaving peptide, DBD = DNA Binding Domain, R/FR = Red light/Far-red light.



A Plasmid maps

B Imaging endogenously produced PCB in mammalian cells

A. Experiment 1: Fd and Heme dependence.

We determined experimentally the rate limiting factors are Fd, followed by Fd+FNR and finally heme. To model this experimental result, we performed a sweep over initial concentrations of Fd ([Fd](0)), heme ([Heme](0)), and rate of renewal of Fd by FNR (k_{13}). The result of those sweeps are shown in Figure 7A (Heme concentration vs. Fd renewal) and Figure 7B (Heme concentration vs. Fd concentration). The resulting graphs show the dependency of PCB production on those molecules, and how the initial condition of each affects the rate of production of PCB.

B. Experiment 2: 2E vs 4E.

Our experimental results show that PCB is only produced to high levels under the presence of Fd, PcyA, and HO1. To model this experimental result, we modified the following parameters to simulate the lack of compatible Fd, namely a “two enzyme” (2E) case, that limits the production of PCB versus the output of the pathway when all four enzymes (4E) are present. For the 2E case, we set $[Fd_{red,oxi}](0)$ to zero (Figure 8A).

C. Experiment 3: Species Specificity as Demonstrated by Different Binding Coefficients.

To demonstrate how the species specificity between Fd and HO1/PcyA plays a pivotal role in the amount of PCB produced, we performed a decreasing sweep through the parameters k_3 and k_9, which control binding of HO1 and PcyA to Fd respectively. The sweeps were started at the parameter’s value as described in Appendix C to 1e-3 $\boldsymbol{c/s^{-1}}$. The resulting graph is shown in Figure 8B.

D. Experiment 4: Variable Levels of Heme.

In this experiment, we performed a sweep over a range of Heme concentrations, from 100, 10, 5, 1 and 0.1 $\boldsymbol{c}$. This experiment, similar to Figure 7, shows the heme dependency of PCB production. The respective graph is shown in Figure 8C.

1.4 Results

1.4.1 Regulation of PCB Production in Mammalian Cells.

Given that previous studies have shown that PCB production can be limited by heme, Fd or FNR (Nicole Frankenberg & Lagarias, 2003; Gambetta & Lagarias, 2001), we tested limiting factors of PCB production in mammalian cells using combinations of these components in excess. Zinc-PAGE PhyB immunoprecipitation assays in Human Embryonic Kidney (HEK293) cells were used to test PCB production with metabolic enzymes from two species: *Synechococcus sp. PCC 7002* (SYNP2/sPcyA) or *Thermosynechococcus elongatus* (THEEB/tPcyA). We tested PCB production under two conditions, either mitochondrial-HO1+PcyA (M2) or mitochondrial-HO1+PcyA+Fd+FNR (M4), (Figure 2). When either species of HO1+PcyA enzymes were expressed, we detected low levels of PCB (Figure 2, M2). However, when all four enzymes HO1+PcyA+Fd+FNR (M4) were expressed, we observed a striking increase in PCB levels (Figure 2).

To exclude the possibility that this was specific to cyanobacterial enzymes, we also produced the plant chromophore PΦB*,* by replacing the cyanobacterial PcyA with a plant homolog *Arabidopsis* HY2. PcyA and HY2 showed the same Fd+FNR dependence (Figure 2, M2-asHY2 *versus* M4-asHY2). It is noteworthy that the Fd+FNR-dependent increase in PΦB production was still observed when plant HY2 was used along with cyanobacterial HO1/Fd/FNR. We chose SYNP2 Fd+FNR for recycling HY2 because SYNP2 Fd was more similar than THEEB Fd in amino acid sequence identity to *Arabidopsis* Fds and specifically the major ferredoxin that recycles HY2 in *Arabidopsis* (Appendix D) (Chiu, Chen, & Tu, 2010). However, PΦB production may be further increased by employing *Arabidopsis* Fd+FNR enzymes. It may be possible to predict compatibility of a transplanted ferredoxin-dependent pathway to the host cells Fd based on sequence similarity as shown in Appendix D. These findings show that excess Fd+FNR activity can increase PCB or PΦB production in mammalian cells (Figure 2).

Next, we delineated the limiting factors for the endogenous production of chromophores in mammalian cells. We decided to test PCB production in both the cytoplasm and mitochondria because the endogenous ferredoxin system of mammalian cells is localized in the mitochondria; therefore, we considered the cytoplasmic enzyme localization as a condition with negligible endogenous Fd+FNR activity. We show in Figure 3A that expression of cytoplasmic-PcyA+HO1 (C2) is not sufficient to produce significant levels of PCB (lane 3 vs. lane 2). When cytoplasmic-PcyA+HO1 was co-transfected along with cytoplasmic Fd+FNR (C4) higher, but statistically non-significant levels of PCB were detected (lane 3 vs. 4, $p > 0.05$). Similarly, when PcyA+HO1 were localized to the mitochondria (M2), very low levels of PCB were detected (lane 5). However, when PcyA+HO1 and Fd+FNR were all localized to the mitochondria (M4), PCB production was significantly increased when compared to PcyA+HO1 only (M2) (lane 5 vs. 6, $p<0.001$). These findings were corroborated by imaging PhyB-bound PCB using the Cy-5 channel (blue) (Figure 9). These results demonstrate that the Fd+FNR system is the primary limiting factor of the PCB production pathway in mammalian mitochondria, but it is not sufficient for high levels of PCB production when expressed in the cytoplasm.

Since heme is a metabolic precursor in the PCB production pathway, we systematically tested if it was limiting for PCB production in either the cytoplasm or in the mitochondria. We hypothesized that if heme was a limiting factor for PCB production in the cytoplasm, then the addition of excess heme would increase production. While a faint band was visible in C2+heme (Figure 3A lane 9), it was indistinguishable from cells transfected with PhyB and no enzymes and given excess heme (Figure 3A lane 8). However, excess heme significantly increased levels of PCB production in the C4 condition (lanes 4 and 10, p<0.01). In addition, we found that Fd+FNR was limiting when comparing C2+heme to C4+heme (lanes 9 and 10, p<0.01). This demonstrates that heme is the limiting factor for PCB production when an excess of Fd+FNR is present in the cytoplasm. Importantly, PCB production was not influenced by excess heme when enzymes were localized to the mitochondria (M4-heme and M4+heme, lanes 6 and 12). This confirms that Fd+FNR is primarily limiting in both the cytoplasm and the mitochondria and that heme is secondarily limiting only in the cytoplasm.

To further investigate the PCB production dependence on Fd, we transfected cells with two, three or all four enzymes in the pathway: PcyA-HO1 (M2), PcyA+HO1+Fd (M3), or PcyA+HO1+Fd+FNR (M4), along with PhyB for all conditions (Figure 3B). We show in Figure 3B that the addition of Fd to PcyA+HO1 (M3) significantly increased PCB production compared to PcyA+HO1 alone (M2) (p<0.05). Importantly, Fd+FNR (M4) produces significantly more PCB than adding Fd alone (p<0.01), demonstrating that for maximum PCB production both Fd and FNR are required.

While we considered testing the overexpression of the host cell's Fd+FNR, there are noteworthy advantages to using orthogonal Fd+FNR matching the species of the transplanted metabolic pathway. The mammalian Fd+FNR may be able to reduce BV bound to PcyA but only at a fraction of the rate of the cyanobacterial Fd+FNR. The required overexpression needed for the host cell's system to perform at the same production rate would therefore more likely disturb the cell's metabolism. Using an orthogonal system would be more efficient and would also less likely interact with the host cell's metabolic proteins. Matching the orthogonal enzyme species thus allows for minimal perturbation of the normal host cell physiology and at the same time maximize production rates.

1.4.2 Effects of Enzyme Stoichiometry on PCB Production Levels.

Okada *et al.* (Okada, 2009) demonstrated that Fd forms stable complexes with both HO1 and PcyA. Therefore, we hypothesized that PCB production may be further optimized through enzyme stoichiometry. We transfected separate PcyA+HO1 and Fd+FNR plasmids at different ratios and observed that PCB production was highly dependent on the ratio between PcyA+HO1 and Fd+FNR (Figure 4A). Considering this, to serve as a quantitative guide for optimizing PCB production, we developed computational models of this pathway using coupled ordinary differential equations. We tested the enzyme stoichiometry using a functional PhyB/PIF33 luciferase gene expression system adapted from Shimizu Sato *et al.* (Shimizu-Sato *et al.*, 2002) (Figure 4B). This was done by transfecting different ratios of the PcyA+HO1 and Fd+FNR plasmids and illuminating the cells with red light for 24 hours, followed by a luciferase assay to compare gene induction levels. We found that gene activation levels were also highly dependent on enzyme stoichiometry, with only the 17:1 PcyA+HO1:Fd+FNR showing any measurable response to light (Figure 4C and 4D, $p<0.01$). This demonstrates how chromophore levels influence the performance of PhyB optogenetic systems.

1.5 Summary

Phytochromes are promising candidates for improving light delivery for imaging and optical control of biology. We have shown that the Fd+FNR system is the rate-limiting factor for the production of the chromophores PCB and PΦB in the mitochondria of mammalian cells, and is limited by the Fd+FNR system followed by heme in the cytoplasm. The ability to produce PCB and PΦB with PcyA and HY2, respectively, suggests that matching reduction systems that efficiently supply electrons to a metabolic pathway can also enhance the production of other bilins and other classes of molecules. This finding creates new opportunities for engineering synthetic systems to produce these chromophores, along with many other molecules. This has potential industrial applications in decreasing costs of crop production, producing plant molecules in microbes, or delivering therapeutic molecules via genetically encoded pathways.

Genetically encoding endogenous production of chromophores like PCB also enables the use of several existing and compatible optogenetic tools to regulate cell signaling (Levskaya *et al.*, 2009; Toettcher, Gong, Lim, & Weiner, 2011), cell migration (Levskaya *et al.*, 2009), or protein localization (Levskaya *et al.*, 2009) without the addition of exogenous chemicals. This makes possible the use of PhyB when constant levels of PCB are required, facilitating potential *in vivo* applications, or when the addition of PCB to samples is not practical (such as when samples are in a sealed container or for long illumination times). This study achieves the long-sought goals in optogenetics of enabling high-level production of the chromophores PCB and PΦB in mammalian cells and demonstrates a more general method for efficiently producing molecules from one species in another.

In the following chapter, we show how genetically encoding mammalian cells to produce these chromophores enabled us to develop a robust NIR gene switch that is fully genetically encoded, removing these barriers for *in vivo* applications. In addition, to demonstrate the utility of increased chromophore production for optogenetic applications, we chose a PhyB-based optogenetic system, which utilizes PCB and has been used to control a wide array of biological processes. Since the light sensitivity of PhyB is proportional to the amount of chromophore in the cell, to apply PhyB optogenetic tools in transgenic animal models, it will be essential to genetically encode a high level of chromophore production.

Chapter 1, in part, is a reprint of the material as it appears in ACS Synthetic Biology, 2018. “Biosynthesis of Orthogonal Molecules Using Ferredoxin and Ferredoxin-NADP+ Reductase Systems Enables Genetically Encoded PhyB Optogenetics" Phillip Kyriakakis, Marianne Catanho, Nicole Hoffner, Walter Thavarajah, Vincent Jian-Yu Hu, Syh-Shiuan Chao, Athena Hsu, Vivian Pham, Ladan Naghavian, Lara E. Dozier, Gentry Patrick and Todd P. Coleman. DOI: 10.1021/acssynbio.7b00413.

CHAPTER 2 OPTIMIZED OPTOGENETIC SWITCH

Reproduced in part with permission from:

Kyriakakis, P., Catanho, M., Hoffner, N., Thavarajah, W., Jian-Yu, V., Chao, S.-S., … Coleman, T. (2018). Biosynthesis of Orthogonal Molecules Using Ferredoxin and Ferredoxin-NADP+ Reductase Systems Enables Genetically Encoded PhyB Optogenetics. ACS Synthetic Biology. doi:10.1021/acssynbio.7b00413

2.1 Abstract

Using light to tune cellular activity of genes and proteins represents a very attractive methodology for producing various temporal interventions in living systems. Phytochromes are red and far-red light photochromic biliprotein photoreceptors known to bind directly to the transcription factor PIF3. In this heterodimer form PhyB/PIF33 has been utilized in several gene expression systems in which light is used to induce conformational changes and direct regulation of gene expression and protein production in plant and animal cells. Expressing the FD-FNR system from bacteria and plants, along with phycocyanobilin:ferredoxin oxidoreductase (PcyA) and heme oxygenase-1 HO1, enabled the production of these chromophores in mammalian cells and the development of a fully endogenous PhyB/PIF33 red-light activated gene switch.

We further characterized the fully endogenous PhyB/PIF33 optogenetic gene system in several mammalian cell lines. We were able to control genes with low background, high dynamic range, and orders of magnitude less light than any other optogenetic system. More importantly, we found that the light-switchable gene system

remains active for several hours upon illumination, even with a short light pulse and requires very small amounts of light for maximal activation. By combining the ability of red light to penetrate deeply into tissue with the low light requirements for maximal activation of the PhyB/PIF33 optogenetic gene switch, our methodology enables unprecedented control of genes through light both *in vitro* and *in vivo*. This system has great potential in animal studies and light-modulated gene therapies, and to enable new areas of synthetic biology.

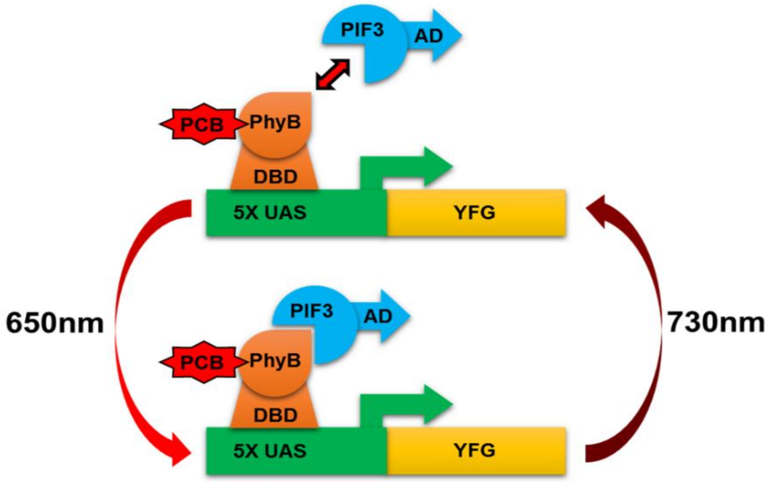


Figure 10: Schematic of the PhyB/PIF33 light switch. PhyB is fused to a DNA Binding Domain (DBD) and bound to a light sensitive chromophore (PCB). The PhyB-DBD fusion remains bound to the UAS promoter. PIF3 is fused to an Activation Domain (AD). Upon absorption of a red photon (650nm), PhyB changes conformation and recruits PIF3 to the promoter region. The AD fused to PIF3 then activates the gene downstream of the promoter. Upon absorption of a far-red photon (740nm), PhyB changes conformation that leads to PIF3 unbinding, removing the AD from the promoter, shutting the downstream gene off.

2.2 Introduction

Optical control of biology holds great promise as a tool for studying gene function, developmental biology, gene therapies and tissue engineering. The exquisite temporal and spatial precision achieved through optics has been used to develop an assortment of tools to control biological functions such as gene expression (Folcher *et al.*, 2014; Kaberniuk *et al.*, 2016; Müller, Engesser, Metzger, *et al.*, 2013; Pathak *et al.*, 2014; Shimizu-Sato *et al.*, 2002; X. Wang *et al.*, 2012), neural activity (Boyden *et al.*, 2005; John Y Lin *et al.*, 2013), cell signaling (Levskaya *et al.*, 2009), secretion (D. Chen *et al.*, 2013), peroxisomal trafficking (Spiltoir *et al.*, 2016), and protein activity (X. X. Zhou *et al.*, 2012). However, most of these existing systems have significant limitations. Particularly, they are either not very robust (Folcher *et al.*, 2014; Kaberniuk *et al.*, 2016; Müller, Zurbriggen, & Weber, 2014; Pathak *et al.*, 2014), require sufficient presence of light-absorbing chromophores (Kawano, Suzuki, Furuya, & Sato, 2015; Konermann *et al.*, 2013; Pathak *et al.*, 2014), interfere with the cells intracellular signaling pathways (Folcher *et al.*, 2014), or the wavelength of light used penetrates tissue poorly (Kawano *et al.*, 2015; Konermann *et al.*, 2013).

Phytochromes-based optogenetic systems have been shown to be ideal candidates to address those shortcomings, having evolved to require minimal light for activation and to absorb light in the NIR window. These are inherent properties of phytochromes and many proteins with a bilin chromophore because: i) the chromophores are very sensitive to light (high absorbance/extinction coefficient) and ii) the chromophores bound to the phytochrome can have a long-lived activation state, ranging from tens of minutes to hours (J Y Lin, Lin, Steinbach, & Tsien, 2009; Mattis *et al.*, 2011; Smith *et al.*, 2016; Yizhar, Fenno, Davidson, Mogri, & Deisseroth, 2011). However, *Arabidopsis* *thaliana*'s Phytochrome B (PhyB), the most characterized phytochrome, has these optical characteristics and has been shown to be very robust compared to other switches (Folcher *et al.*, 2014; Kaberniuk *et al.*, 2016; Müller, Engesser, Metzger, *et al.*, 2013; Pathak *et al.*, 2014). It still requires external addition of a chromophore, limiting them to *in vitro* applications.

PhyB is known to interact with the transcription factor phytochrome interacting factor 3 (PIF3) in a light-dependent way (Shimizu-Sato *et al.*, 2002; Tyszkiewicz & Muir, 2008): it covalently binds to the chromophore PCB, which activates a light-induced conformational change and enables its interaction with the phytochrome interacting factor PIF3(Li *et al.*, 2011; Milias-Argeitis *et al.*, 2011). Together, PhyB/PIF33 compose a light-dependent two-hybrid system, associating and dissociating in response to red (650nm) and far-red (730nm). The basic principle of this interaction is shown in Figure 10. Originally create to explore protein-protein interactions, this photoreversible gene system based on the PhyB/PIF33 interaction has been used to induce gene transcription (Hughes *et al.*, 2012; Li *et al.*, 2011), control the activity of proteins at the post-translational level (Tyszkiewicz & Muir, 2008), regulate intracellular pathways (Zhang & Cui, 2015), nuclear translocation of synthetic transcription factors (Beyer *et al.*, 2015), angiogenesis in chicken embryos (Müller, Engesser, Metzger, *et al.*, 2013), among others.

As shown in Chapter 1, we are now able to genetically encode mammalian cells to produce the chromophores needed for phytochrome systems. Building upon this endogenous availability of chromophores in mammalian cells, we have developed a fully genetically encoded and robust NIR gene switch based on the light-responsive PhyB/PIF33 module (M. Chen *et al.*, 2005; Elich & Chory, 1997; Kunkel, Speth, Büche, & Schäfer, 1995; Li *et al.*, 2011; Remberg, Ruddat, Braslavsky, Gärtner, & Schaffner, 1998), removing the barriers for *in vivo* applications. We demonstrate the utility of coupling matching reduction systems, such as Fd+FNR, with the PCB metabolic pathway by developing a phytochrome based tool to control biological processes with NIR light in mammalian cells. In addition, we show that increasing the production of PCB in mammalian cells enables the development of a robust genetically encoded Red-light Activated Gene Switch, compatible with PhyB based optogenetic systems.

To our knowledge, the genetic tool presented in this work is the most light sensitive optogenetic system to date: the peak intensity required for maximal activation is at most 2nWatts/mm^2. For comparison, it requires 500,000X-2,500,000X less light than the peak activation for stimulating neurons with ChR2 (Yizhar *et al.*, 2011) and is 50-100X more sensitive than other phytochrome-based gene switches in yeast and mammalian cells (Müller *et al.*, 2014; Shimizu-Sato *et al.*, 2002). By combining the ability of red light to penetrate deeply into tissue with the low light requirements for maximal activation of our system, it will now be possible to use light to control genes deeper into tissues than ever before. Our red-light activated gene switch has great potential in animal studies and light-modulated gene therapies.

2.3 Methodology

2.3.1 Cell Culture, Transfection, Light Induction and Reporter Gene Assays

Human Embryonic Kidney 293 cells (HEK293, ATCC CRL-1573) were cultivated in Dulbecco's Modified Eagle Medium (DMEM, Gibco, 11965-092) supplemented with 10% fetal bovine serum (FBS, Omega Scientific, FB-02) and 100 U/ml of penicillin and 0.1 mg/ml of streptomycin (Gibco, 11548876). All cells were cultured under 5% CO2 at 37°C. Cells were seeded at 100,000 HEK293 cells per well in 24-well plates, 24 hours before transfection. Transfection of plasmids was achieved through lipofection following the manufacturer’s instructions and protocol (Lipofectamine 2000, ThermoFisher, 11668019). For each transfection reaction, a total of 0.5µg of plasmid DNA was combined with specific plasmid ratios for each experiment as detailed in Appendix A and B.

A construct with Renilla luciferase reporter plasmid DNA was included as an internal transfection control in all transfections. The culture medium was replaced with fresh medium 24 hours after transfection and the plates were placed inside black boxes (Hammond Manufacturing Company, 1591ESBK) for the remainder of the experimental procedure. For conditions where external PCB is added, 15µM of PCB (Frontier Scientific, P14137) from a 20 mM stock dissolved in DMSO (Santa Cruz Biotechnology, sc-202581) was supplemented in fresh medium 24 hours after transfection (Figure 11A).

Light induction was programmed to start 12 hours after medium replacement. Each black box was equipped with a circuit consisting of six red LEDs (660nm, Thorlabs, M660L3), except for the dark boxes and far-red boxes which had no LEDs or a single far-red LED (735nm, Thorlabs, M735L2), respectively. In addition, each black box circuit was designed to allow for fine adjustment of light intensity (circuitry is shown Figure 5 and 6), from 0.0008 to 200 µmol/m^2/s. Light intensity was measured in µW at the cell level, converted to µmol/m^2/s (light sensor area = 63.6mm^2), and adjusted for each experiment design using Sper Scientific Direct’s Laser Power Meter (SSD, 8400). Detailed information on wavelengths, illumination intensity, and duration used for each experimental procedure and data shown are detailed in Appendix B. Pulse duration and total illumination times were electronically controlled via a LabVIEW computer driving an Arduino microprocessor and custom-made circuits (see Section 1.3.6).

2.3.2 Luciferase Activity Assay

Luciferase assays were carried out using the Dual-Luciferase Assay system (Promega, PRE1960), and following the manufacturer's protocol. Cells were lysed immediately after removing from the incubator using the manufacturer's instructions. Firefly and Renilla Luciferase activities were measured from cell lysates using the luminometer module of the Infinite 200 PRO multimode reader (Tecan). Results of luciferase activity assays are expressed as a ratio of firefly luciferase (Fluc) activity to Renilla luciferase (Rluc) activity.

2.4 Results

2.4.1 PCB production in mammalian cells enables genetically encoded PhyB/PIF33 based optogenetic systems

After identifying the requirements for high levels of endogenous PCB production, we sought to encode all four biosynthetic enzymes on a single plasmid. Our original four enzyme plasmid (pPKm-245) contained all PCB biosynthetic enzymes separated by P2A sequences to achieve a 1:1:1:1 expression level of each enzyme (Szymczak *et al.*, 2004). However, the results in Figures 4A-4D suggested that PCB production could be further optimized by modifying the plasmid's expression stoichiometry. To this end, we replaced one of the P2A sequences with an Internal Ribosomal Entry Site (IRES), which typically gives one order of magnitude lower expression to the gene following the IRES sequence (Bochkov & Palmenberg, 2006; Licursi, Christian, Pongnopparat, & Hirasawa, 2011; Mizuguchi, Xu, Ishii-Watabe, Uchida, & Hayakawa, 2000). The plasmid pPKm-244 was generated by placing an IRES between *pcyA* and *Fd*, leading to higher PcyA-HO1 levels and lower Fd+FNR levels (Figure 11A). We also constructed a plasmid, pPKm-248, containing *HO1*, *Fd*, and *FNR* all placed after the IRES sequence. This plasmid results in minimized heme oxygenase and Fd+FNR activity while keeping higher levels of PcyA (Figure 11A). Using the experiment timeline in Figure 12, we found that lowering HO1 and Fd+FNR levels with the pPKm-248 plasmid produced 1.8-fold ($p<0.05$) and 2.2-fold ($p<0.01$) higher gene activation levels than pPKm-244 and pPKm-245 respectively (Figure 11B). In addition to producing more PCB, lower expression of HO1, Fd and FNR should provide maximal PCB levels with minimal interference in the host cells metabolism.

Adapted from Shimizu-Sato *et al.* and Müller *et al.* (Müller, Engesser, Metzger, *et al.*, 2013; Shimizu-Sato *et al.*, 2002), we constructed several versions of the PHYB/PIF33 gene switch to optimize gene induction in mammalian cells (Figure 12). In the presence of exogenous PCB, we compared red light gene activation using two strong synthetic activation domains, MTAD and VPR (Chavez *et al.*, 2015; Tachikawa *et al.*, 2004). The VPR domain activated luciferase 2.6 fold more than MTAD (Figure 12C). To find the optimal configuration for the activation domain, we also compared C-terminal and N-terminal fusions of VPR to PIF3. VPR on the C-terminus produced 2.4 fold higher luciferase activation compared to the N-terminal fusion (Figure 12C).

Next, we compared the leakiness of promoter constructs containing CMV minimal promoter with 13X TET-UAS from Müller *et al.* to Fluc and CMV minimal promoters with 5X Gal4-UAS and to cells transfected with Renilla alone (Müller *et al.*, 2014). The 13X TET-UAS gave a signal 172.6 fold higher than the Renilla only control, and both Fluc and CMV Gal4-UAS constructs had similar levels of leakiness with 16.0 and 14.2 fold activation, respectively, above the Renilla only control (Figure 13A). As an additional test, we also measured transcription levels of the entire gene switch in the off state, under far-red light. The Fluc and CMV minimal promoters gave a luciferase signal 6.2 fold and 31.4 fold higher than the Renilla alone, respectively (Figure 13A).

The decrease in leakiness with the entire switch under far-red light means that for phytochrome-based gene switches, there are two useful ways to define leakiness: (1) the basal transcription rate when cells contain with reporter and control plasmids alone (UAS-Luciferase and Renilla, for our experiments) and (2) the basal transcription rate when cells contain the complete switch and illuminated with far-red light. We also tested maximal activation levels of the Gal4 UAS reporters Fluc and CMV by using Gal4-VP16. The CMV minimal promoter had 3.4 fold higher the activation levels than the Fluc promoter (Figure 13B). Together these promoter constructs allow for modularity for higher activation levels at the expense of leakiness. Depending on the application where low leakiness is essential, Fluc can be used or where high activation levels are required, the CMV minimal promoter or other UAS constructs such as the 13X-TET-UAS can be employed.

Figure 11: Stoichiometry of PCB production constructs. (A) Three construct designs consisting of all four biosynthetic enzymes on a single plasmid and a single plasmid for PIF3 and PhyB. (B) Testing gene activation comparing single plasmid biosynthetic plasmids (n=7). ho1 = heme oxygenase, pcyA = Phycocyanobilin:ferredoxin oxidoreductase, petF = ferredoxin, petH = ferredoxin:oxidoreductase/FNR, MTS = Mitochondrial Targeting Sequence, P2A = 2A self-cleaving peptide, NLS = Nuclear Localization Sequence, IRES = Internal Ribosome Entry Site, AD = Activation Domain, DBD = DNA Binding Domain, R/FR = Red light/Far-red light. Error bars = Standard Deviation, (*) = $p<0.05$, (**) = $p<0.01$. Statistics were calculated using one-way ANOVA with Bonferroni post-test using GraphPad Prism 5.01. n = individual experiments.

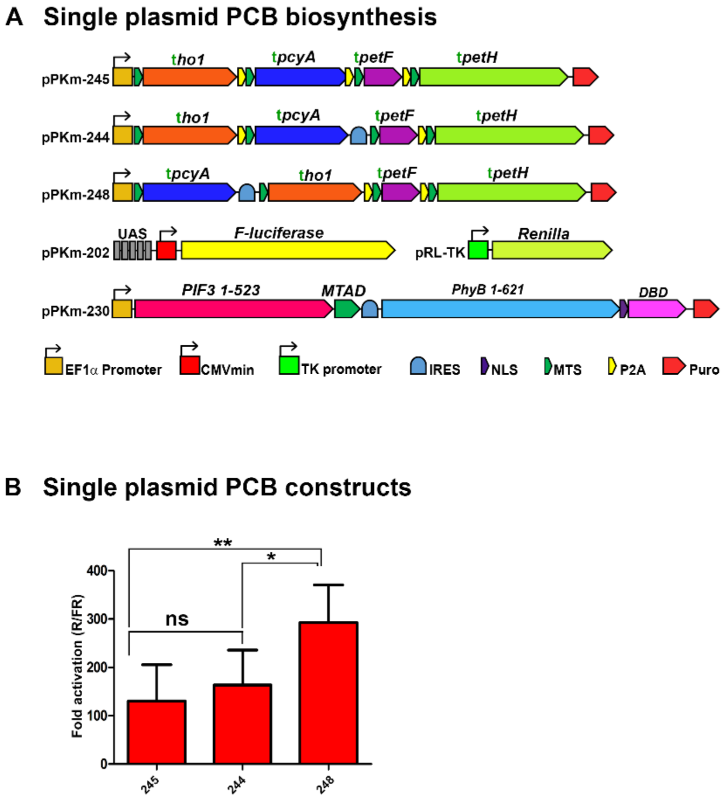

A Single plasmid PCB biosynthesis
pPKm-245
tho1
tpcyA
tpetF
tpetH
pPKm-244
tho1
tpcyA
tpetF
tpetH
pPKm-248
tpcyA
tho1
tpetF
tpetH
pPKm-202
UAS
F-luciferase
pRL-TK
Renilla
pPKm-230
PIF3 1-523
MTAD
PhyB 1-621
DBD
EF1α Promoter
CMVmin
TK promoter
IRES
NLS
MTS
P2A
Puro
B Single plasmid PCB constructs
Fold activation (R/FR)
0
100
200
300
400
**
*
ns
245
244
248

Figure 12: Activation domain optimization. (A) Timeline for experiments where HEK293 cells were transfected and illuminated for 24 hours. (B) Plasmid maps for constructs with MTAD and VPR activation domains fused to the C-terminal or N-terminal of PIF3 (C) Comparison of MTAD and VPR fusions with PIF3 effects of luciferase gene activation. Fold gene expression was calculated comparing cells incubated in red light to cells incubated in far-red light, after normalizing to a Renilla control (n=3). DBD = DNA Binding Domain, AD = Activation Domain, MTAD = Minimal Trans-Activation Domain, VPR = VP64+P65+RTA, R/FR = Red light/Far-red light. (Error bars = s.d. (***) = p<0.001, Statistics were calculated using one-way ANOVA with Bonferroni post-test using GraphPad Prism 5.01).

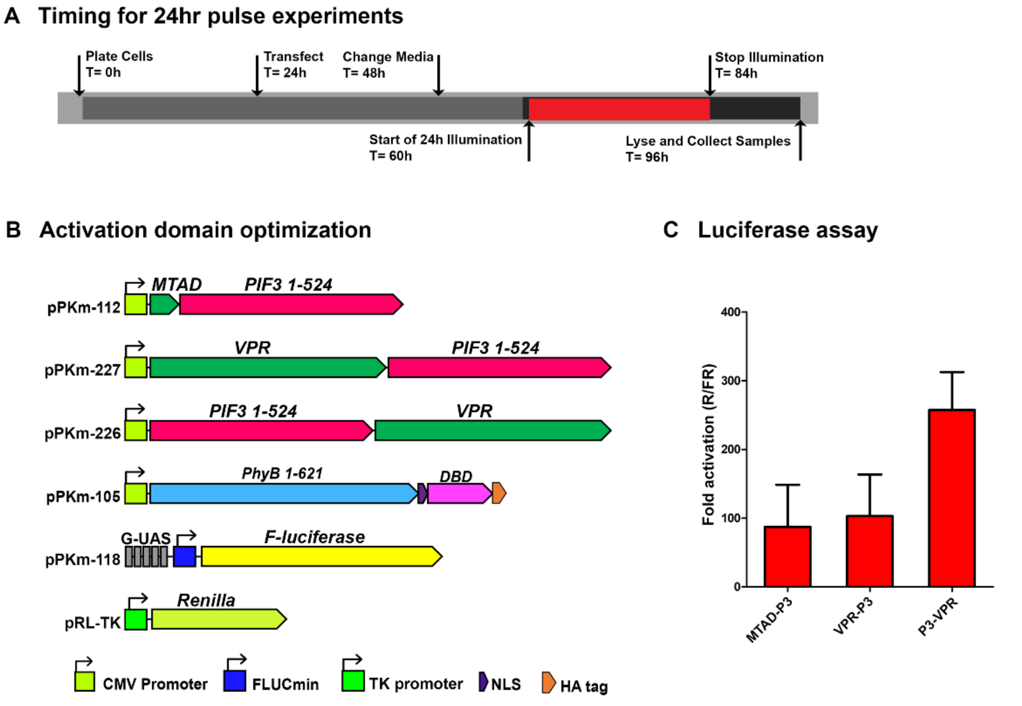
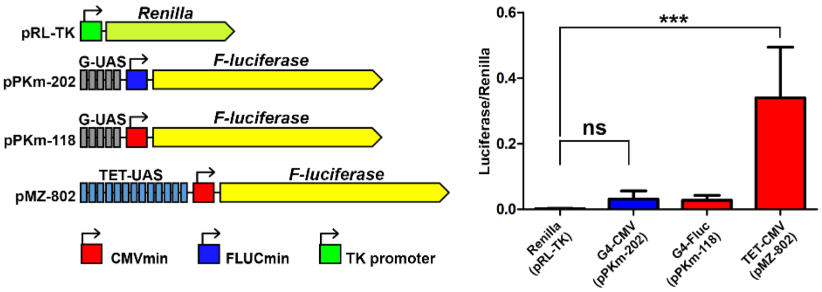

A Timing for 24hr pulse experiments
Plate Cells T= 0h
Transfect T= 24h
Change Media T= 48h
Stop Illumination T= 84h
Start of 24h Illumination T= 60h
Lyse and Collect Samples T= 96h
B Activation domain optimization
pPKm-112
MTAD
PIF3 1-524
pPKm-227
VPR
PIF3 1-524
pPKm-226
PIF3 1-524
VPR
pPKm-105
PhyB 1-621
DBD
pPKm-118
G-UAS
F-luciferase
pRL-TK
Renilla
CMV Promoter
FLUCmin
TK promoter
NLS
HA tag
C Luciferase assay
Fold activation (R/FR)
400
300
200
100
0
MTAD-P3
VPR-P3
P3-VPR

Figure 13: Comparing reporter constructs. (A) Leakiness analysis comparing different reporter vectors. HEK293 cells were transfected using the reporter vector along with Renilla (pRL-TK) alone or with Renilla+filler DNA (pRL-TK +pPKm-102) plasmids. Leaky luciferase values were compared to Renilla alone (n=5) (B) Activation level comparison of Gal4 UAS and TET UAS reporters. HEK293 cells transfected with pPKm-202 or pMZ-802 along with pPKm-292 or pPKm-293 respectively (n=3). G-UAS = Gal4 UAS, TET-UAS = TET UAS, GDBD = Gal4 DNA Binding Domain, TETDBD = TET DNA Binding Domain. (Error bars = s.d. (***) = p<0.001, Statistics were calculated using one-way ANOVA with Bonferroni post-test using GraphPad Prism 5.01).

A Promoter leakiness assessment



B Promoter activation level comparison

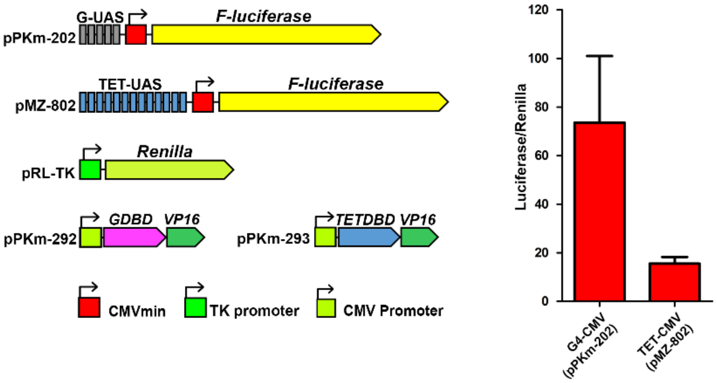
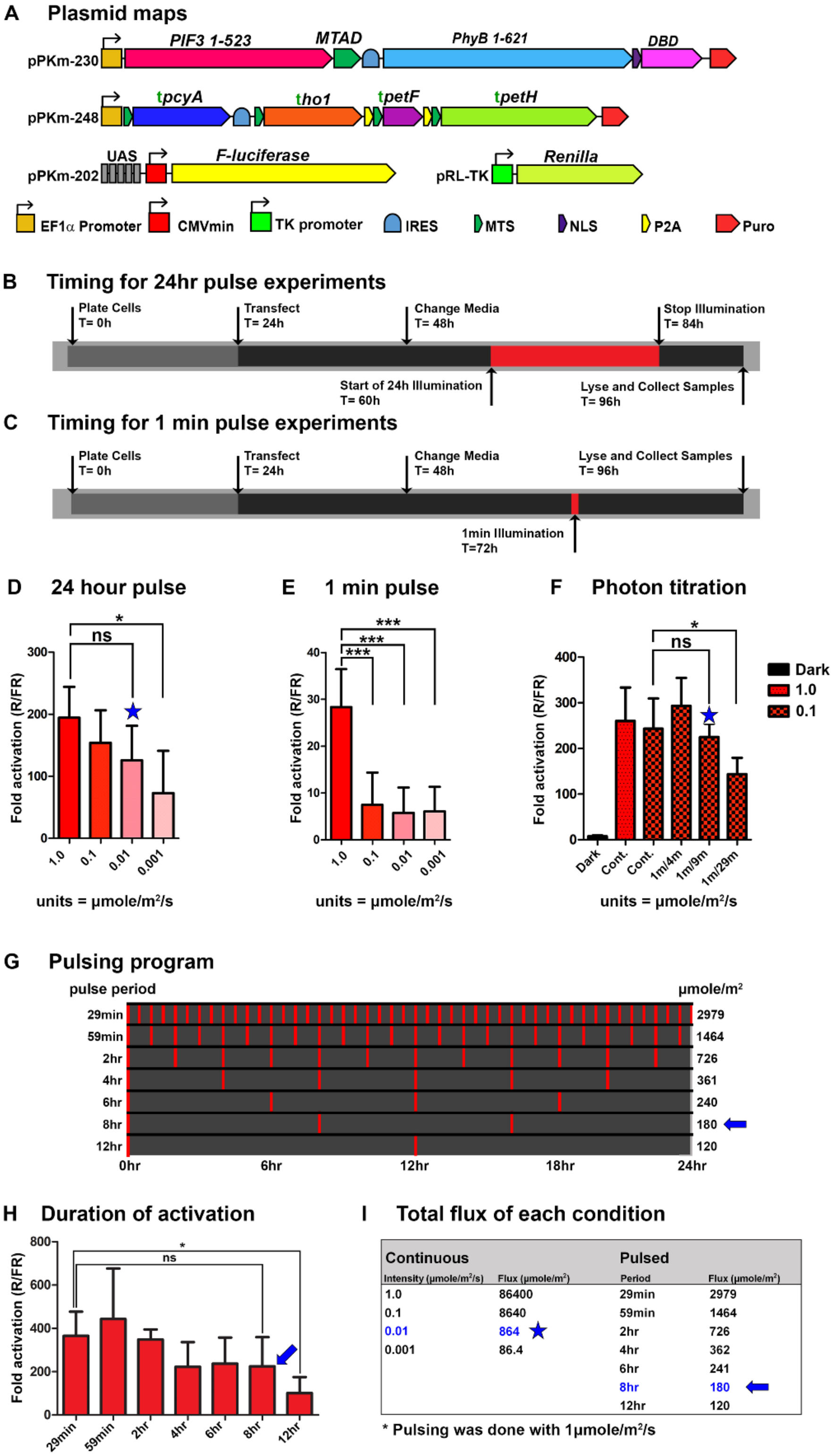


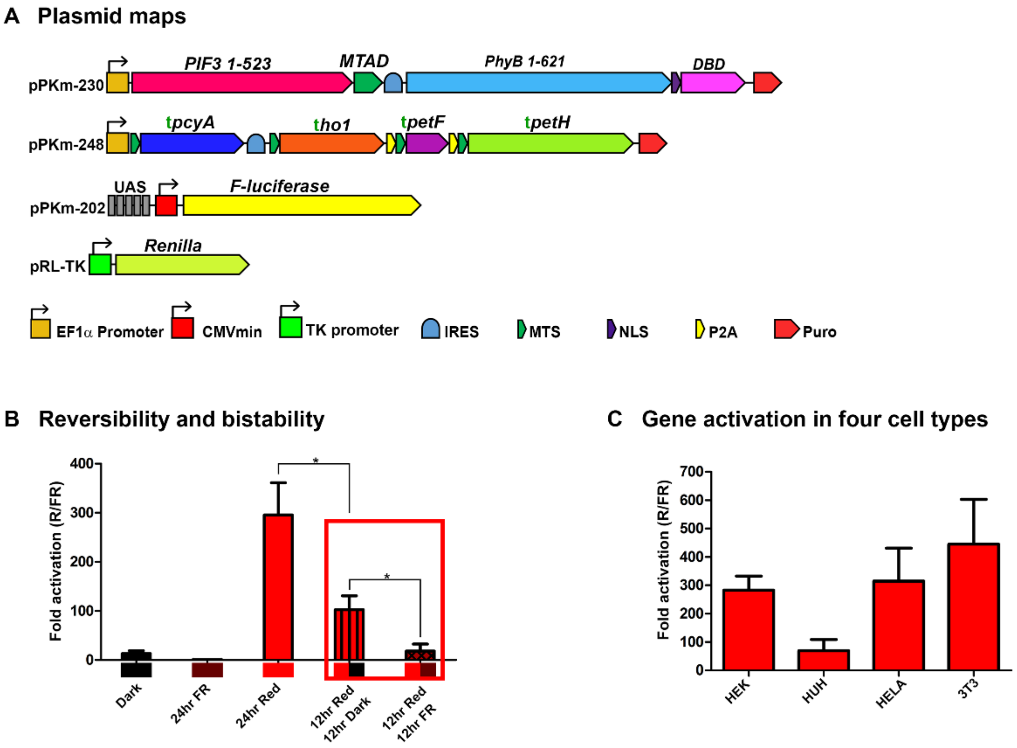
Figure 14: Light sensitivity of the genetically encoded PhyB-PIF3 switch. (A) Plasmids optimized for an endogenous PhyB-PIF3 light switchable promoter. (B) Pulsing program for 24-hour illumination experiments. (C) Pulsing program for one-minute illumination experiments. (D) Gene response to a 24-hour pulse with several light intensities (n = 4). (E) Gene response to a one-minute pulse with several light intensities (n = 4). (F) Gene activation responses using 1µmole/m^2/sec or 0.1µmole/m^2/sec of continuous light compared with using 0.1umole light at different pulse intervals for 24 hours. (n=3) – The blue star indicates minimal light doses for 24-hour illuminations. (G) Pulsing program for testing the duration of activation. Pulsing was done as in 4B. (H) Gene response to pulsing at increasing intervals. Cells were pulsed for one minute using 1µmole/m^2/sec 660nm light, followed by darkness for the indicated times for a total of 24 hours (n=5). Cont. = continuous illumination, 1m/4m = one-minute red light, 4 minutes darkness, 1m/9m = one minute red light, 9 minutes darkness, 1m/29m = one minute red light, 29 minutes darkness. ho1 = heme oxygenase, pcyA = Phycocyanobilin:ferredoxin oxidoreductase, petF = ferredoxin, petH = ferredoxin:oxidoreductase/FNR IRES = Internal Ribosome Entry Site, MTS = Mitochondrial Targeting Sequence, NLS = Nuclear Localization Sequence, P2A = 2A self-cleaving peptide, AD = Activation Domain, DBD = DNA Binding Domain, R/FR = Red light/Far-red light. (Error bars = Standard Deviation, (*) = $p<0.05$, (***) = $p<0.001$. Statistics were calculated using one-way ANOVA with Bonferroni post-test using GraphPad Prism 5.01). n = individual experiments.



Continuous		Pulsed	
Intensity (μmole/m²/s)	Flux (μmole/m²)	Period	Flux (μmole/m²)
1.0	86400	29min	2979
0.1	8640	59min	1464
0.01	864 ★	2hr	726
0.001	86.4	4hr	362
		6hr	241
		8hr	180 ⬅
		12hr	120

* Pulsing was done with 1μmole/m²/s

Figure 15: The PhyB-PIF3 light switch bistability and reversibility with far-red light and performance in several cell types. (A) Plasmids optimized for an endogenous PhyB-PIF3 light switchable promoter. (B) Testing the reversibility of the PhyB-PIF3 light-switchable promoter in mammalian cells. Cells were in darkness, illuminated with 735nm far-red light, 660nm red light for 24 hours, or with 12 hours or red light followed by darkness or followed by far-red light (n=3). (C) Testing the PhyB-PIF3 light switch in four different cell types. Cells were transfected, then illuminated with red light for 24 hours as shown in Figure 3C. (n=4). ho1 = heme oxygenase, pcyA = phycocyanobilin:ferredoxin oxidoreductase, petF = ferredoxin, petH = ferredoxin:oxidoreductase/fnr, IRES = Internal Ribosome Entry Site, MTS = Mitochondrial Targeting Sequence, NLS = Nuclear Localization Sequence, P2A = 2A self-cleaving peptide, AD = Activation Domain, DBD = DNA Binding Domain, R/FR = Red light/Far-red light. (Error bars=s.d., (*) = $p<0.05$, Statistics were calculated using one-way ANOVA with Bonferroni post-test using GraphPad Prism 5.01). n = individual experiments.


A Plasmid maps
pPKm-230
PIF3 1-523
MTAD
PhyB 1-621
DBD
pPKm-248
tpcyA
tho1
tpetF
tpetH
pPKm-202
UAS
F-luciferase
pRL-TK
Renilla
EF1α Promoter
CMVmin
TK promoter
IRES
MTS
NLS
P2A
Puro
B Reversibility and bistability
Fold activation (R/FR)
400
300
200
100
0
*
*
Dark
24hr FR
24hr Red
12hr Red 12hr Dark
12hr Red 12hr FR
C Gene activation in four cell types
Fold activation (R/FR)
700
600
500
400
300
200
100
0
HEK
HUH
HELA
3T3

2.4.2 Light Sensitivity of the Mammalian PhyB/PIF33 Gene Switch Using Endogenously Produced PCB

PhyB/PIF3 optogenetic systems in animal cells have mostly been characterized in conditions where PCB is added externally. However, PCB degrades rapidly in cell culture media (Müller, Engesser, Metzger, *et al.*, 2013), which affects PhyB's light sensitivity over long time spans (Li *et al.*, 2011). Since our constructs enable constant endogenous production of PCB, we sought to test the light sensitivity of the PhyB/PIF33 switch (pPKm-230) with the endogenously produced chromophore. We illuminated transfected cells with the activating red light, at different intensities for 24 hours, and found that light intensities of 1.00μmol/m^2/s, 0.1μmol/m^2/s, and 0.01μmol/m^2/s achieved similar high levels of gene activation (Figure 14B and 14D). In contrast, transfected cells illuminated with a light intensity of 0.001μmol/m^2/s had a significantly lower gene response ($p<0.05$).

Since the system is bistable (Smith *et al.*, 2016), we reasoned that activating with intensities between 1.0-0.01 μmol/m^2/s, which activate the system over a long time span (24 hours), may not represent saturating amounts of light for shorter illumination times (Mattis *et al.*, 2011). To test this hypothesis, we characterized the gene switch using these same light intensities, but with a single one-minute pulse of red light (Figure 14C and 14E). Unlike the 24-hour illumination experiment, we found that when we illuminated the cells with red light for 1 minute, light intensities of 0.1μmol/m^2/s and 0.01μmol/m^2/s had a significantly lower gene response than an intensity of 1.0μmol/m^2/s ($p<0.001$). This finding highlights that for characterizing these light responsive bistable proteins, we should consider both the light intensity and duration of illumination. For example, our results using 0.1μmol/m^2/s and 0.01μmol/m^2/s show that those intensities are not saturating with a one-minute pulse, but those same intensities induce saturating activation levels over 24 hours

(Figure 14D and 14E). This is expected from a system that is bistable with a long-lived activation state (Mattis *et al.*, 2011), inactive molecules not activated in the first minute will be activated later if light is continuously applied, eventually activating all of the light-sensitive molecules.

2.4.3 Endogenous Mammalian PhyB/PIF33 Gene Switch Bistability and Reversibility with Far-red Light

We further tested the light sensitivity and bistability by shining activating red light at different pulse intervals (Figure 14F). As controls, we illuminated HEK293 cells with continuous 1.0μmol/m^2/s or 0.1μmol/m^2/s red light for 24 hours and found they reach similar levels of gene activation. In addition to continuous illumination, we utilized alternating light/dark cycles composed of 1 minute of red light and 4, 9, or 29 minutes of darkness (1m/4m, 1m/9m, 1m/29m respectively) for 24 hours. Continuous red light at 0.1μmol/m^2/s, as well as the 1m/4m and 1m/9m conditions, did not produce statistically different activation levels (Figure 14F). In contrast, the condition with 0.1μmol/m^2/s of red light pulsed at 1m/29m had significantly lower activation levels than continuous light and pulsed light in the 1m/4m and 1m/9m conditions (Figure 14F, $p<0.05$). Because the 1m/9m (blue star) condition has one-tenth the number of photons as 0.1μmol/m^2/s in total photon flux, it is equivalent in the number of photons to 0.01μmol/m^2/s of continuous illumination or 183nW/cm^2 for 660nm red light. This agrees with the result where the same total amount of light is applied continuously, suggesting that the activation state of PhyB is much longer than the 9-minute dark interval (Figure 14D and 14F).

Interestingly, we also found that cells containing the PhyB/PIF33 system had a slightly higher level of gene activation in the darkness than cells in the presence of far-red light, potentially due to the bistability of the protein (Figures 14F). Thermodynamically, in darkness, a mixed population of species (*Pf* and *Pfr* forms) is the expected nature of a bistable molecule, since some PhyB molecules can spontaneously switch to the "activated state". Therefore, the proportion of activated PhyB molecules should be higher in darkness than when PhyB is illuminated with a deactivating far-red light.

Since pulsing the light on a minute time scale achieved similar levels of activation as continuous light (Figure 14F), we decided to test the duration of the activated state of PCB bound PhyB (PhyB·PCB) by increasing the spacing between red light pulses as shown in Figure 14G. Our results show similar levels of gene activation for red light pulses delivered for one minute every eight hours, six hours, four hours, two hours, one hour, and a half hour at 1 μmol/m^2/s (Figure 14H). However, a pulse delivered every 12 hours (a total of two pulses in the 24 hour period) produced significantly lower gene activation than the pulses delivered in the shorter intervals (Figure 14H). It is possible that those two pulses in the 24-hour period delivered too little total amount of light to fully activate the system (Figure 14I). However, this data still supports that the switch effectively stays "on" for at least eight hours following a one-minute pulse of 1μmol/m^2/s of red light (Figure 14H, blue arrow). In terms of total light delivery (μmol/m^2), the one-minute pulses every 8 hours using 1.0μmol/m^2/s is effectively equivalent to the number of photons with continuous light at 0.0021μmol/m^2/s or 38nW/cm^2 for 660nm light, which is a strikingly small amount of light and speaks to the high sensitivity of this system.

One hallmark of PhyB based optogenetic switches is their conformational reversibility upon absorption of another photon of a different wavelength (Smith *et al.*, 2016). While the ability for PCB bound PhyB (PhyB·PCB) to isomerize upon red light absorption and reverse upon far-red light absorption has been previously shown (Shimizu-Sato *et al.*, 2002), whether the PhyB(1-621)-DBD and PIF3(1-524)-AD interaction was reversible by far-red light when expressed in mammalian cells has not been tested (Beyer *et al.*, 2015; Levskaya *et al.*, 2009). To test the reversibility of the switch, we exposed HEK293 cells, transfected with the PhyB/PIF33 switch and endogenously producing PCB constructs (Figure 15A), to either 24 hours of red light, 12 hours of red light followed by 12 hours of darkness, or 12 hours of red light followed by 12 hours of far-red light (Figure 15B). Luciferase expression was significantly lower in cells shifted into darkness after 12 hours of continuous red light than cells exposed to 24 hours of light ($p<0.05$), indicating PhyB reversed to its inactive state once red-light illumination ended.

Compared to switching from red light to darkness, switching from red to far-red light showed significantly lower luciferase expression, indicating that the far-red light inactivated the gene switch (red box, $p<0.05$). This result indicates that after red light activation, the switch remains on for some time in the darkness and that it can be switched off with far-red light. This finding has important implications for the switch's ability to control genes since it shows that the gene expression levels can be titrated temporally by timing the duration of red light or by red light followed by far-red light. Thus, this system can be used for spatial control by patterning red and far-red light for targeted localization of gene activation (Adrian, Nijenhuis, Hoogstraaten, Willems, & Kapitein, 2017).

2.4.4 Genetically Encoded PhyB/PIF33 Gene Switch in Several Mammalian Cell Lines

We also tested the PhyB/PIF33 gene switch performance in different cell types containing endogenously produced PCB. We transfected HEK293, hepato-cellular carcinoma (HUH-7), HeLa, and mouse fibroblasts (3T3) cells with the PhyB/PIF33 gene switch and HO1+PcyA+Fd+FNR plasmids (pPKm-230 and pPKm-248, respectively). We used 1 μmol/m^2/s of red light illumination in a cycle composed of 1-minute pulses of red light followed by 4 minutes of darkness, for a duration of 24 hours (Figure 15C). The PhyB/PIF33 switch with endogenously produced PCB activated luciferase about 280-fold in HEK293 cells, 70-fold in HUH-7 cells, 300-fold in HeLa cells and 440-fold in 3T3 cells. These findings show that the system is effective in producing PCB and activating different mammalian cell types.

While we have highly optimized the PhyB/PIF33 light switch with endogenously produced PCB, there are several ways to customize the levels of activation or leakiness to tailor it to specific cell types and applications. For example, different activation or repression domains could be used (Figure 12). In addition, there are still other permutations of gene fusions that can be tested in future studies that may further enhance this system, such as using a DBD on the N-terminus of PhyB or optimizing linker sequences. Using a stronger or tissue-specific promoter to drive expression of PCB or PΦB biosynthetic enzymes may also lead to higher activation levels or can restrict light sensitivity to specific cell types (Qin *et al.*, 2010). As presented in this research, using wavelengths that are optimal for tissue penetration (Kaberniuk *et al.*, 2016; John Y Lin *et al.*, 2013), the PhyB(1-621)-PIF3 gene switch with endogenously produced PCB is among the most light-sensitive optogenetic switches.

2.5 Summary

We demonstrate the utility of coupling the PCB metabolic pathway and a phytochrome based tool to control biological processes with NIR light in mammalian cells. We were able to genetically encode mammalian cells to produce the chromophores needed for phytochrome systems and to deploy a red-light activated gene switch based on the light-responsive PhyB/PIF33 module. While we have highly optimized the fully endogenous red-light gene switch, we have also demonstrated several ways through which customization of activation levels or leakiness is possible, to tailor it for specific applications.

As stated above, modifying the promoter can greatly affect the level of activation at the expense of leakiness. We also found that the VPR activation domain is a stronger activator than MTAD, however, VPR is 8.9X larger than MTAD, which can create problems in transfection efficiency. There are still other permutations of gene fusions that were not tested in our study that may further enhance this gene switch, such as DBD on the N-terminus of PhyB or optimizing linker sequences. It may also be the case that using a stronger or tissue specific promoter to drive expression of PCB or PΦB biosynthetic enzymes could lead to higher activation levels or restrict light sensitivity to specific cell types (Qin *et al.*, 2010).

Our fully genetically encoded system works robustly in several cell types and can be used widely in optogenetics. For example, with our switch, it is possible to make light-sensitive model organisms to instantaneously control genes deep into tissue due to NIR's tissue penetration properties. The endogenous production of chromophores like PCB enables the *in vivo* use of several existing and compatible optogentic tools to regulate cell signalling (Levskaya *et al.*, 2009; Toettcher *et al.*, 2011), cell migration (Levskaya *et al.*, 2009), or protein localization (Levskaya *et al.*, 2009). Complementarily, our red-light activated gene switch, a fully endogenous NIR-PhyB switch with Fd+FNR matching, provides long sought goals for non-invasive optogenetics and genetically-efficient encoded production of a multitude of molecules from one species in another.

Chapter 2, in part, is a reprint of the material as it appears in ACS Synthetic Biology, 2018. “Biosynthesis of Orthogonal Molecules Using Ferredoxin and Ferredoxin-NADP+ Reductase Systems Enables Genetically Encoded PhyB Optogenetics" Phillip

Kyriakakis, Marianne Catanho, Nicole Hoffner, Walter Thavarajah, Vincent Jian-Yu Hu, Syh-Shiuan Chao, Athena Hsu, Vivian Pham, Ladan Naghavian, Lara E. Dozier, Gentry Patrick and Todd P. Coleman. DOI: 10.1021/acssynbio.7b00413.

CHAPTER 3 HIERARCHICAL ATTENTION NETWORKS APPLIED TO PROTEIN SEQUENCES

Reproduced in part with permission from:

Catanho, M., Gao, S., Kyriakakis, P., Coleman, T.P., Ramanathan, A.

“Discovering sequence coevolution signatures with hierarchical attention networks”. In preparation for submission.

3.1 Abstract

One of the keys to understanding life at the molecular level is to understand the components driving a protein’s function. Sequencing technologies have improved greatly over the last few years and with it an increase in millions of functionally uncharacterized proteins. In comparison, experimental assays are still lacking, greatly restricting studies that depend on functionally annotated proteins. Given the volume of data and lack of annotations, automated analysis and predictions of important residues for protein function and structure have emerged as a promising trend in proteomics. Here, I proposed a Hierarchical Attention Network for Proteins (HANprot), deploying a bidirectional long-short-term memory unit, capable of predicting residues and sectors of importance for a protein’s function. HANprot uses multiple sequence alignments as input and requires minimum preprocessing. Applied to the PDZ protein family, the analysis of the residues identified by HANprot utilizing on multiple sequence alignments indicates physically-connected protein residues when checked against three-dimensional structures. Functionally relevant residues, identified based on database annotations, are identified as well. The attention residues highlighted by my methodology can be used as a tool select

few residues or short stretches of amino acids of long protein sequences, enabling quicker exploration of perturbation analyses, drug response, and enzyme kinetics, among other possibilities.

3.2 Introduction

Physics and evolution come together to generate a complex set of relationships in proteins, of which several experimental and theoretical experiments have sought to qualify. In 1973, Anfinsen *et al.* showed that the function of a protein could be predicted from its amino acid sequence, and demonstrated the intimate relationship between native tertiary structures and functionality (Anfinsen, 1973; Sadowski & Jones, 2009). An important aspect of a protein's functional characterization is the determination of important residues mediating its function (Sadowski & Jones, 2009). These residues exist in short regions of a protein's sequence and three-dimensional structure (Fischer, Mayer, & Söding, 2008; Watson, Laskowski, & Thornton, 2005). Often, these residues obey a specific arrangement and, given evolutionary constraints, they remain conserved over time, forming active sites or binding sites for other molecules, or enabling protein-protein interactions (Fischer *et al.*, 2008; Ouzounis, Pérez-Irratxeta, Sander, & Valencia, 1998; Watson *et al.*, 2005).

In their review of computational methods of protein function determination, Watson *et al.* argued that the function of some proteins is enabled by a small number of those residues, usually grouped in a region of the three-dimensional protein structure (Watson *et al.*, 2005). In DNA-binding proteins, for example, residues localized in the surface of the protein's structure enabled it to bind to a DNA sequence (Sadowski & Jones, 2009). In enzymes, a small number of residues in the active site are responsible for the enzyme's catalytic function (Di Lena, Nagata, & Baldi, 2012; Nagao, Nagano, & Mizuguchi, 2014).

Those residues remain conserved through evolution, in a process named residue coevolution (Lockless & Ranganathan, 1999; Ouzounis *et al.*, 1998).

Protein coevolution follows similar principles as Hebb's rule, commonly known as the "fire together wire together" of neurobiology: residues that have undergone evolutionary constraints together are likely to be conserved together (Ribeiro & Ortiz, 2015). It often indicates the residues' importance for the protein's function, structural stability and chemical activity. As such, protein structure and function depend on cooperation between residues. As pointed by Halabi *et al.*, the specific position or the distribution of residues at specific positions are not to be considered independent of one another (Halabi, Rivoire, Leibler, & Ranganathan, 2009). Halabi *et al.* also posited that residues contribute unequally, but cooperate to form a protein's structure and function (Halabi *et al.*, 2009). Furthermore, sequence conservation is not only determined by local interactions, but is also influenced by interactions of residues that are further apart (Rost & Sander, 1993). This has driven a central problem in computational biology: discriminating between residues that are conserved for strictly for structural reasons from those that are conserved for other functional reasons (Sadowski & Jones, 2009).

Identifying active site residues strictly from protein three-dimensional structure has been shown to be a difficult task and structural information is only available for a small fraction of proteins (Fischer *et al.*, 2008). Additionally, large-scale sequencing has resulted in an exploding widening of the sequence-structure-function gap (Ofran & Rost, 2003; Rost & Sander, 1993, 1994). Methods relying on the relationship between sequence and structural information have been shown to be largely complementary (Rost & Sander, 1994). Predictions that are structure-focused can fail to recognize the binding sites, especially when those sites undergo a drastic conformational change during ligand binding

(Fischer *et al.*, 2008). Another shortfall of predicting allosteric function from a protein structure occurs when remote pockets on the surface of a protein act as binding site, over time replacing major clefts as the most important site for the protein's allosteric control (Fischer *et al.*, 2008; Lockless & Ranganathan, 1999). Combined with the increase in knowledge on the relationship between structure and function, those factors have motivated the development of computational methods to predict different characteristics of proteins using sequences (de Juan, Pazos, & Valencia, 2013; Elloumi, Iliopoulos, Wang, & Zomaya, 2015; Rost & Sander, 1993; S. Wang, Peng, Ma, & Xu, 2016).

Historically, one of the first strategies for predictions that relied on sequences used non-annotated databases, even though some relationships, such as between homology and function, are not highlighted (Rost & Sander, 1993). Other methods have used multiple sequence alignments (MSAs) to infer protein familial relationships and its approximate structure (Binkowski, Adamian, & Liang, 2003). Overall, it is widely accepted that MSAs often contain more information about the protein, its structure and function than a single sequence (Rost & Sander, 1993, 1994; Sadowski & Jones, 2009; Watson *et al.*, 2005). On the other hand, studies based on MSA have revealed that structure is more conserved than sequence, and that different sequences can adopt the same structure (Rost & Sander, 1993). Multiple sequence alignments are an evolutionary record of unlikeliness: even evolutionarily-linked protein residues in different proteins can have identical structure and dissimilar sequence (Rost & Sander, 1993). Computational tools that rely on co-evolution to predict functional sites focus on changes happening between interacting residues, which can provide information on the protein stability, function and folding (Sadowski & Jones, 2009).

Even when experimental determination of important residues in a protein is unattainable at first, sequence similarity can often generate meaningful predictions. As such, sequence conservation is still considered to be the greater single contributor to assigning function to protein structures and a powerful predictor of functional residues (Capra & Singh, 2007; Lockless & Ranganathan, 1999; Ouzounis *et al.*, 1998; Rost & Sander, 1994). Overall, sequence alignments have been used to (i) improve protein secondary structure prediction (Cuff & Barton, 2000), (ii) predicting damaging missense mutations (Adzhubei *et al.*, 2010), (iii) determine similarities between target proteins (Baker & Sali, 2001), (iv) predict physicochemical properties (Tung & Ho, 2007), (v) evolutionary relationships (Chenna *et al.*, 2003; Feng & Doolittle, 1987; Strimmer & von Haeseler, 1997), (vi) map protein–protein interaction networks (Ofran & Rost, 2003; Yan, Dobbs, & Honavar, 2004), among others.

Identifying residues of importance or functional sites through computational methods has been addressed by several groups (Hamilton, Burrage, Ragan, & Huber, 2004; Lockless & Ranganathan, 1999; Mihalek, Res, & Lichtarge, 2004; Morcos *et al.*, 2011), but most have focused mostly on protein-protein interactions. The methodologies used include mutual information and information theoretic approaches (Morcos *et al.*, 2011; Yan *et al.*, 2004), machine learning (Pugalenthi, Kumar, Suganthan, & Gangal, 2008; Somarowthu & Ondrechen, 2012; Somarowthu, Yang, Hildebrand, & Ondrechen, 2011), statistical approaches (Hamilton *et al.*, 2004; Mihalek *et al.*, 2004; Morcos *et al.*, 2011), continuum electrostatics (Elcock, 2001), and sequence conservation (Capra & Singh, 2007; Lockless & Ranganathan, 1999). Those methods leveraged the evolutionary conservation, inherently captured by sequence alignments, and structural information as basis for the predictions (de Juan *et al.*, 2013).

One residue coevolution approach, called Statistical Coupling Analysis (SCA) aimed to characterize protein co-evolution and was shown to identify groups of coevolving residues (Ranganathan & Ross, 1997). Those residues have been associated to functionally relevant, conserved sectors of the protein (Halabi *et al.*, 2009; Lockless & Ranganathan, 1999; Ranganathan & Ross, 1997). A similar method of residue co-evolution, called Direct Coupling Analysis (DCA) utilizes a premise similar to SCA to infer the statistical correlation between co-evolving pairs of residues in a protein (Morcos *et al.*, 2011). In DCA, correlated residues are used as a starting point to identify direct and indirect interactions using a global inference approach implemented by an algorithm that relies on a maximum-entropy statistical model for entire protein sequences and use the conditional mutual information to obviate spurious relationships. DCA, SCA and similar algorithms are limited by their reliance of protein structural data to generate meaningful predictions, the ambiguity in the results and are particularly prone to returning long lists of false positive matches (Min, Lee, & Yoon, 2017; Zvelebil & Baum, 2008).

As evidenced by those methods, evolutionary information stored in multiple sequence alignments can be used as input to neural networks to predict residues of importance with increased accuracy. Machine learning methods are poised for such approaches, being able to learn relationships from data and derive predictive models without the need for an *a priori* definition or strong assumptions about underlying mechanisms (Angermueller, Pärnamaa, Parts, & Stegle, 2016). For proteins, this is particularly beneficial, since they often have unknown or insufficiently defined mechanisms, or their information is stored in non-annotated databases. The most accurate models of gene expression levels (Lamb *et al.*, 2006), genomics (Park & Kellis, 2015; Quang, Chen, & Xie, 2015), proteomics (Jo, Hou, Eickholt, & Cheng, 2015; Spencer, Eickholt, & Jianlin Cheng, 2015; S. Wang *et al.*, 2016), metabolomics (Aggio, Villas-Bôas, & Ruggiero, 2011; Min *et al.*, 2017), all rely on machine learning (Angermueller *et al.*, 2016). These methods usually employ convolutional neural networks (CNN) or support vector machines (SVM) (Sønderby, Sønderby, Nielsen, & Winther, 2015). A comprehensive review of machine learning approaches in bioinformatics can be found here (Angermueller *et al.*, 2016; Min *et al.*, 2017; Park & Kellis, 2015).

Differently from CNNs and SVMs, bidirectional long-short-term memory (BLSTM) units are a class of recurrent neural networks (RNNs) designed to handle sequential data. RNNs' units share identical weights at each time step, which allows for information to flow across a sequence through recurrent weights placed between each hidden layer (Sønderby *et al.*, 2015). Even though RNNs have been used in bioinformatics in contact map prediction (Di Lena *et al.*, 2012), and to solve protein secondary structure (Magnan & Baldi, 2014), they have been shown to be difficult to train due to vanishing and exploding gradients (Sønderby *et al.*, 2015). This caveat has made RNNs unreliable when exploitation of long-range dependencies is needed (D. Wang & Nyberg, 2015). To address this shortcoming, Hochreiter *et al.* developed BLSTMs, which rely on input, modulation, forget and output gates memory cells instead of the standard sigmoid or tangent units used by RNNs (Graves, Mohamed, & Hinton, 2013; Sønderby *et al.*, 2015). Figure 16 shows diagrammatically the difference between RNNs and LSTMs. Hierarchical Attention Networks (HAN) build upon the work of Graves *et al.*, and were first proposed for document classification (Graves *et al.*, 2013; Yang *et al.*, 2016). When used in natural language processing (NLP), HAN architectures assume that documents have hierarchical structure (words, sentences, documents) and that different words are inherently different in the information they convey (Yang *et al*., 2016).

Here, I employ a hierarchical attention network for protein sequences (HANprot), trained on public, non-annotated datasets from NCBI and PFAM, to identify residues of importance in the third PDZ domain of the PSD95 protein (PSD95-PDZ3, hereby ‘PDZ’), Phytochrome B (PhyB), and other protein families. A detailed analysis is provided for PDZ and PhyB. PDZ was chosen as a proof of concept protein, due to its small size (81 amino acids), well-defined structure, and varied conservation throughout its sequence. Further, I compared HAN results of highly annotated proteins and found it over-performs a comparable method (SCA) in most proteins families analyzed in this work.

HANprot deploys a hierarchical structure with a bidirectional long-short term memory unit and an attention mechanism for protein sequence recognition, which explicitly captures both local and global interaction information in an end to end process. Similar to the assumptions applied to HANs in NLP applications, I assume that MSAs have hierarchical structure: sequences (documents) have segments (sentences), often domains or motifs, that are composed of amino acid residues (words). Depending on where a residue (a word) is located, the level of information it conveys can change. Therefore, similarly to what was proposed by Yang *et al.*, the importance of residues (words) and sequences (sentences) are highly context-dependent (Yang *et al.*, 2016). My hypothesis is that by capturing which residues are of importance through the HAN mechanism, I am effectively ranking residues on their contribution to a protein’s function.

My findings highlight the importance of residues at specific positions in a protein sequence and help define a metric for residues’ contribution to the protein structure, function, and evolution. By narrowing into which residues are important for a previously non-annotated sequence, the methodology proposed here can motivate theoretical and experimental analysis of sequence positions, towards elucidating how protein sequences encode the basic conserved biological features of a protein family. These findings can motivate a deeper theoretical and experimental analysis of deep learning architectures with the goal of understanding how protein sequences encode conserved biological properties.

3.3 Methodology

3.3.1 Protein Family Datasets

Training of HANprot was performed using multiple sequence alignments obtained from PFAM (Jones, Buchan, Cozzetto, & Pontil, 2012), or collected from the NCBI non-redundant database, utilizing an expected threshold of 0.1 and other parameters as default (National Center for Biotechnology Information, 1988; Waterhouse, Procter, Martin, Clamp, & Barton, 2009). Jalview (Version 2) was used for preprocessing of NCBI alignments and to remove 95% of sequence redundancy for those alignments. Specific protein sequences from those alignments (hereby a “reference protein”), whose sequence was fully covered within their respective PFAM alignment, was used to generate attention residue predictions. A solved 3D structure of those proteins was used to generate images depicting spatial distribution of identified residues. Appendix E: Table 5 lists the protein families used in this work, alignment size and other information.

3.3.2 Amino acid and Sequence Encoding

Integers (1 through 21) were assigned to the 20-letter amino acid alphabet, arranged from most correlated (known to naturally form groups with similar physiochemical properties) to least, following Murphy *et al.* (Murphy, Wallqvist, & Levy, 2000). Accordingly, the following order was observed in numbering the amino acids: L V E M C A G S T P F Y W E D N Q K R H. Gaps, denoted “-” in the alignments, were given a value of zero. As such, for an alignment with S sequences of length L, each amino acid is given by $x_{ij} \in [0{,}21]$, with $i \in [0, S]$ and $j \in [0, L]$ (Figure 17A).

Training labels, or targets, were formulated from Henikoff weights (Henikoff & Henikoff, 1994), henceforth referred to as the Henikoff Sequence Weight, or HSW. Henikoff weights, based on the Rumelhart backpropagation of errors method, are often used in bioinformatics for their simplicity and broad applicability, requiring minimum preprocessing (Elloumi *et al.*, 2015). Each HSW is given by a conservation score of a specific sequence given an alignment. Because HANprot inputs are from non-annotated databases or MSAs, which do not contain annotations of important residues, this is ideal since sequencing weighting methods compensate for over-representation in MSAs, often being tree-based or pairwise-distance based (Henikoff & Henikoff, 1994). HSWs were assigned to individual alignment columns, and normalized to generate a sequence weight (Figure 18). For an alignment of size $[N, L]$, where N is the number of sequences and L is the length of the sequences, assume that n_{ij}, $i = [0, N]$ and $j = [0, L]$, is the number of times residue $x_{ij} = [0{,}21]$, appears in column j, and that there are d_j different types of residues in column j. For a specific sequence i, each residue is given a weight equal to $w_{ij} = 1/(n_{ij}d_j)$. The sequence weight is given by the sum of the individual residue weight (Zvelebil & Baum, 2008). Assume that the residue weights are labeled w_{ij} for residue j of sequence i, the weight of sequence i will be given by:

$$w_i = \sum_j w_{ij} = \sum_j \frac{1}{n_{ij}d_j} \tag{1}$$

Averaging over all columns, will give the final HSW:

$$HSW_i \ = \frac{w_i}{N} \tag{2}$$

As such, the dataset's labels are given by the vector HSW_i, with $i \in [0, N]$.

3.3.3 BLSTM

BLSTMs are an extension of LSTM algorithms (Hochreiter & Schmidhuber, 1997) (Figure 16B and 16C). An LSTM is a set of recurrently connected blocks (memory blocks), containing one or more memory cells that are recurrently connected and three multiplicative units (input, output, forget). Given a sequence i with residues x_{it} with $t \in [1, L]$, embedded as mentioned above, the precise gating mechanism of updates are as follows:

$$i_t = \sigma(W_x i x_t + W_h i h_t - 1 + W_C i C_t - 1 + b_i) \tag{3}$$

$$f_t = \sigma\left(W_x f x_t + W_h f h_t - 1 + W_C f C_t - 1 + b_f\right) \tag{4}$$

$$C_t = f_t C_t - 1 + i_t \tanh(W_x C x_t + W_h C h_t - 1 + b_C) \tag{5}$$

$$o_t = \sigma(W_x o x_t + W_h o h_t - 1 + W_C o C_t - 1 + b_o) \tag{6}$$

$$h_t = o_t \tanh(C_t) \tag{7}$$

LSTMs are an efficient way to approach sequence learning since it relies on a casual structure: the state at time t only receives information from the present (x_t) and the past ($x_1, \dots, x_{t-1}$). However, for a protein sequence of length L, a residue at position x_m, $m \in L$, can be closely related to both the residues in its "past" ($t < m$, with $t, m \in L$) and its future ($t > m$, with $t, m \in L$). In other words, if we consider the protein sequence to be like a time sequence, both future and past, that is the whole protein sequence, can provide insight about hidden states that can help predict important characteristics of a protein. As shown in Figure 16B, Equation 7 for LSTMs is replaced by Equations 8 and 9 for BLSTMs, since there are now forward hidden units h_1, which reads the sequence s_i from x_{i1} to x_{iL}, and backward hidden units h_2, which reads the sequence s_i from x_{iL} to x_{i1}:

$$h_{i1} = f\left(W_{x_i} h_{i1} x_{it} + W_{h_{i1}} h_{i1} h_{i1t} - 1 + b_{h_{i1}}\right) \tag{8}$$

$$h_{i2} = f\left(W_{x_i} h_{i2} x_{it} + W_{h_{i2}} h_{i2} h_{i2t} - 1 + b_{h_{i2}}\right) \tag{9}$$

The final h_{it} is given by the sum of h_1 and h_2.

3.3.4 HANprot Architecture

The structure of the HAN deployed in this work follows that of Yang *et al.* (Yang *et al.*, 2016), which assumes that not all words (residues) contribute equally to the representation of the sentence (sequence) meaning (function). The attention mechanism is adopted on the input features (amino acid residues) with the BLSTM, each with 500 hidden units, to learn the important regions in a protein sequence and the crucial sequences in an alignment. The proposed architecture projects the protein's MSA into a vector representation. Assume that each protein MSA has S sequences s_i, with $i \in [1, S]$, and each sequence contains L residues: x_{it}, with $t \in [1, L]$, represents the residues in the $\imath$th sequence. Figure 19 shows the HANprot architecture in detail.

Two types of inputs were tested:

(i) Whole sequence, where x_{it} represents the residues in the $\imath$th sequence of length L.

(ii) Overlapping windows (Figure 17B), o_{ki} (here I abuse the notation), in which each sequence i was divided into windows of size k with an m-residue overlap. For this work, $k \in [5, 12]$ and $m \in [1, 4]$. HSW (labels) are assigned per sequence, such that labels across methods are identical for a sequence i or for a group of windows generated from a sequence i.

The attention mechanism, adopted from (Ahmed, 2017) extracts the residues of importance and aggregates their representation to form a sequence vector:

$$u_{it} = \tanh(W_w h_{it} + b_w) \tag{10}$$

$$\alpha_{it} = \frac{\exp\left(u_{it}^T u_w\right)}{\sum_t \exp\left(u_{it}^T u_w\right)} \tag{11}$$

$$p_i = \sum_t \alpha_{it} h_{it} \tag{12}$$

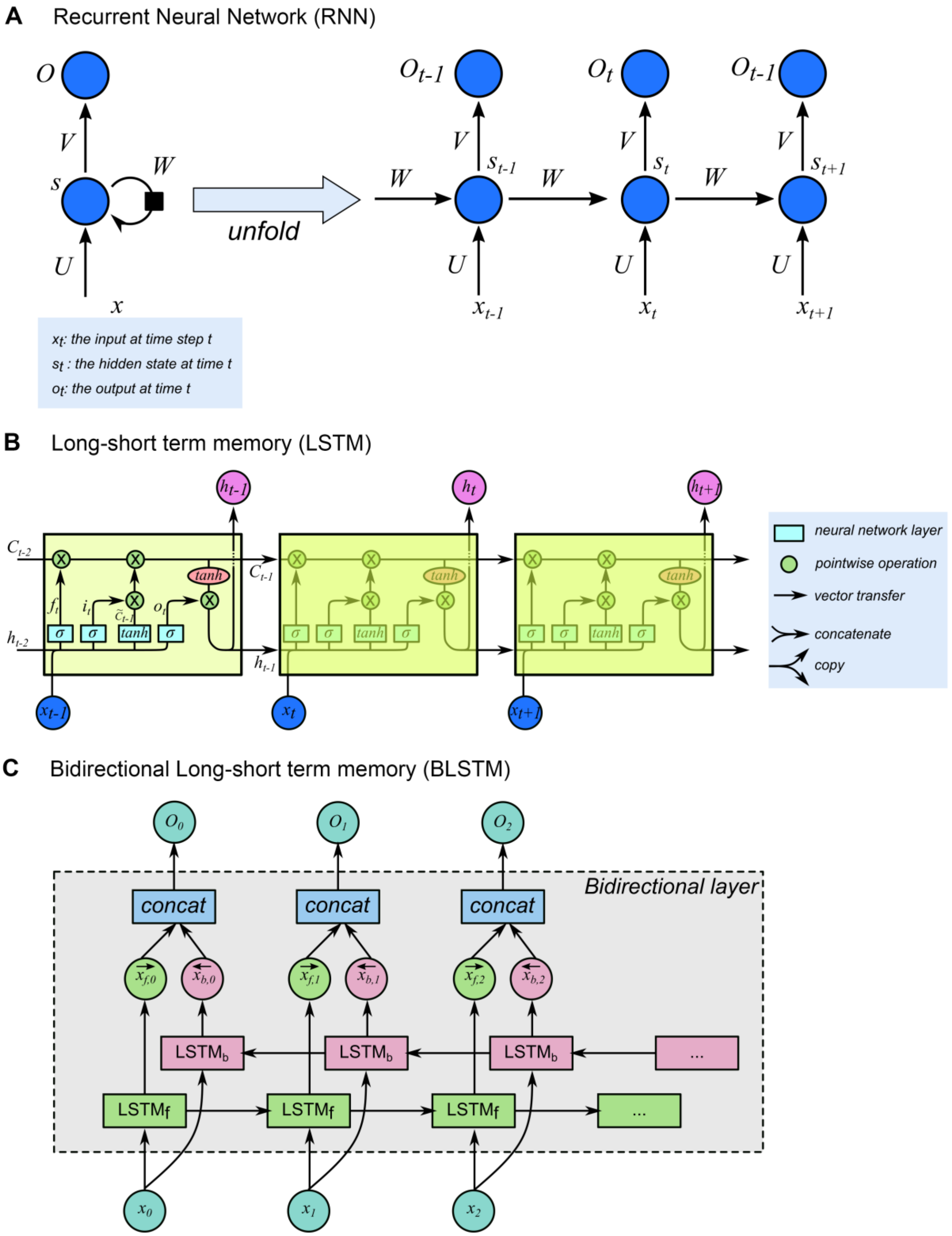
As such, the residue annotation h_{it} is feed through the two layers of the BLSTM to get u_{it} (hidden representation of h_{it}). I measure importance of a residue as the similarity of u_{it} with a residue level context vector u_w (randomly initialized and jointly learned during training), resulting in a normalized importance weight α_{it} through a softmax function. I then compute a sequence vector p_i, given by the weighted sum of the residue annotations built on the weights. In a way, u_w is a representation of a fixed query “What are the most important residues?” over the residues in the sequence, and p_i is a vector representation of each residue and its relevance to the model.

The same calculations are performed at the sentence level, with the only difference being that the layers learn the encoding of the whole sequence, as given by the Henikoff weight. Since I have trained the network on non-annotated datasets, I do not perform a classification task, but rather fit the model using a final sigmoid layer:

$$q \ = \ sigmoid \left(W_q s + b_q\right) \tag{13}$$

I minimize the loss using a mean squared error function and an adaptive learning rate method (Tieleman & Hinton, 2012).

Figure 16: RNNs and LSTMs architectures. (A) A simple folder and unfolded RNN architecture.(B) Three common LSTMs, composed of cell, input, output and forget gates, (C) BLSTM with a concatenated output.



A Recurrent Neural Network (RNN)

x_t: the input at time step t
s_t : the hidden state at time t
o_t: the output at time t

B Long-short term memory (LSTM)

C Bidirectional Long-short term memory (BLSTM)

Figure 17: Residue encoding and windows. (A) Each residue is assigned a value in between zero and 21, with gap (‘-‘) being zero. The entire numerical vector is then used as input to HANprot. (B) For the same sequence, I show how windows are determined, using a window size example of five and an overlap size of two. Each sequence is then given by a group of windows (color segments), which is the input for the window method adaptation of HANprot.

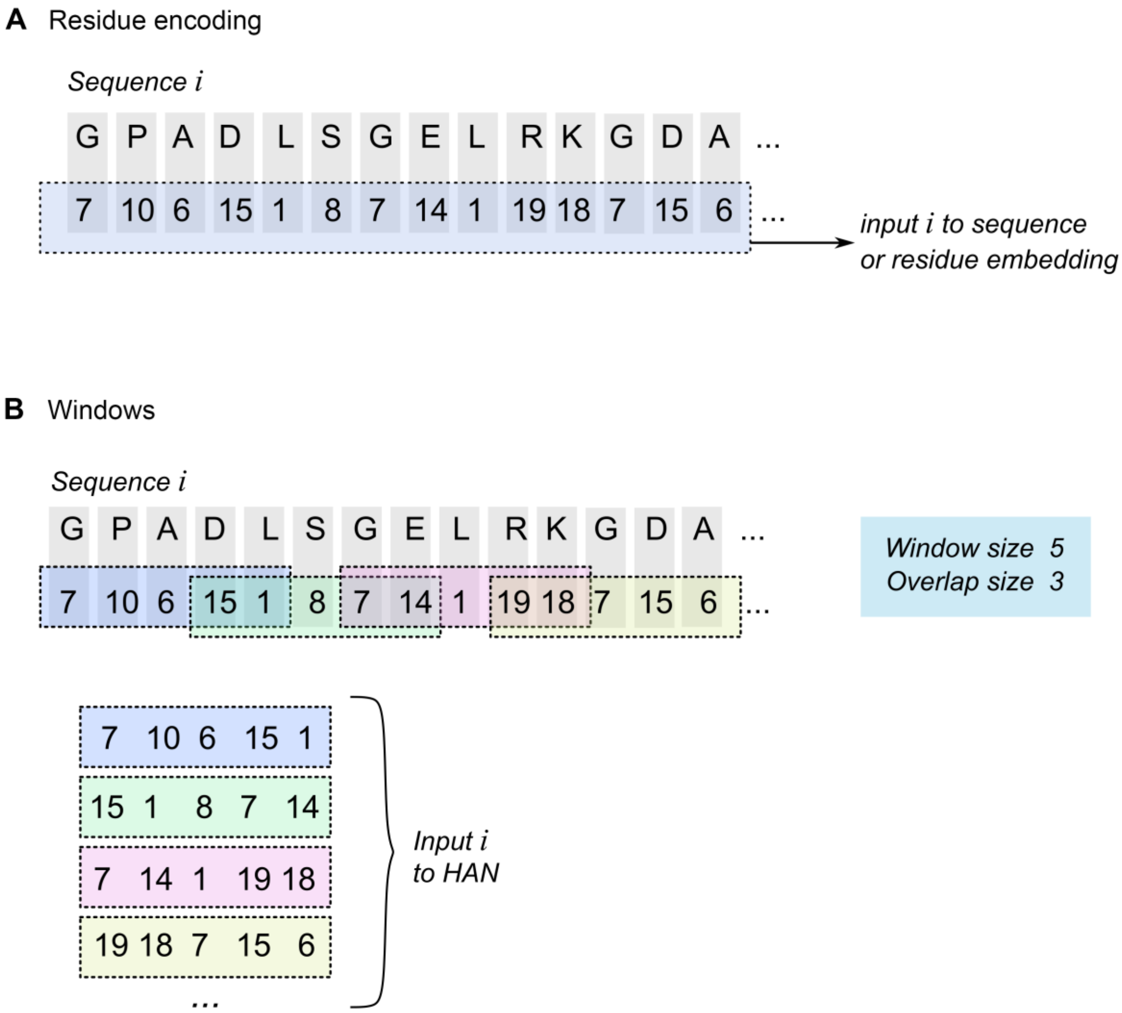

A Residue encoding
Sequence i
G P A D L S G E L R K G D A ...
7 10 6 15 1 8 7 14 1 19 18 7 15 6 ...
input i to sequence
or residue embedding
B Windows
Sequence i
G P A D L S G E L R K G D A ...
7 10 6 15 1 8 7 14 1 19 18 7 15 6 ...
Window size 5
Overlap size 3
7 10 6 15 1
15 1 8 7 14
7 14 1 19 18
19 18 7 15 6
...
Input i
to HAN

Figure 18: Henikoff weight calculation. For each position j, each residue in a sequence i is given a weight $\mathrm{w_{ij}}$ equal to $1/(\mathrm{n_{ij}d_j})$, where $\mathrm{n_{ij}}$ is number of times that particular residue occurs in that position, and $\mathrm{d_j}$ is the number of different residues in position j, with $j = [1, N]$, where N is the total number of sequences, and L the length of the sequence. The Henikoff sequence weight is given by the sum of each weight, and normalized by the total number of positions (L)in the sequence.

$\boldsymbol{i}$	**Sequence**	**Position (j)**						**Total**	***HSW***
		1	2	3	4	5	6		
		Position weight w_{ij}						w_i	$\boldsymbol{HSW_i = {w_i}/{N}}$
1	GDQGID	1/(1*5)	1/(1*5)	1/(3*1)	1/(1*5)	1/(2*3)	1/(3*2)	1.267	***0.211***
2	GDRGIN	1/(1*5)	1/(1*5)	1/(3*3)	1/(1*5)	1/(2*3)	1/(3*2)	1.044	***0.174***
3	GDLGVN	1/(1*5)	1/(1*5)	1/(3*1)	1/(1*5)	1/(2*2)	1/(3*2)	1.350	***0.225***
4	GDRGVQ	1/(1*5)	1/(1*5)	1/(3*3)	1/(1*5)	1/(2*2)	1/(3*1)	1.294	***0.216***
5	GDRGID	1/(1*5)	1/(1*5)	1/(3*3)	1/(1*5)	1/(2*3)	1/(3*2)	1.044	***0.174***
N=5	**Total**	**1**	**1**	**1**	**1**	**1**	**1**	**L=6**	***1***

Figure 19: HANprot layers. Residues encoded numerically are fed in windows or as a full sequence to the first layer of HANprot. A residue-level BLSTM learns the patterns of residues being inputted, and its output is inputted into the sequence-level BLSTM. The sequence-level BLSTM learns features of the whole alignment. Those features are not explored in this work. Finally, a dense layer followed by a sigmoid is used for loss minimization during training.

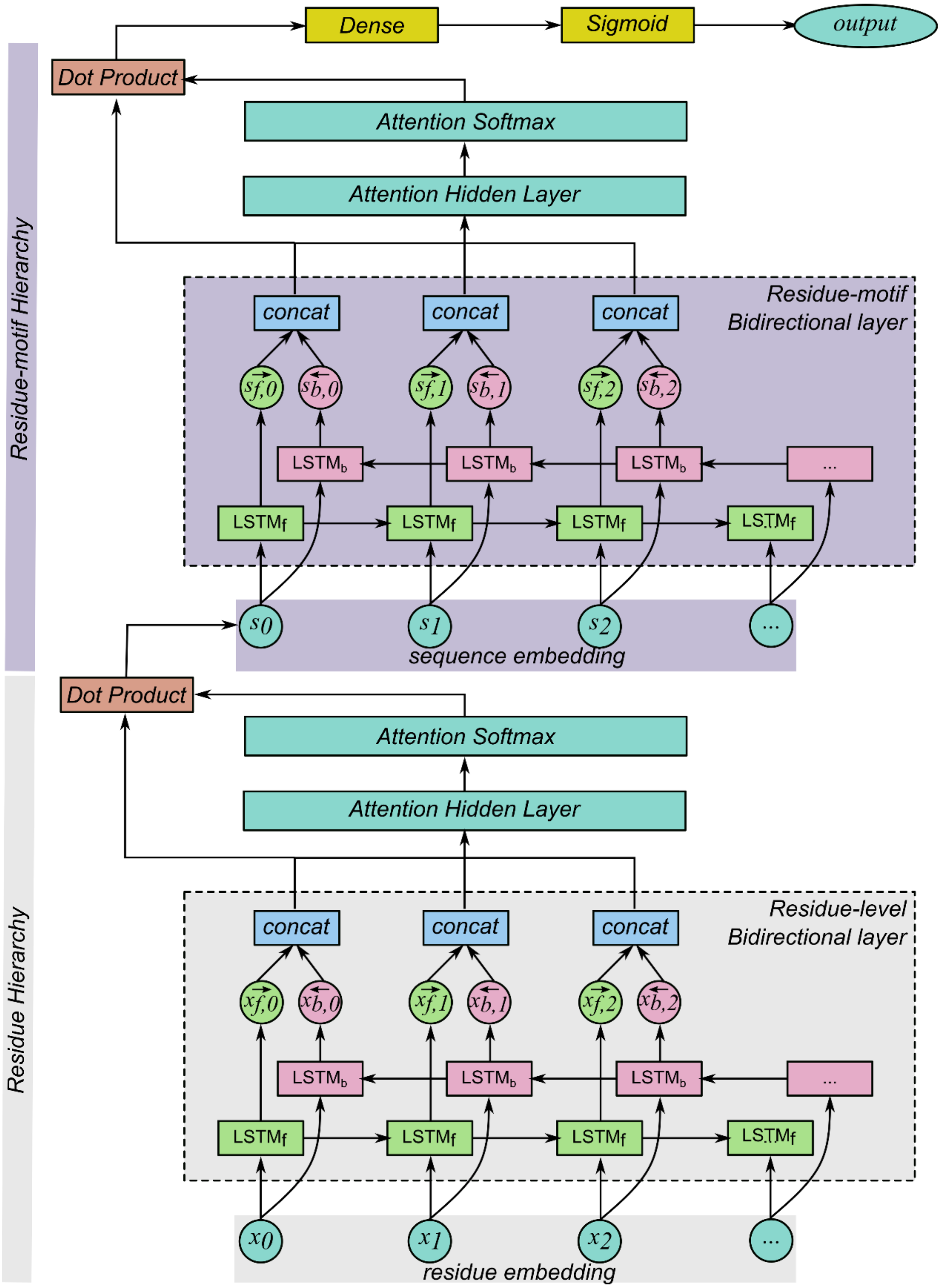
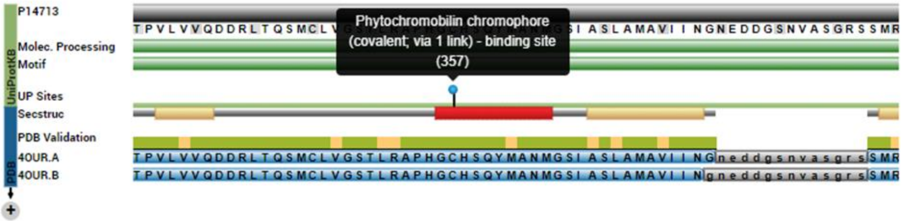
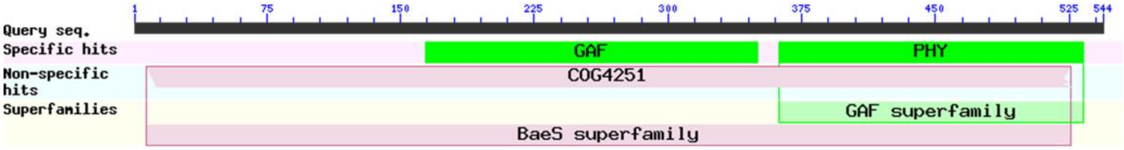

Dense
Sigmoid
output
Dot Product
Attention Softmax
Attention Hidden Layer
Residue-motif Hierarchy
Residue-motif
Bidirectional layer
concat
concat
concat
$\overrightarrow{s_{f,0}}$
$\overleftarrow{s_{b,0}}$
$\overrightarrow{s_{f,1}}$
$\overleftarrow{s_{b,1}}$
$\overrightarrow{s_{f,2}}$
$\overleftarrow{s_{b,2}}$
LSTM$_b$
LSTM$_b$
LSTM$_b$
...
LSTMf
LSTMf
LSTMf
LSTMf
s_0
s_1
s_2
...
sequence embedding
Dot Product
Attention Softmax
Attention Hidden Layer
Residue Hierarchy
Residue-level
Bidirectional layer
concat
concat
concat
$\overrightarrow{x_{f,0}}$
$\overleftarrow{x_{b,0}}$
$\overrightarrow{x_{f,1}}$
$\overleftarrow{x_{b,1}}$
$\overrightarrow{x_{f,2}}$
$\overleftarrow{x_{b,2}}$
LSTM$_b$
LSTM$_b$
LSTM$_b$
...
LSTMf
LSTMf
LSTMf
LSTMf
x_0
x_1
x_2
...
residue embedding

Figure 20: Example of annotations extracted from published databases: (A) RCSB, (B) NCBI and (C) UniProt.

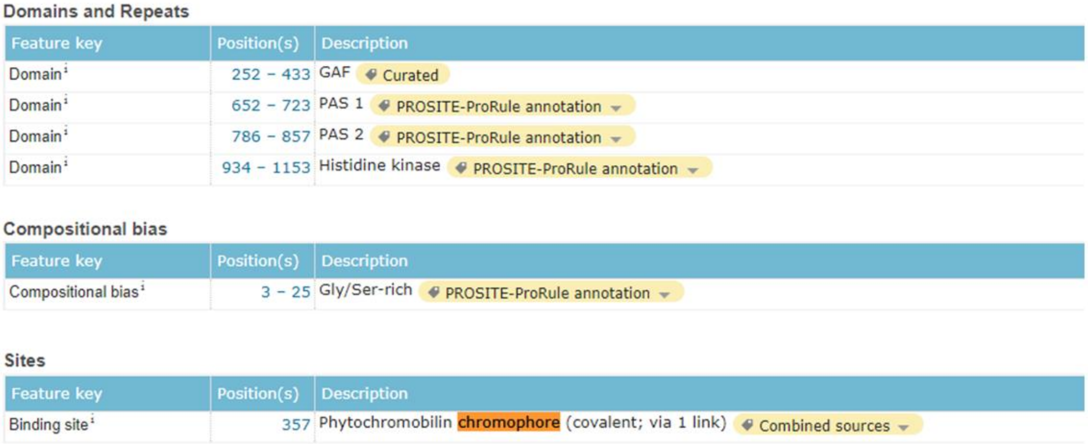
A RCSB Annotations for PhyB (PDB ID: 4OUR)


P14713
TPVLVVQDDRLTQSMCLV
SIASLAMAVIINGNEDDGSNVASGRSSMR
Phytochromobilin chromophore (covalent; via 1 link) - binding site (357)
UniProtKB
Molec. Processing
Motif
UP Sites
Secstruc
PDB Validation
4OUR.A
TPVLVVQDDRLTQSMCLVGSTLRAPHGCHSQYMANMGSIASLAMAVIINGneddgsnvasgrsSMR
4OUR.B
TPVLVVQDDRLTQSMCLVGSTLRAPHGCHSQYMANMGSIASLAMAVIINgneddgsnvasgrsSMR
PDB
+

B NCBI Annotations for PhyB (PDB ID: 4OUR)



C UniProt Annotations for PhyB (PDB ID: 4OUR)



Domains and Repeats

Feature key	Position(s)	Description
Domain[i]	252 – 433	GAF Curated
Domain[i]	652 – 723	PAS 1 PROSITE-ProRule annotation
Domain[i]	786 – 857	PAS 2 PROSITE-ProRule annotation
Domain[i]	934 – 1153	Histidine kinase PROSITE-ProRule annotation

Compositional bias

Feature key	Position(s)	Description
Compositional bias[i]	3 – 25	Gly/Ser-rich PROSITE-ProRule annotation

Sites

Feature key	Position(s)	Description
Binding site[i]	357	Phytochromobilin chromophore (covalent; via 1 link) Combined sources

3.3.5 Training and Validation

HANprot was trained for 200 epochs for each protein family, using batches of 300 sequences, with a dropout rate of 30%, learning rate of 0.1, and using an RMSprop optimizer (Tieleman & Hinton, 2012). Each dataset was randomized, and split into a 0.6/0.3/0.1 partition for train/validation/test sets. Each BLSTM was composed of 500 hidden units, as mentioned previously. The network weights at the epoch with highest validation performance is saved and used to evaluate the model performance on the test set. For each protein family, attention values for the full sequence of the protein being tested were extracted using the reference protein's sequence, as a vector p_i, with $p_i \in [0,1]$, for $i \in [0, L]$. Attention residues are determined by a threshold, given by the values above one standard deviation of the mean of p_i.

Using each reference protein's sequence, important residues were compiled from NCBI, RCSB and UniProt, and are hereby referred to as annotated binding sites (National Center for Biotechnology Information, 1988; The UniProt Consortium, 2017). Due to the lack of standardized annotation in protein databases to date, annotated binding sites or residues were defined as residues that are involved in coenzyme binding, enzymatic catalysis, effector interactions, ligand binding, etc. (Ouzounis *et al.*, 1998). Figure 20 shows annotations from those databases for PhyB, as an example. In PhyB's case specifically, to circumvent lack of annotations in those databases, as seen in Figure 20, residues of relevance were extracted from previously published data) (Burgie, Bussell, Walker, Dubiel, & Vierstra, 2014; Burgie & Vierstra, 2014). Appendix F shows similar screenshots for the other protein families used in this work. Other than PhyB, all proteins used annotated binding sites or residues from UniProt, NCBI and RCSB.

Pertinent literature was reviewed and used to extracted motif, domain and binding pocket annotations, and used in the three-dimensional structure images in this work. The collection of these manually extracted residues was used to validate the results (or in other terms, the true labels). As another method of comparison, residues determined as relevant by SCA were obtained through direct application of the SCA methodology, as instructed and utilizing MATLAB toolbox distributed by the authors (Halabi *et al.*, 2009). Since SCA does not discriminate residues with regards to their importance, function, or relevance to the protein function or structure, each residue output was considered important (e.g., received an attention value or 1).

The receiver operating characteristic's area under the curve (AUC) analysis was chosen as a measure to assess the ability of HANprot to identify important residues. This allows for uniform performance assessment. AUC scores were calculated for HANprot and SCA using the manually extracted residues (from NCBI, RCSB and UniProt) as the true label. F1 scores are also reported, and are given by:

$$Recall = TP/(TP + FN) \tag{12}$$

$$Precision = \frac{TP}{TP + FP} \tag{13}$$

$$F1 = 2 * Precision * Recall/(Precision + Recall) \tag{14}$$

Where TP is the number of true positives, TN true negatives, FP false positives and FN false negatives. For ROC AUC, given by the True Positive Rate (TPR) against the False Positive Rate (FPR).

$$TPR = \frac{TP}{FP + FN} \tag{15}$$

$$TPR = \frac{TP}{FP + FN} \tag{16}$$

3.4 Results

I analyzed existing experimental datasets to compare the functional significance of attention residues to that of annotated residues from public databases (NCBI, RCSB, UniProt), or ‘annotated binding sites’. Datasets discussed below are PDZ and PhyB. In addition, two other protein families were analyzed: Cadherin and HSP70. See Appendix E: Table 5 for dataset information. Summarized AUC results for all families are shown in Appendix G: Table 6. Here, I show that in all these cases, HANprot identifies functional positions effectively, obtaining better AUC scores than a previously established methods (SCA), in both window and full sequence methods, for 4 out of 5 protein families analyzed. HANprot not only identifies important residues for the protein’s function, but it assigns each residue a score based on the attention mechanism and perceived relevance of the residue.

3.4.1 PDZ Domains

PDZ domains are found in highly divergent species and are known to regulate diverse biological activities, having abundant protein-protein interactions (Lee & Zheng, 2010). In the mouse genome, for example, PDZ domains are present in over 300 proteins, accounting for over 900 of the domains (Lee & Zheng, 2010). PDZ domains are typically composed of 5 to 6 β-strands and 2 or 3 α-helices, often displaying a short and a long α-helix in canonical domains. Those canonical domains also present a highly conserved fold. Although they’re known to recognize the C-terminus of proteins, PDZ domains also have a single binding site known to bind to internal motifs in target proteins (Lee & Zheng, 2010). This single binding site exists in a groove between an α-helix and a β-strand (Figure 21). In this groove, several residues are located in the α-helix (α2 in Figure 21, with sequence ‘HEQAAIALKN’), and the remaining are part of a highly conserved carboxylate-binding loop, (R/K-XXX-G-Φ-G-Φ motif, X represents any amino acid residue, and Φ hydrophobic residues), whose side chains allow for hydrophobic binding of ligands, in the β2 sheet (Du, Meng, Wang, Long, & Huang, 2011). In that motif, the second Glycine residue (Gly, G) is highly conserved, whereas the first Gly is often replaced by a serine, threonine or phenylalanine (Lee & Zheng, 2010; Ranganathan & Ross, 1997).

Utilizing the PDZ dataset from PFAM, HANprot was trained using windows of sequence residues and using full sequences. Structure and sequence of the PDZ domain from the synaptic protein PSD-95 was used for visualization. Additionally, HANprot was trained with a dataset sourced from NCBI, to demonstrate the architecture's robustness. This result is discussed in Section 3.4.3.

Figure 21 Structure of PDZ domain (PDB ID: 1BE9) (Doyle et al., 1996). Ligands are known to bind to a surface groove formed in the α2-β2 groove. Residues in that groove can determine ligand affinity and enable recognition of specific amino acid sequences of its binding partner. Blue segments in the three-dimensional structure represent annotated binding sites.

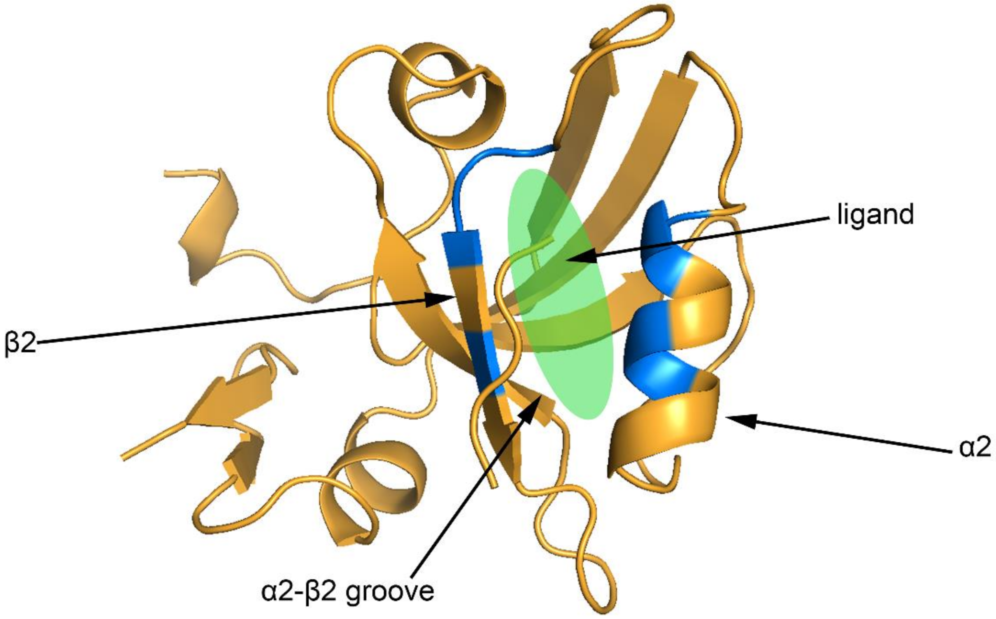
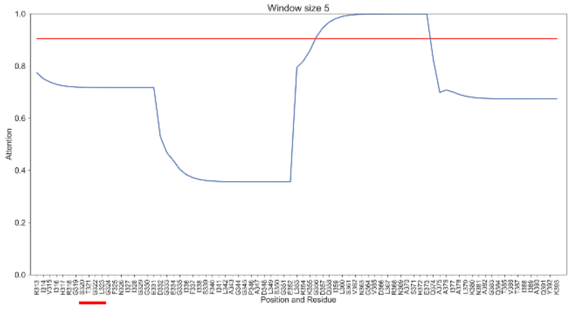
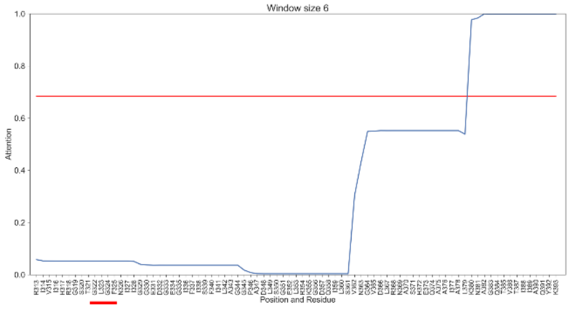
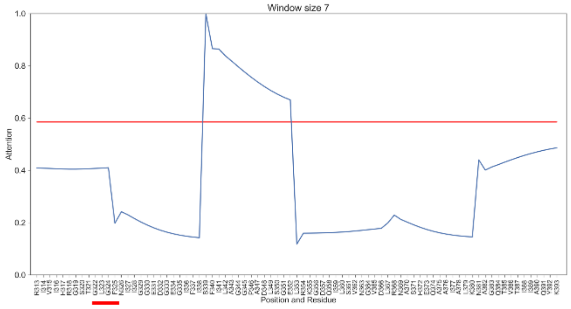
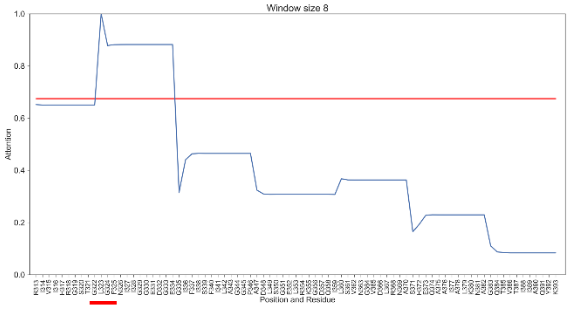

ligand
β2
α2
α2-β2 groove

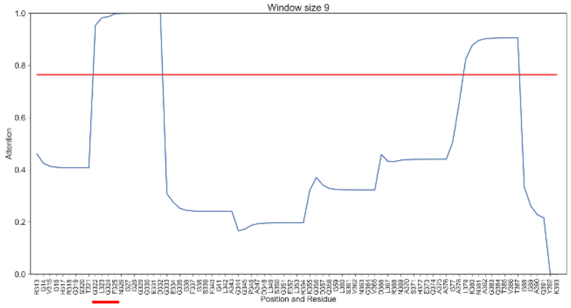
Figure 22: HANprot results for the window method, using windows of size 5 through 12 and an overlap size of 1. Different features are identified by the network throughout the different window sizes. Window size 9 achieved the highest AUC score (0.799), with the GLGF domain (red bar under axis) receiving the highest attention scores. Red line indicates attention threshold, given by the values above one standard deviation of the mean of the sequence’s attention score (blue).

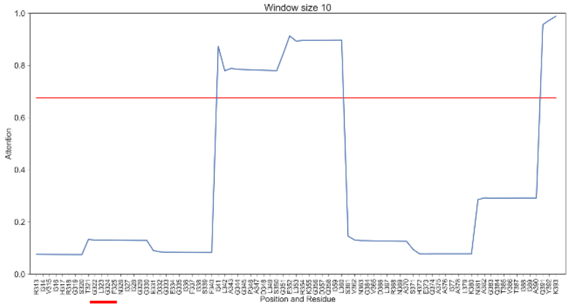

Window size 5
Attention
1.0
0.8
0.6
0.4
0.2
0.0
Position and Residue

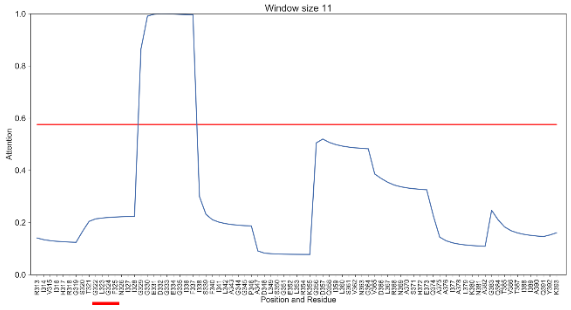

Window size 6
Attention
1.0
0.8
0.6
0.4
0.2
0.0
Position and Residue


Window size 7
Attention
1.0
0.8
0.6
0.4
0.2
0.0
Position and Residue


Window size 8
Attention
1.0
0.8
0.6
0.4
0.2
0.0
Position and Residue


Window size 9
Attention
1.0
0.8
0.6
0.4
0.2
0.0
Position and Residue


Window size 10
Attention
1.0
0.8
0.6
0.4
0.2
0.0
Position and Residue


Window size 11
Attention
1.0
0.8
0.6
0.4
0.2
0.0
Position and Residue

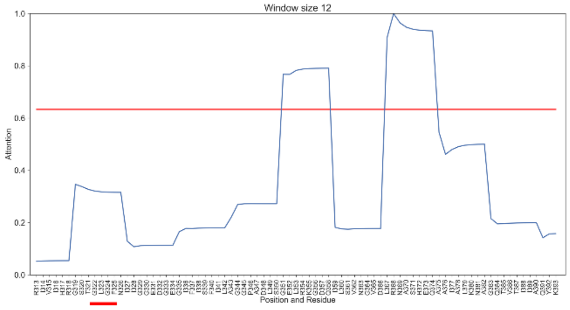
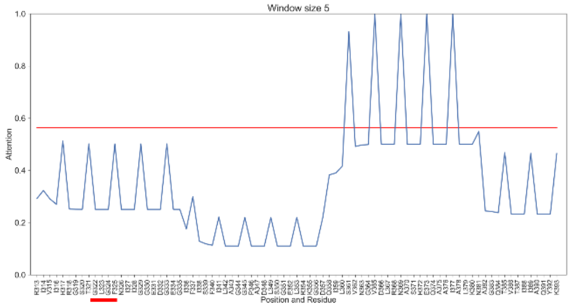
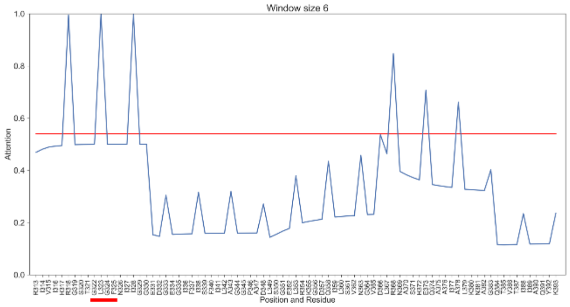
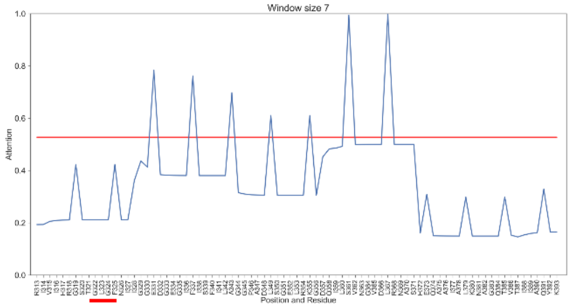
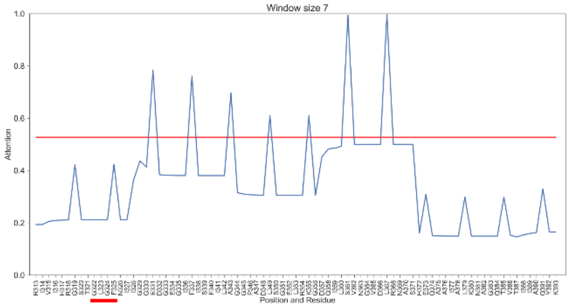

Window size 12
Attention
1.0
0.8
0.6
0.4
0.2
0.0
Position and Residue

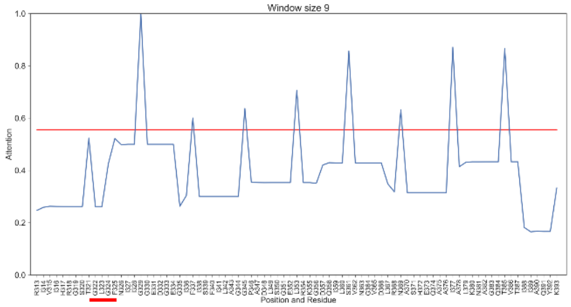
Figure 23: HANprot results for the window method, using windows of size 5 through 12 and an overlap size of 2. Different features are identified by the network throughout the different window sizes. Window size 12 achieved the highest AUC score (0.632), with the GLGF domain (red bar under axis) receiving the second highest attention scores. The highest attention scores are obtained by the second set of relevant residues, located in α2. Red line indicates attention threshold, given by the values above one standard deviation of the mean of the sequence's attention score (blue).

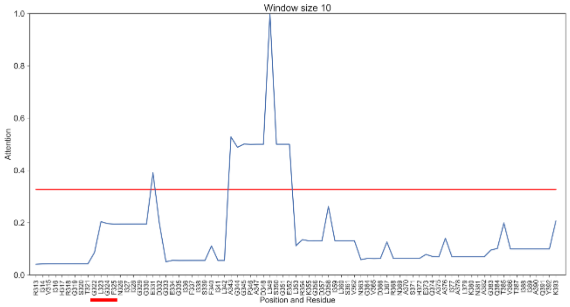

Window size 5
Attention
1.0
0.8
0.6
0.4
0.2
0.0
Position and Residue

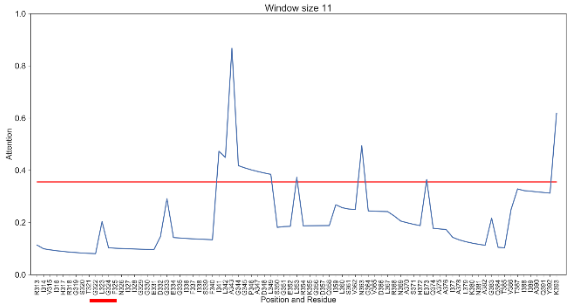

Window size 6
Attention
1.0
0.8
0.6
0.4
0.2
0.0
Position and Residue


Window size 7
Attention
1.0
0.8
0.6
0.4
0.2
0.0
Position and Residue


Window size 7
Attention
1.0
0.8
0.6
0.4
0.2
0.0
Position and Residue


Window size 9
Attention
1.0
0.8
0.6
0.4
0.2
0.0
Position and Residue


Window size 10
Attention
1.0
0.8
0.6
0.4
0.2
0.0
Position and Residue


Window size 11
Attention
1.0
0.8
0.6
0.4
0.2
0.0
Position and Residue

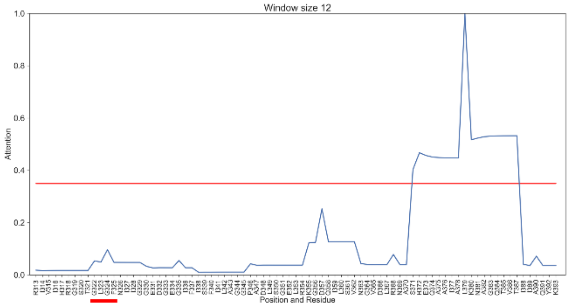

Window size 12
Attention
1.0
0.8
0.6
0.4
0.2
0.0
Position and Residue

Figure 24: HANprot results for the window method, using windows of size 5 through 12 and an overlap size of 3. Different features are identified by the network throughout the different window sizes. The highest attention scores are obtained by the second set of relevant residues located in the α2. Red line indicates attention threshold, given by the values above one standard deviation of the mean of the sequence’s attention score (blue).

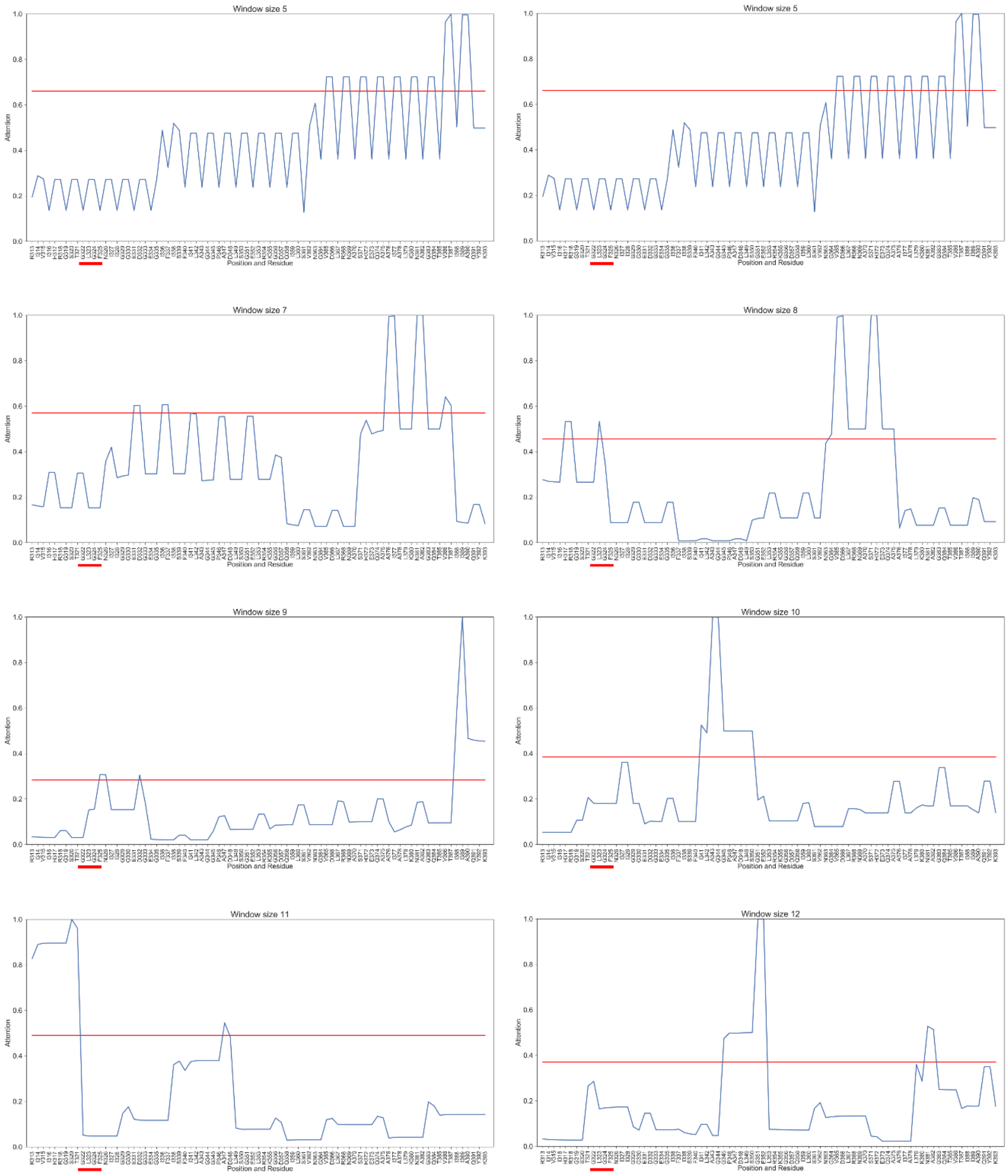
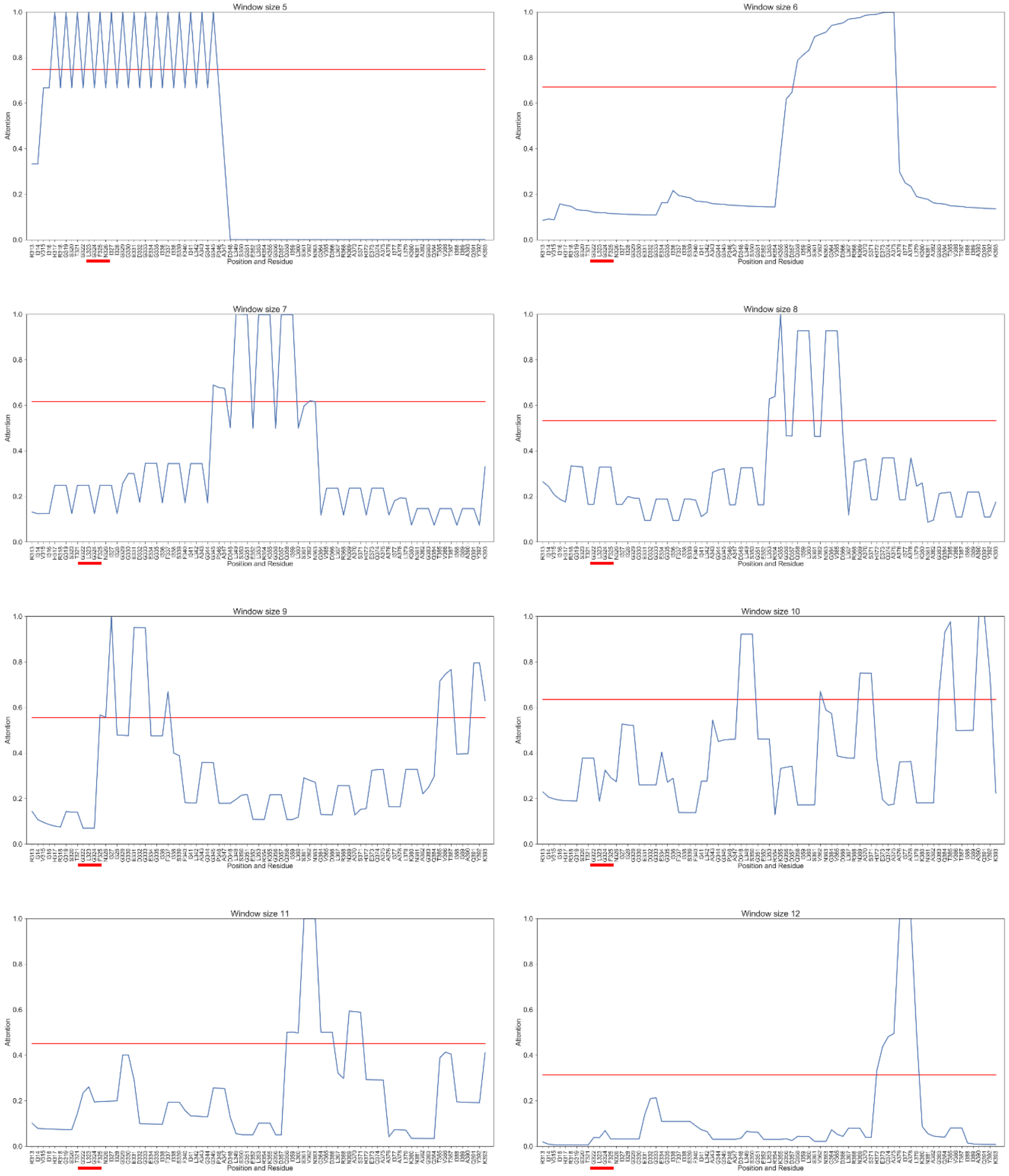

Window size 5
Window size 5
Window size 7
Window size 8
Window size 9
Window size 10
Window size 11
Window size 12
Attention
Position and Residue
0.0
0.2
0.4
0.6
0.8
1.0

Figure 25: HANprot results for the window method, using windows of size 5 through 12 and an overlap size of 4. Different features are identified by the network throughout the different window sizes. In this case, window size 12 achieved the highest AUC score (0.631). Red line indicates attention threshold, given by the values above one standard deviation of the mean of the sequence’s attention score (blue).


Window size 5

Window size 6

Window size 7

Window size 8

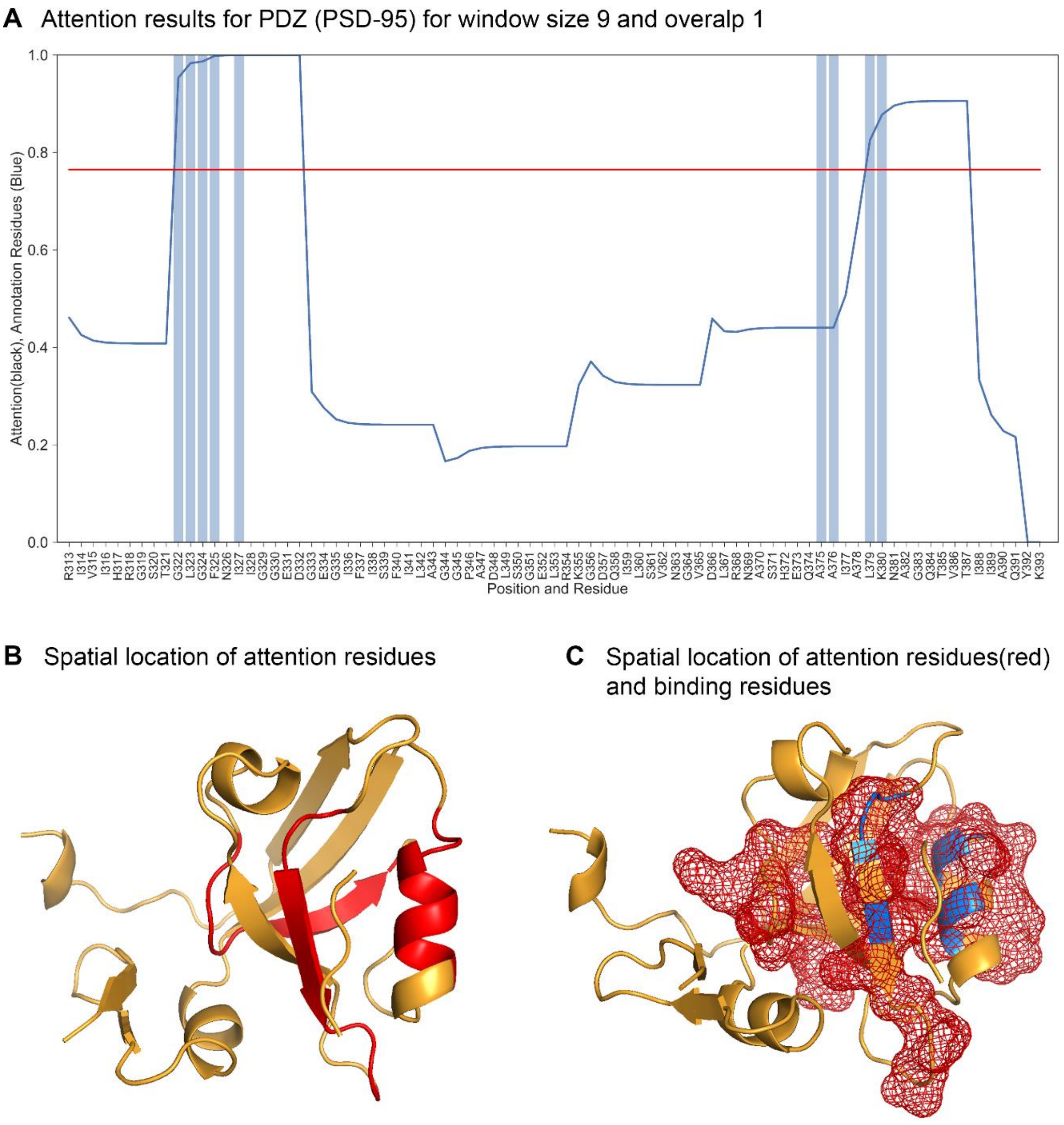
Window size 9

Window size 10

Window size 11

Window size 12

Figure 26: HANprot results for PDZ, using a window size of 9 and overlap of 1. (A) Attention scores for each residue are plotted (blue line). Blue bars represent annotated binding residues. Red line indicates attention threshold, given by the values above one standard deviation of the mean of the sequence's attention score (blue line). (B) In the 1BE9 PDZ structure (yellow), HANprot attention residues are shown in red. (C) Red mesh = attention residues, blue residues = annotated binding sites.



A Attention results for PDZ (PSD-95) for window size 9 and overalp 1

B Spatial location of attention residues

C Spatial location of attention residues(red) and binding residues

3.4.2 PDZ, trained using the window input

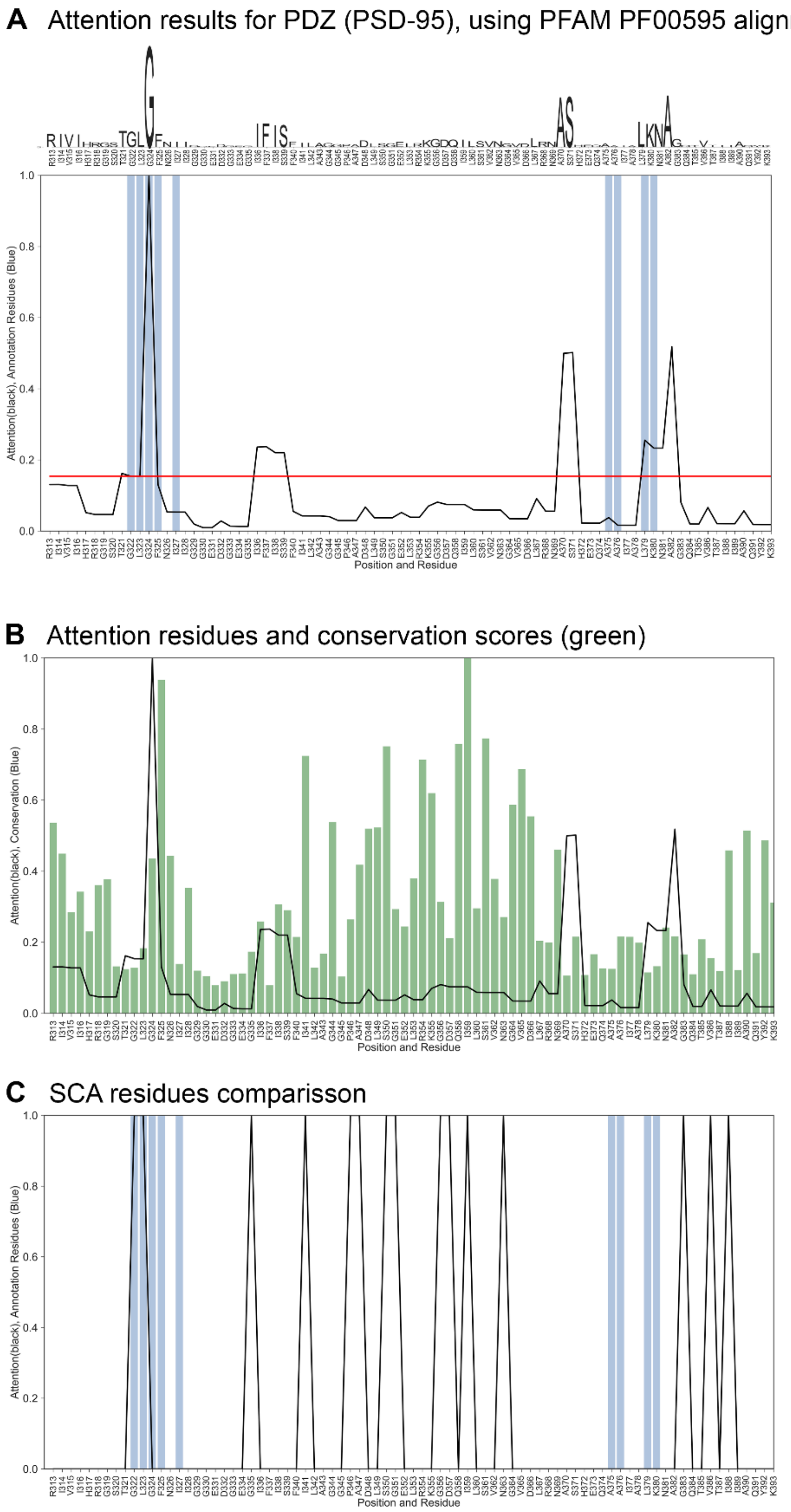
HANprot was trained using 8 configurations of windows (Figure 17B) and 3 different overlaps. Figures 22, 23, 24 and 25 show the normalized attention values for overlaps of size 1,2,3,4 (respectively) under different window lengths. Attention threshold is indicated by red line, and given by the values above one standard deviation of the mean of the sequence's attention score (blue). Appendix H: Table 7 shows the AUC and F1 scores computed for each condition. The highest AUC score is obtained by window of size 9 with an overlap of size 1 (Figure 22). Figure 26 shows HANprot results for PDZ (1BE9) displayed in structure for windows size 9 and overlaps 1, which achieved the highest AUC score for the window trainings (0.799). HANprot identifies residues (marked in red in Figure 26B and as a red mesh in 26C) in the carboxylate binding site (Figure 26C, blue residues in α-helix), among other residues in the vicinity. In addition, HANprot identifies other binding sites (Figure 26C, blue in β-sheet) in the groove where the binding site is located. However, for overlaps of size 2 and 4, windows of size 12 achieve higher scores (Figure 23 and 25). Any other larger window size seldom obtained higher scores than the other sizes. Smaller window with different overlaps size, in particular with overlaps of 2 and 3, achieve high AUC scores. However, we note that there is a lack of discernible pattern across the different combinations of windows and overlaps for AUC scores lower than 0.55. This can serve as an indicator for future analysis, when little information about the protein's functional residues is available.

The two highest AUC scores are with window size of 9 and 8 with scores of 0.799 and 0.667, respectively. Smaller windows can capture local interactions more accurately, whereas long-range interactions, although present, and attenuated when windows of larger size are used. This indicates that HANprot captures protein interactions over small and medium alignment segments better than over long segments, when segments of the protein sequence are used as input in the window method. However, when the full sequence is used as input, which effectively compares to a window of size L, HANprot identifies important residues, achieving AUC scores of over 0.65, as discussed in the following section. This suggests an optimal window size through which both short-range or long-range interactions are captured by the network, which we expect to investigate in future works.

3.4.3 PDZ, Full sequence input

When trained in the full sequence instead of windows, HANprot achieves ROC of 0.715 for NCBI alignments and 0.660 for PFAM (Figure 27 and Figure 28). SCA obtains a ROC of 0.520 in comparison. As highlighted previously, HANprot can not only determine important residues for the protein's function, but it assigns each residue a score (0, 1) based on the attention mechanism. For example, the highest attention score for PDZ is given to a Glycine (G), known to be the most conserved G in that motif in the GLGF motif (Lee & Zheng, 2010). Figure 27A and 28A shows the normalized attention over the whole PDZ sequence, followed by its comparison with annotated binding sites for PFAM and NCBI PDZ MSA inputs, respectively. Figure 27B and 28B, show attention scores and conservation scores (calculated according to Halabi *et al.*), for PFAM and NCBI PDZ MSA inputs, respectively (Halabi *et al.*, 2009). Figure 27C shows residues identified through the SCA method for PDZ. Figure 29A shows the pairwise distance of residues in PDZ, with attention residues superimposed, and Figure 29B shows the three-dimensional structure and the attention residues for PDZ MSA sourced from PFAM. Similarly, Figure 30 shows these same results for PDZ MSA sourced from NCBI.

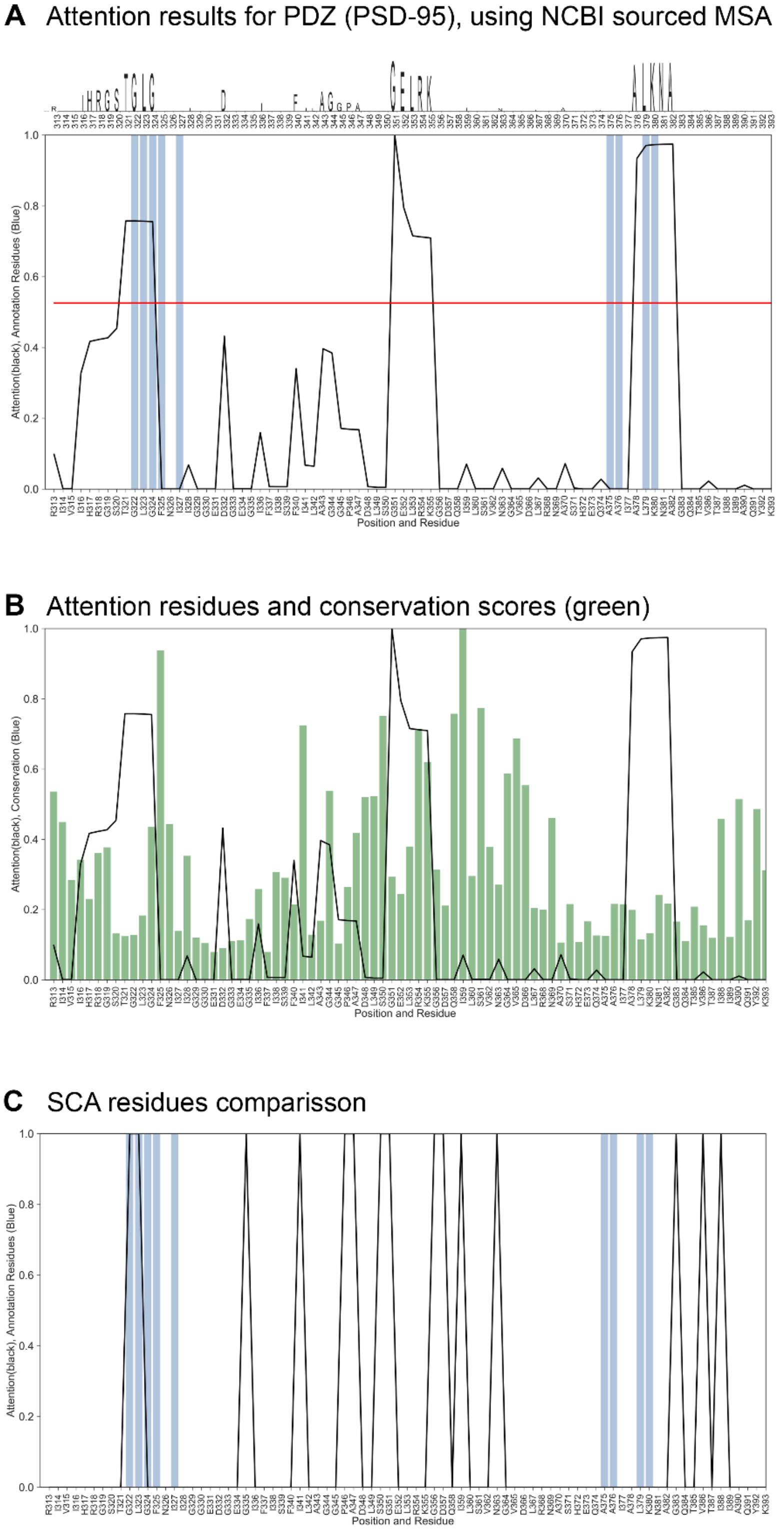
Figure 27: PDZ attention residues for full sequence inputs (PFAM alignment). (A) In logo form, attention levels for each residue are shown in black. Attention scores are plotted (black) with the blue bars represent annotated binding sites. Red line indicates attention threshold, given by the values above one standard deviation of the mean of the sequence's attention score (black line). (B) Attention scores versus conservation scores (Halabi *et al.*, 2009). (C) Residues identified through the SCA method are shown in black and annotated binding sites in blue.


A Attention results for PDZ (PSD-95), using PFAM PF00595 alignment

B Attention residues and conservation scores (green)

C SCA residues comparisson

Figure 28: PDZ attention residues for full sequence inputs (NCBI alignment). (A) In logo form, attention levels for each residue are shown in black. Attention levels are plotted (black) with the blue bars represent annotated binding sites. Red line indicates attention threshold, given by the values above one standard deviation of the mean of the sequence's attention score (black line). (B) Attention levels versus conservation scores (Halabi *et al.*, 2009). (C) Residues identified through the SCA method are shown in black and annotated binding sites in blue.



A Attention results for PDZ (PSD-95), using NCBI sourced MSA

B Attention residues and conservation scores (green)

C SCA residues comparisson

Figure 29: Structural details for PDZ attention residues for full sequence inputs (PFAM alignment). (A) Inter-residue distance versus attention residues. Note that attention residues have varying inter-residue distances, indicating possible short and long-range interactions. (B) Three-dimensional visualization of attention residues. Red segments/mesh = attention residues, blue segments = annotated binding sites.

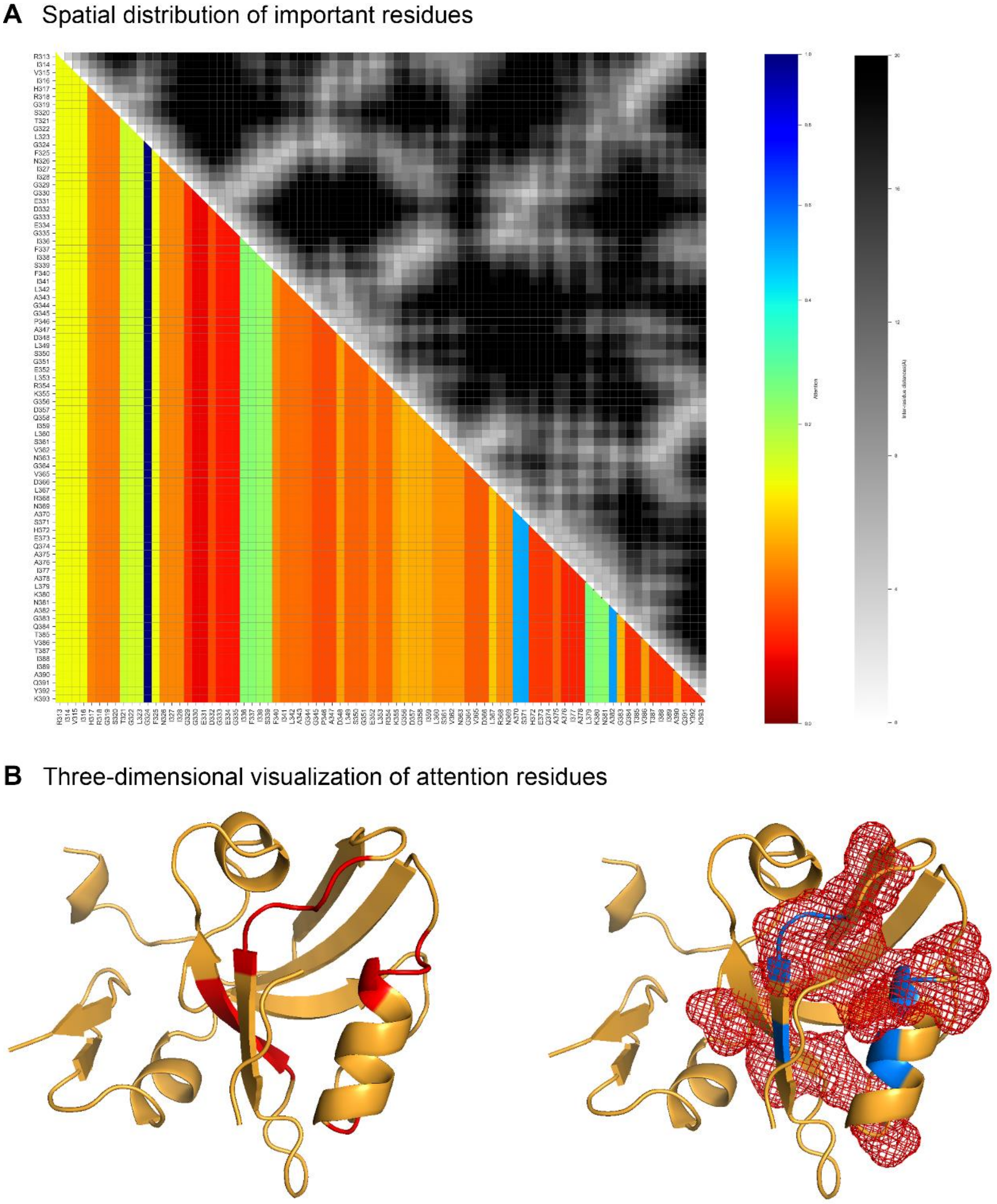

A Spatial distribution of important residues
R313
I314
V315
I316
H317
R318
G319
S320
T321
G322
L323
G324
F325
N326
I327
I328
G329
G330
E331
D332
G333
E334
G335
I336
F337
I338
S339
F340
I341
L342
A343
G344
G345
P346
A347
D348
L349
S350
G351
E352
L353
R354
K355
G356
D357
Q358
I359
L360
S361
V362
N363
G364
V365
D366
L367
R368
N369
A370
S371
H372
E373
Q374
A375
A376
I377
A378
L379
K380
N381
A382
G383
Q384
T385
V386
T387
I388
I389
A390
Q391
Y392
K393
1.0
0.8
0.6
0.4
0.2
0.0
Attention
20
16
12
8
4
0
Inter-residue distance(Å)
B Three-dimensional visualization of attention residues

Figure 30: Structural details for PDZ attention residues for full sequence inputs (NCBI alignment). (A) Inter-residue distance versus attention residues. Note that attention residues have varying inter-residue distances, indicating possible short and long-range interactions. High attention values (dark blue) display short to mid inter-residue distances. This is further confirmed by the proximity displayed by those residue segments in the structural figure that follows. (B) Three-dimensional visualization of attention residues. Red segments/mesh = attention residues, blue segments = annotated binding sites.

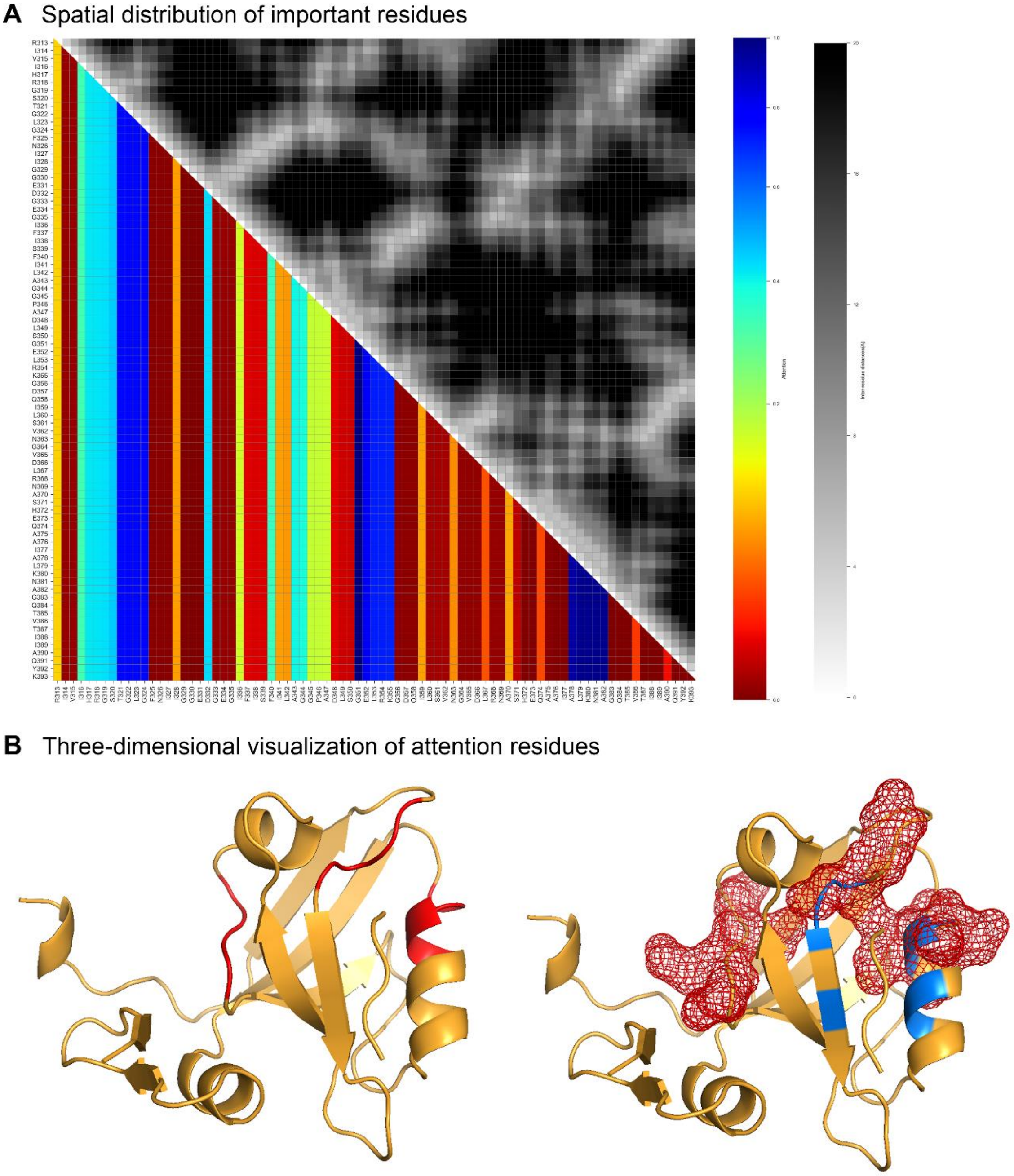
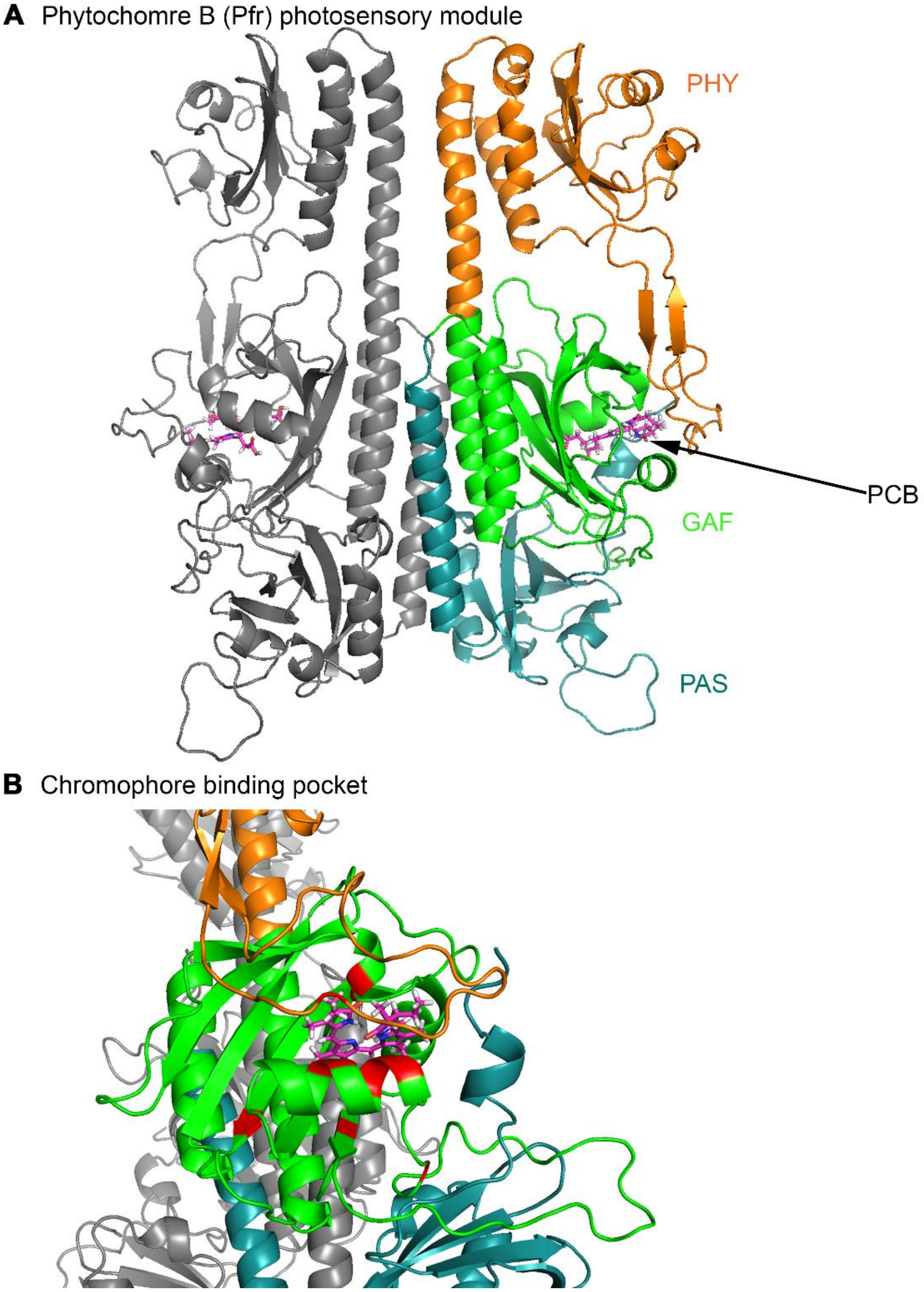
A
Spatial distribution of important residues
R313 I314 V315 I316 H317 R318 G319 S320 T321 G322 L323 G324 F325 N326 I327 I328 G329 G330 E331 D332 G333 E334 G335 I336 F337 I338 S339 F340 I341 L342 A343 G344 G345 P346 A347 D348 L349 S350 G351 E352 L353 R354 K355 G356 D357 Q358 I359 L360 S361 V362 N363 G364 V365 D366 L367 R368 N369 A370 S371 H372 E373 Q374 A375 A376 I377 A378 L379 K380 N381 A382 G383 Q384 T385 V386 T387 I388 I389 A390 Q391 Y392 K393
Attention
1.0
0.8
0.6
0.4
0.2
0.0
Inter-residue distance(Å)
20
16
12
8
4
0
B
Three-dimensional visualization of attention residues

Figure 31: Structure of *Arabidopsis thaliana* phytochrome B photosensory module (PDB ID 4OUR) (Burgie et al., 2014). (A) Ribbon diagrams of PHY (orange), PAS (cyan) and GAF (green) domains and the chromophore (PCB) (in pink, indicated by arrow). (B) Closeup of the chromophore binding pocket, with interacting residues in red, according to Burgie *et al.* (Burgie et al., 2014; Burgie & Vierstra, 2014).



A Phytochomre B (Pfr) photosensory module

B Chromophore binding pocket

3.4.4 PhyB

Phytochromes are dimeric chromoproteins, composed of two polypeptides (each ~125-kD) each carrying a covalently linked tetrapyrrole chromophore in the N-terminal domain (Quail *et al.*, 1995; Sakamoto & Nagatani, 1996). When bound to a chromophore, these signaling proteins undergo photoisomerization upon red and far-red illumination, a light induced molecular change that plants and algae use for “measuring” light (Cerdán & Chory, 2003; Yanovsky & Kay, 2003). The photoisomerization event is linked to an allosteric transition in the phytochrome between two spectrally distinct conformational states, called *Pr* (red absorbing), and *Pfr* (far-red-absorbing) (Li *et al.*, 2011; Rockwell *et al.*, 2006). Thus, light acts as a switch between these two forms, and the transition between the *Pr* and *Pfr* states is reversible upon sequential absorption of red (R) and far-red (FR) light (Cerdán & Chory, 2003; Yanovsky & Kay, 2003). Understanding more about how these proteins interact with light will enable us to design protein optical properties, creating unique opportunities for light controlled systems that have potential applications in medicine, imaging and synthetic biological systems including biofuel-producing species.

Plant phytochomres contain an N-terminal photosensing module (PSM) (Figure 31A), responsible for the dimerization and signal transduction (Burgie *et al.*, 2014; Burgie & Vierstra, 2014; Kikis, Oka, Hudson, Nagatani, & Quail, 2009). The PSM contains a PAS (Period/Arnt/Single-Minded) domain, a GAF (cGMP phosphodiesterase/adenylyl cyclase/FhIA) domain where the chromophore binding pocket is located (Figure 31B), and a PHY (Phy-specific) domain, also involved in photoisomerization (Burgie *et al.*, 2014). Several groups have shown that the GAF domain forms most of the chromophore binding pocket in PhyB (Kikis *et al.*, 2009; Velázquez Escobar *et al.*, 2017; von Horsten *et al.*, 2016), through a covalent bond and several hydrogen bonds (Burgie *et al.*, 2014). The hydrogen bonds promote stabilization in the *Pr* form or red light absorbing form, are several arginine and histidine residues, among others (e.g., using 4OUR numbering: R252, R222, J257, Y261, D207, R482, H303) (Burgie *et al.*, 2014; Burgie & Vierstra, 2014).

Mutational analyses with PhyB confirmed the importance of several other residues around the binding pocket. Studies have also shown that mutations to residues in that pocket are disruptive to chromophore binding, even if the chromophore attachment site (C256) is not modified. Furthermore, mutations to the binding pocket are known to not only disrupt chromophore binding but photoreversibility, stability, and disruptions in the PIF3 binding (Kikis *et al.*, 2009). Although there are extensive publications relating the residues involved in chromophore binding, only annotated binding residues from NCBI, RCSB and UniProt were considered for validation of HANprot.

HANprot identifies the amino acids that make up the chromophore binding pocket and assigns it the highest attention for the sequence, including amino acids that are far apart in their primary structure (Figure 32A, also showing a zoom into the segment with highest attention scores). SCA results for PhyB are largely nonspecific (Figure 32C) and obtains a AUC score of 0.5620 compared to HANprot’s score of 0.778. Figure 33 shows the results for PhyB when HANprot is trained using full sequences. Figure 33A shows the shows the three-dimensional structural distances between the attention residues (red), and the spatial distribution of those residues over the chromophore binding site in the PhyB homology model-generated structure. In addition, in Figure 33B, the segment with the second highest level of attention is shown in red. The residues in this segment span the junction between the PAS and GAF domains, which is involved in PIF binding and other *Pfr*-specific interactions (Burgie & Vierstra, 2014; Kikis *et al.*, 2009). Figure 34 shows the results for PhyB when HANprot is trained in the windows method.

3.4.5 Effect of Dataset quality on residue identification

There is a noticeable the variability in the AUC scores for the different protein families. This is expected since not every protein family has the same level of diversity or conservation due to evolutionary constraints that have been imposed to the protein. In addition, the length and number of sequences available for each protein family will impact the learning accuracy of the deep learning algorithm. I hypothesized that the dataset's quality (number of sequences available, number of gaps, etc.) affected the conservation scores as calculated by the Henikoff weights. In turn, I expected variation in those scores to impact the training of the dataset. Indeed, when I compared the AUC scores and the mean HSW, as a measure of the dataset's quality. High HSWs are associated with small proteins (short sequences), or limited datasets (small number of sequences) (Appendix I).

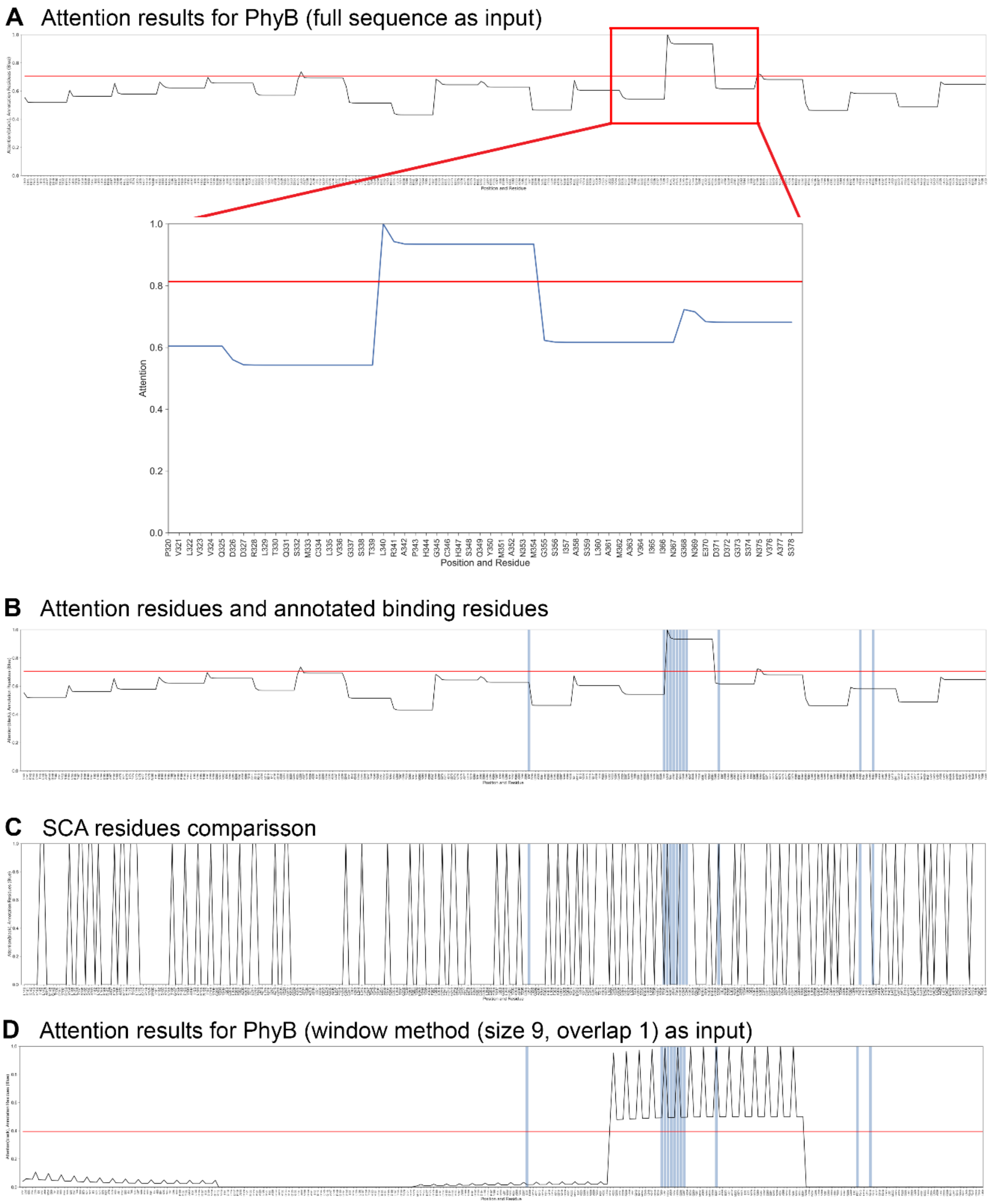
3.5 Summary

The biological properties displayed by proteins can arise from interactions among amino acid residues and from the basic chemical properties of polypeptide chains, but new methods to identify relevant residues are still lacking. Here, I show that by applying principles of natural language processing machine learning methodologies and sequence conservation, HANprot can identify positions that control different properties of a protein's function, and scores better than previous methods. However, significant technical challenges remain for proper identification and ranking of those residues. The novelty of my proposal will require validation at the intersection of many approaches: a combination of molecular dynamics, machine learning, and biological experimentation.

Limitations in the known properties or functional importance of other residues identified in my method highlight the need for further functional studies and more detailed database annotations. Nevertheless, visualization of the attention vectors generated by HANprot illustrates its effectiveness in identifying important residues or short segments of residues. And, the fact that attention residues identified correspond to important functional properties of the proteins discussed provides strong support for HANProts' biological relevance. As pointed out by Halibi *et al.*, residues identified by SCA and similar methods that compare sequences throughout evolution often form sparse, physically connected and functionally independent groupings. This was evident in the three-dimensional structures presented in this work. More poignantly, several studies have also pointed out that amino acids contribute unequally, but still cooperatively to a protein's structure and function (Halabi *et al.*, 2009). This highlights the importance of methods like HANprot, were relevance of a residue can be predicted.

As such, I argue that HANprot represents the first step towards identifying important residues and their level of contribution using a deep learning approach. Moreover, my results suggest that deep learning methodologies and architectures can be translated for use in proteomics, and to identify a pattern of functional residues in a protein sequence. The automated identification of important residues or sectors in a protein sequence provides a basis to direct further experiments or as a scientific/experimental gateway to explore perturbation analyses, drug response, protein mechanisms and protein engineering, among other possibilities.

Figure 32: HANprot results for PhyB. (A) Attention levels for each residue are shown in black, and a focus in a segment with the highest attention is shown. This segment contains the chromophore binding pocket. (B) Attention levels are plotted (black) with the blue bars represent annotated binding sites. (C) SCA-determined relevant residues. (D) For window method results, attention levels for each residue are shown in black, with the blue bars represent annotated binding sites. Red line indicates attention threshold, given by the values above one standard deviation of the mean of the sequence's attention score (blue lines in (A) and lack lines in (B) and (D) plots).



A Attention results for PhyB (full sequence as input)

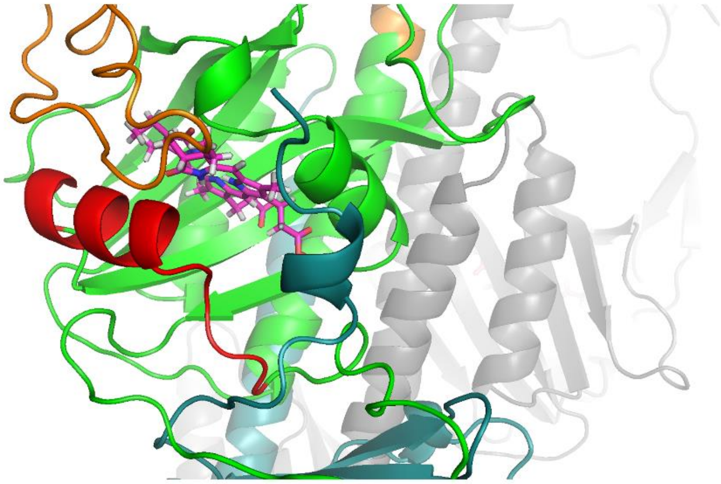
B Attention residues and annotated binding residues

C SCA residues comparisson

D Attention results for PhyB (window method (size 9, overlap 1) as input)

Figure 33: Three-dimensional localization of attention residues. (A) Red residues correspond to those in Figure 32B with highest attention scores. (B) Red residues correspond to a segment of residues that obtained second highest attention scores. In Figure 32B, the segment is circled in green.

A Attention residues (from segment with highest attention scores)



B Attention residues (from segment with second highest attention scores)

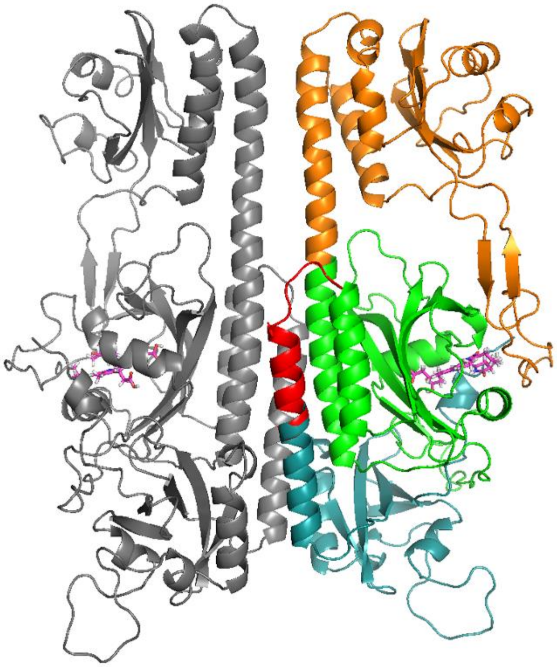
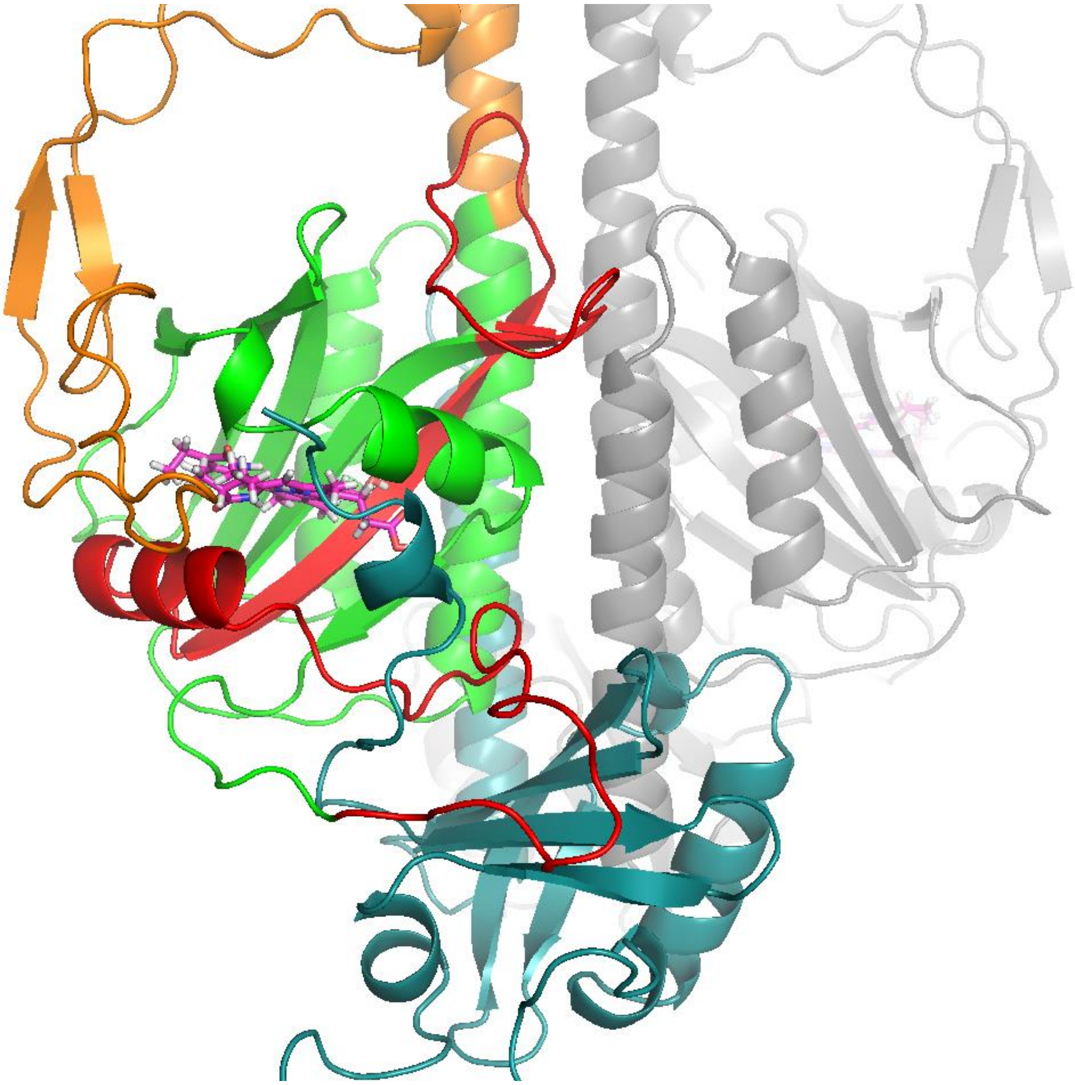


Figure 34: Three-dimensional localization of attention residues, when HANprot is trained in the window method. Red residues correspond to those in Figure 32D with highest attention scores.



The results of this study can guide efforts to modify, disable or disrupt a protein's function. For example, for PhyB, the results shown introduced possible sites for customizing the protein's photoswitchable properties. While we already knew the binding pocket from the annotations and structure, the results from HANprot highlight that this tool is useful for identifying important sites when structures are not available. The integration of these techniques holds great promise, but also brings forth new challenges that must be met for this platform to realize its full potential. Regardless, utilizing different types sequence lengths and segments, my method performs relatively well in both cases. Even though SCA reaches high, albeit in average, lower scores than HANprot, I suspect it is in part due to the large number of residues it identifies, increasing the chances of a positive match.

For future work, exploration of different architectures and network parameters will be needed. In addition, it is important to validate this work through experiments and through the training of larger protein sequences.

Chapter 3, in part, is currently being prepared for submission for publication of the material by Catanho, Marianne; Gao, Shang; Kyriakakis, Phillip; Coleman, Todd P.; Ramanathan, Arvind. "Discovering sequence coevolution signatures with hierarchical attention networks". The dissertation author is the primary investigator and author of this material.

APPENDICES

Appendix A: Table 1 - Plasmids used in this work

Genes for enzymes were synthesized by Genscript and Integrated DNA Technologies.

Plasmids and sequences will be made available on Addgene or upon request.

Plasmid Number	Description	Source	Addgene Plasmid ID
pMZ-802	FLuc under control of pTet (tetO13-CMVmin-FLuc-pA)	Müller *et al.*	N/A
pPKm-102	pcDNA3 - mOrange	This study	90493
pPKm-105	pcDNA3 - PhyB NT - GBD,	This study	104853
pPKm-112	pcDNA3 - MTAD - PIF3,	This study	90494
pPKm-113	pcDNA3 - MTAD - PIF6,	This study	90495
pPKm-118	pcDNA3 - 5X UAS - pFR Luciferase	This study	90491
pPKm-145	Empty plasmid, pSIN-EF1-alpha-IRES-puro	This study	90505
pPKm-163	pcDNA3 - PIF3 - GBD,	This study	104854
pPKm-195	pcDNA3 - PhyB NT - MTAD	This study	90496
pPKm-196	pcDNA3 - PIF6-DBD	This study	90511
pPKm-202	pcDNA3 – CMVmin 5X UAS - pFR - Luciferase	This study	90492
pPKm-226	pcDNA3 - PIF3 – VPR	This study	90497

Appendix A: Table 1 - Plasmids used in this work, Continued.

pPKm-227	pcDNA3 - VPR - PIF3	This study	90498
pPKm-230	pSIN - EF1-alpha - PIF3 - MTAD - IRES - PhyB - GBD	This study	90499
pPKm-231	pSIN - EF1-alpha - MTS - tFd - P2A - MTS - tFNR, encoding for mitochondrial-tagged *Thermosynechococcus elongatus* Ferredoxin (Fd) and Ferredoxin-NADP(+) oxi0doreductase (FNR)	This study	90500
pPKm-232	pSIN - EF1-alpha - MTS tHO1 - P2A - MTS - tPCYA, encoding for mitochondrial-tagged *Thermosynechococcus elongatus* Heme Oxygenase-1 (HO1) and phycocyanobilin:ferredoxin oxidoreductase (PcyA)	This study	90501
pPKm-233	pSIN - EF1-alpha - sFD - P2A - MTS - sFNR, encoding for *Synechococcus sp.* Ferredoxin (Fd) and Ferredoxin-oxidoreductase (FNR)	This study	90508
pPKm-234	pSIN - EF1-alpha - MTS sHO1 - P2A - MTS - sPCYA, encoding for mitochondrial-tagged *Synechococcus sp.* Heme Oxygenase (HO1) and phycocyanobilin:ferredoxin oxidoreductase (PcyA),	This study	90507
pPKm-235	pSIN - EF-1alpha - MTS sHO1 - P2A - MTS - sPCYA, encoding for mitochondrial-tagged *Synechococcus sp.* Heme Oxygenase-1 (HO1) and *Arabidopsis thaliana* phytochromobilin:ferredoxin oxidoreductase (Hy2) replacing the chloroplastic targeting sequence with a MTS	This study	90509

Appendix A: Table 1 - Plasmids used in this work, Continued.

pPKm-240	pSIN - EF1-alpha cyto-sHO1-P2A – cyto-sPcyA, encoding for cytoplasmic-tagged *Synechococcus sp* HO1 and PcyA	This study	90510
pPKm-241	pSIN - EF1-alpha - cyto-sFd - P2A - cyto-sFNR, vector encoding for cytoplasmic-tagged *Synechococcus sp* Fd and FNR	This study	104855
pPKm-243	pSIN - EF1-alpha - mOrange-P2A-mitosfGFP, mOrange and mitochondrial-tagged sfGFP	This study	90506
pPKm-244	pSIN – EF1-alpha - MTS - tHO1 - P2A - MTS - tPCYA - IRES - MTS - tFD - P2A - MTS - tFNR	This study	90502
pPKm-245	pSIN - EF1-alpha - MTS - tHO1 - P2A - MTS - tPCYA - P2A - MTS - tFD - P2A - MTS - tFNR	This study	90503
pPKm-248	pSIN - EF1-alpha - MTS - tPCYA - IRES - MTS - tHO1 - P2A - MTS - tFD - P2A - MTS - tFNR	This study	90504
pPKm-292	pcDNA3 – GAL4_DNA BD -MTAD	This study	105816
pPKm-293	pcDNA3 – TET DNA BD -MTAD	This study	105817
pPKm-300	pSIN - EF1-alpha - MTS - tFd, encoding for mitochondrial-tagged *Thermosynechococcus elongatus* Ferredoxin (Fd)	This study	104626
pRL-TK	Control reporter for constitutive expression of wildtype Renilla luciferase (Rluc) under pRL-TK	Promega	E2241

Appendix B: Table 2 - Transfection and illumination details for each figure

Each experiment described in this work was transfected according to the following table. To the best of our ability, ratios and concentrations were kept identical for comparable experiments.

Figure 2	HEK293 cells were transfected 24 hours after plating. Calculations are for each well. Transfected in a 6 well plate. Cells were harvested 44 hours post-transfection followed by Immunoprecipitation and Zn-PAGE as described in methods.

NE control

Plasmid	DNA mass (ng)	DNA Ratio
pPKm-105	125	1/20
pPKm-102	125	1/20
pPKm-145	1125	18/20

M2-sPcyA

Plasmid	DNA mass (ng)	DNA Ratio
pPKm-105	125	1/20
pPKm-243	125	1/20
pPKm-234	1125	9/20
pPKm-145	1125	9/20

M4-sPcyA

Plasmid	DNA mass (ng)	DNA Ratio
pPKm-105	125	1/20
pPKm-243	125	1/20
pPKm-234	1125	9/20
pPKm-233	1125	9/20

M2-tPcyA

Plasmid	DNA mass (ng)	DNA Ratio
pPKm-105	125	1/20
pPKm-243	125	1/20
pPKm-232	1125	9/20
pPKm-145	1125	9/20

M4-tPcyA

Plasmid	DNA mass (ng)	DNA Ratio
pPKm-105	125	1/20
pPKm-243	125	1/20
pPKm-232	1125	9/20

Appendix B: Table 2 - Transfection and illumination details for each figure, Continued.

	pPKm-231	1125	9/20
M2-Hy2	Plasmid	DNA mass (ng)	DNA Ratio
	pPKm-105	125	1/20
	pPKm-243	125	1/20
	pPKm-235	1125	9/20
	pPKm-145	1125	9/20
M4-Hy2	Plasmid	DNA mass (ng)	DNA Ratio
	pPKm-105	125	1/20
	pPKm-243	125	1/20
	pPKm-235	1125	9/20
	pPKm-233	1125	9/20
Figure 3A	HEK293 cells were transfected 24 hours after plating. Calculations are for each well. Transfected two of each in a 6 well plate, one with and one without heme. 10µM (Frontier scientific) was added 18 hours and 43 hours post-transfection. Cells were harvested 44 hours post transfection followed by Immunoprecipitation and Zn-PAGE as described in methods.		
NE control	Plasmid	DNA mass (ng)	DNA Ratio
	pPKm-105	125	1/20
	pPKm-243	125	1/20
	pPKm-145	1125	18/20
C2	Plasmid	DNA mass (ng)	DNA Ratio
	pPKm-105	125	1/20
	pPKm-243	125	1/20
	pPKm-240	1125	9/20
	pPKm-145	1125	9/20
C4	Plasmid	DNA mass (ng)	DNA Ratio
	pPKm-105	125	1/20
	pPKm-243	125	1/20
	pPKm-240	1125	9/20
	pPKm-241	1125	9/20

Appendix B: Table 2 - Transfection and illumination details for each figure, Continued.

M2

Plasmid	DNA mass (ng)	DNA Ratio
pPKm-105	125	1/20
pPKm-243	125	1/20
pPKm-234	1125	9/20
pPKm-145	1125	9/20

M4

Plasmid	DNA mass (ng)	DNA Ratio
pPKm-105	125	1/20
pPKm-243	125	1/20
pPKm-234	1125	9/20
pPKm-233	1125	9/20

Figure 3B

HEK293 cells were transfected 24 hours after plating. Calculations are for each well in a 6-well plate. Cells were harvested 44 hours post transfection followed by Immunoprecipitation and Zn-PAGE as described in methods.

M2

Plasmid	DNA mass (ng)	DNA ratio
pPKm-105	125	1/20
pPKm-243	125	1/20
pPKm-232	1125	9/20
pPKm-145	1125	9/20

M3

Plasmid	DNA mass (ng)	DNA ratio
pPKm-105	125	1/20
pPKm-243	125	1/20
pPKm-232	1125	9/20
pPKm-300	1125	9/20

M4

Plasmid	DNA mass (ng)	DNA ratio
pPKm-105	125	1/20
pPKm-243	125	1/20
pPKm-232	1125	9/20

Appendix B: Table 2 - Transfection and illumination details for each figure, Continued.

	pPKm-231	1125	9/20
NE	Plasmid	DNA mass (ng)	DNA ratio
	pPKm-105	125	1/20
	pPKm-243	125	1/20
	pPKm-145	2250	18/20
Figure 4	HEK293 cells were transfected 24h after plating, followed by a medium change 24h after transfection. For illumination, 1µmol/m2/s 1-minute pulses of red light were delivered for 24h, starting 12h after the medium change. Cells were kept in darkness before and after illumination. Lysis was performed 72h after transfection, and samples stored in -20C until assayed.		
	Plasmid	DNA mass (ng)	DNA ratio
9HP:9EV (1:1 ratio HP:EV)	pPKm-102	425.0	25.5/30
	pPKm-105	16.7	1/30
	pPKm-112	16.7	1/30
	pPKm-232	16.7	1/30
	pPKm-202	16.7	1/30
	pRL-TK	8.3	0.5/30
	Plasmid	DNA mass (ng)	DNA ratio
9HP:9FF (1:1 ratio HP:FF)	pPKm-102	408.3	24.5/30
	pPKm-105	16.7	1/30
	pPKm-112	16.7	1/30
	pPKm-232	16.7	1/30
	pPKm-231	16.7	1/30
	pPKm-202	16.7	1/30
	pRL-TK	8.3	0.5/30
	Plasmid	DNA mass (ng)	DNA ratio
17HP:1EV (17:1 ratio HP:EV)	pPKm-102	158.3	9.5/30
	pPKm-105	16.7	1/30
	pPKm-112	16.7	1/30
	pPKm-232	283.3	17/30
	pPKm-202	16.7	1/30
	pRL-TK	8.3	0.5/30

Appendix B: Table 2 - Transfection and illumination details for each figure, Continued.

17HP:1FF

(17:1 ratio HP:FF)

Plasmid	DNA mass (ng)	DNA ratio
pPKm-102	141.7	8.5/30
pPKm-105	16.7	1/30
pPKm-112	16.7	1/30
pPKm-232	283.3	17/30
pPKm-231	16.7	1/30
pPKm-202	16.7	1/30
pRL-TK	8.3	0.5/30

Figure 9

HEK293 Cells were transfected 24h after plating on polylysine-coated coverslips. 43 hours later media was changed with media+5µM PCB (Frontier Scientific P14137) added to the NE+PCB control. One hour later cells were rinsed in PBS and fixed in 4%Paraformaldehyde for 10 minutes. Next cells were incubated in permeabilization buffer (5% BSA + 0.3% TritonX-100 in PBS) for 30min, followed by primary antibodies overnight at 4°C in antibody buffer (2% BSA + 0.2% TritonX-100 in PBS; anti-flag mouse monoclonal 1:1000 (Sigma F3165) anti-HA rabbit polyclonal 1:500 (Santa Cruz Y-11)); Next coverslips were rinsed twice and washed three time in PBS and then incubated in antibody buffer with goat anti-mouse AlexaFluor 488 1:1000 (Thermo-Fisher A11001) goat anti-rabbit AlexaFluor 568 1:1000 (Thermo-Fisher A11011)). Coverslips were then mounted with Fluoromount-G (SouthernBiotech 0100-20). Images were taken using a DeltaVision RT Deconvolution Microscope.

NE control

Plasmid	DNA mass (ng)	DNA Ratio
pPKm-105	100	4/20
pPKm-145	400	16/20

C2

Plasmid	DNA mass (ng)	DNA Ratio
pPKm-105	100	4/20
pPKm-240	375	15/20
pPKm-145	25	1/20

C4

Plasmid	DNA mass (ng)	DNA Ratio
pPKm-105	100	4/20
pPKm-240	375	15/20

Appendix B: Table 2 - Transfection and illumination details for each figure, Continued.

pPKm-241	25	1/20

M2

Plasmid	DNA mass (ng)	DNA Ratio
pPKm-105	100	4/20
pPKm-234	375	15/20
pPKm-145	25	1/20

M4

Plasmid	DNA mass (ng)	DNA Ratio
pPKm-105	100	4/20
pPKm-234	375	15/20
pPKm-233	25	1/20

Figure 11

HEK293 Cells were transfected 24h after plating, followed by a medium change 24h after transfection. For illumination, 1 μmol/m^2/s 1-minute pulses of red light were delivered for 24h, starting 12h after the medium change. Cells were kept in darkness before and after illumination. Cell lysis was performed 72h after transfection, and samples stored in -20C until assayed.

245

Plasmid	DNA mass (ng)	DNA ratio
pPKm-102	10	1/50
pPKm-230	225	22.5/50
pPKm-245	225	22.5/50
pPKm-202	20	2/50
pRL-TK	20	2/50

244

Plasmid	DNA mass (ng)	DNA ratio
pPKm-102	10	1/50
pPKm-230	225	22.5/50
pPKm-244	225	22.5/50
pPKm-202	20	2/50
pRL-TK	20	2/50

248

Plasmid	DNA mass (ng)	DNA ratio
pPKm-102	10	1/50
pPKm-230	225	22.5/50

Appendix B: Table 2 - Transfection and illumination details for each figure, Continued.

pPKm-248	225	22.5/50
pPKm-202	20	2/50
pRL-TK	20	2/50

Figure 12

HEK293 cells were transfected 24h after plating, followed by a medium change 24h after transfection. For this experiment, 15uM of PCB (Frontier Scientific) was added 47h after transfection. Light at 1 μmol/m^2/s in 1-minute pulses of red light was delivered 1h after PCB was added. Cells were kept in darkness before and after illumination. Lysis was performed 72h after transfection, and samples stored in -20C until assayed.

P3-MTAD

Plasmid	DNA mass (ng)	DNA ratio
pPKm-102	325	33/50
pPKm-105	50	5/50
pPKm-112	50	5/50
pPKm-118	50	5/50
pRL-TK	25	2/50

P3-VPR

Plasmid	DNA mass (ng)	DNA ratio
pPKm-102	325	33/50
pPKm-105	50	5/50
pPKm- 226	50	5/50
pPKm-118	50	5/50
pRL-TK	25	2/50

VPR-P3

Plasmid	DNA mass (ng)	DNA ratio
pPKm-102	325	33/50
pPKm-105	50	5/50
pPKm- 227	50	5/50
pPKm-118	50	5/50
pRL-TK	25	2/50

Figure 13A

HEK293 cells were transfected 24h after plating, followed by a medium change 24h after transfection. Cells were lysed 72h after transfection, and samples stored in -20C until assayed.

Appendix B: Table 2 - Transfection and illumination details for each figure, Continued.

Renilla

Plasmid	DNA mass (ng)	DNA ratio
pPKm-102	480	48/50
pRL-TK	20	2/50

TET-UAS-CMVmin

Plasmid	DNA mass (ng)	DNA ratio
pPKm-102	430	43/50
pMZ-802	50	5/50
pRL-TK	20	2/50

G4-UAS-Flucmin

Plasmid	DNA mass (ng)	DNA ratio
pPKm-102	430	43/50
pPKm-118	50	5/50
pRL-TK	20	2/50

G4-UAS-CMVmin

Plasmid	DNA mass (ng)	DNA ratio
pPKm-102	430	43/50
pPKm-202	50	5/50
pRL-TK	20	2/50

Figure 14B

HEK293 cells were transfected 24h after plating, followed by a medium change 24h after transfection. Cells were lysed 72h after transfection, and samples stored in -20C until assayed.

TET-CMV (pMZ-802)

Plasmid	DNA mass (ng)	DNA ratio
pPKm-102	380	38/50
pPKm-293	50	5/50
pMZ-802	50	5/50
pRL-TK	20	2/50

G4-CMV (pPKm-202)

Plasmid	DNA mass (ng)	DNA ratio
pPKm-102	380	38/50
pPKm-292	50	5/50

Appendix B: Table 2 - Transfection and illumination details for each figure, Continued.

<table>
<tr><td></td><td>pPKm-202</td><td>50</td><td>5/50</td></tr>
<tr><td></td><td>pRL-TK</td><td>20</td><td>2/50</td></tr>
<tr><td>Figure 15 an 16</td><td colspan="3">Cells were transfected 24h after plating, followed by a medium change 24h after transfection. In Figure 4D, red light at 1μmol/m^2/s, 0.1μmol/m^2/s, 0.01 μmol/m^2/s and 0.001 μmol/m^2/s were delivered for a total of 24 hours. Similarly, in Figure 4E, continuous illumination for 24h was delivered to the cells, in the intensities listed above. For Figure 4F, red light at 0.1 and 1μmol/m^2/s was continuously delivered or shone for 1-minute pulses every 4 minutes, 9 minutes or 29 minutes, starting 12h after medium change for a total of 24h. For Figures 4G, red light at the intensity of 1μmol/m^2/s was delivered to the cells every 30minutes, every hour, every 2 hours, every 4 hours, 6 hours, 8 hours or every 12 hours. For Figure 5B, cells were kept in darkness, illuminated with far-red light, red light for 24 hours, or with 12 hours or red light followed by darkness or far-red light. For Figure 5C, cells were illuminated with red light at 1 μmol/m^2/s and given a 1 min red light pulse every 5 minutes for 24 hours. In all cases, cells were kept in darkness before and after illumination. Far-red samples were kept under constant illumination starting at medium change. Cell lysis was performed 72h after transfection, and samples stored in -20C until assayed.</td></tr>
<tr><td>All conditions</td><td>Plasmid</td><td>DNA mass (ng)</td><td>DNA ratio</td></tr>
<tr><td></td><td>pPKm-102</td><td>10</td><td>1/50</td></tr>
<tr><td></td><td>pPKm-230</td><td>225</td><td>22.5/50</td></tr>
<tr><td></td><td>pPKm-248</td><td>225</td><td>22.5/50</td></tr>
<tr><td></td><td>pPKm-202</td><td>20</td><td>2/50</td></tr>
<tr><td></td><td>pRL-TK</td><td>20</td><td>2/50</td></tr>
</table>

Appendix C: Table 3 - Parameters for the kinetic model

Parameters used in simulations are detailed bellow. Units are defined in S.I. units with concentrations as the number of molecules for species ($\#molecules$, or c), and parameters as bimolecular rate constants in $\#molecules/s^{-1}$(or c/s^{-1}).

Parameter	Value	Description
k1	0.1228	HO1 and heme binding rate
k2	1e-12	HO1 and heme unbinding rate
k3	0.5687	HO1:Heme and Fd_{red} binding rate
k4	1e-12	HO1:Heme:Fd_{red} unbinding rate
k5	0.2285	Fd_{red}:HO1:Heme unbinding, forming HO1:BV and Fd_{oxi}
k6	0.4750	HO1 unbinding from BV, releasing BV
k7	0.1825	Rate of BV and PcyA binding, forming PcyA:BV
k8	1e-12	PcyA:BV unbinding rate
k9	0.2500	PcyA:BV and Fd_{red} binding rate, forming Fd_{red}:PcyA:BV
k10	1e-12	Unbinding rate of Fd_{red}:PcyA:BV
k11	0.1220	Fd_{red}:PcyA:BV unbinding, forming PcyA:PCB and Fd_{oxi}
k12	0.2667	Unbinding of PcyA:PCB, producing PCB
k13	0.2250	Reduction of Fd_{oxi}, forming Fd_{red}
$k_{deg,PCB}$	0.1567	Degradation of PCB
Heme, at t=0	100	Initial concentration of Heme
HO-1, at t=0	10	Initial concentration of HO-1
PcyA, at t=0	10	Initial concentration of PcyA
$Fd_{red,oxi}$, at t=0	5	Initial concentration of Fd (red and oxi)

Appendix D: Table 4 - Similarity Tables for Ferredoxin and Ferredoxin-dependent Bilin Reductases

(A) The similarity of ferredoxin-dependent bilin reductases and similarity of Fds. (B) The similarity of ferredoxins with eukaryotic sequences containing signal sequences. (C) The similarity of ferredoxins with eukaryotic sequences with signal sequences removed. Sequence alignments were performed using UniProt (http://www.UniProt.org/).

Fd types: 🟫 Cyanobacterial; 🟩 Chloroplastic; 🟥 Mitochondrial.

Species: 🟫 Cyanobacterial, 🟦 Arabidopsis; 🟪 Yeast; 🟨 Human.

A

Ferredoxin-dependent Bilin Reductases		
	THEEB-PCYA	Syn-PCYA
W/SS		
ARATH-Hy2		
% Identity	14.454	8.627
Identical AA	49	49
Similar AA	82	89
W/O SS		
ARATH-Hy2		
% Identity	15.667	15.282
Identical AA	47	46
Similar AA	80	88

B

With Signal Sequence				
	THEEB	SYNP2	ADX_HUMAN	FDX2_HUMAN
THEEB				
% Identity	100	71.429	10.811	12.973
Identical AA	98	70	20	24
Similar AA	0	16	37	36
SYNP2				
% Identity	71.429	100	11.17	12.5

Appendix D: Table 4 - Similarity Tables for Ferredoxin and Ferredoxin-dependent Bilin Reductases, Continued.

Identical AA	70	97	21	23
Similar AA	16	0	37	32
FER1_ARATH	🟩	🟩	🟩	🟩
% Identity	39.597	42.568	16.754	19.565
Identical AA	59	63	32	36
Similar AA	26	23	55	47
FER2_ARATH*	🟩	🟩	🟩	🟩
% Identity	39.597	44.595	17.617	17.857
Identical AA	59	66	34	35
Similar AA	26	20	44	46
FER3_ARATH	🟩	🟩	🟩	🟩
% Identity	40.645	40.645	18.135	17.949
Identical AA	63	63	35	35
Similar AA	25	23	57	49
FER4_ARATH	🟩	🟩	🟩	🟩
% Identity	32.432	40.645	15.426	12.821
Identical AA	48	63	29	25
Similar AA	31	23	57	53
MFDX1_ARATH	🟥	🟥	🟥	🟥
% Identity	12.563	12.183	31.25	32.258
Identical AA	25	24	65	70
Similar AA	37	35	58	56
MFDX2_ARATH	🟥	🟥	🟥	🟥
% Identity	13.568	9.645	30.653	34.653
Identical AA	27	19	61	70
Similar AA	37	40	65	63
ADRX_YEAST	🟥	🟥	🟥	🟥
% Identity	15.517	16.000	29.798	33.333
Identical AA	27	28	59	61
Similar AA	31	28	52	56
ADX_HUMAN	🟥	🟥	🟥	🟥
% Identity	10.811	11.170	100	30.688
Identical AA	20	21	184	58
Similar AA	37	37	0	61

Appendix D: Table 4 - Similarity Tables for Ferredoxin and Ferredoxin-dependent Bilin Reductases, Continued.

FDX2_HUMAN				
% Identity	12.973	12.500	30.688	100
Identical AA	24	23	58	183
Similar AA	36	32	61	0

Without Signal Sequence				
	THEEB	SYNP2	ADX_HUMAN	FDX2_HUMAN
THEEB				
% Identity	100	71.429	16.000	18.045
Identical AA	98	70	20	24
Similar AA	0	16	37	35
SYNP2				
% Identity	71.429	100	16.406	15.909
Identical AA	70	97	21	21
Similar AA	16	0	37	35
FER1_ARATH				
% Identity	59.184	63.918	16.8	18.321
Identical AA	58	62	21	24
Similar AA	25	22	37	31
FER2_ARATH*				
% Identity	59.184	67.010	16.126	18.321
Identical AA	58	65	20	24
Similar AA	26	20	34	31
FER3_ARATH				
% Identity	59.434	59.434	16.794	21.053
Identical AA	63	63	22	28
Similar AA	25	23	36	32
FER4_ARATH				
% Identity	48.485	49.495	15.152	13.74
Identical AA	48	49	20	18
Similar AA	31	29	41	37
MFDX1_ARATH				
% Identity	15.244	14.815	32.927	38.272

Appendix D: Table 4 - Similarity Tables for Ferredoxin and Ferredoxin-dependent Bilin Reductases, Continued.

Identical AA	25	24	54	62
Similar AA	37	35	39	43
MFDX2_ARATH				
% Identity	21.600	15.447	43.2	45.455
Identical AA	27	19	54	60
Similar AA	37	40	38	38
ADRX_YEAST				
% Identity	23.077	22.881	38.71	36.641
Identical AA	27	27	48	48
Similar AA	31	29	35	39
ADX_HUMAN				
% Identity	16.000	16.406	100	31.579
Identical AA	20	21	124	42
Similar AA	37	37	0	46
FDX2_HUMAN				
% Identity	18.045	15.909	31.579	100
Identical AA	24	21	42	131
Similar AA	35	35	46	0

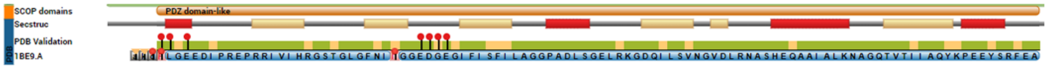
Appendix E: Table 5 - Protein family dataset information

Each protein family used in this work is described below with its PFAM number (not applicable for NCBI-originated MSAs), number of sequences, length of each sequence in the alignment. In addition, the reference protein used to generate visualizations of results, and its PDB ID code is given.

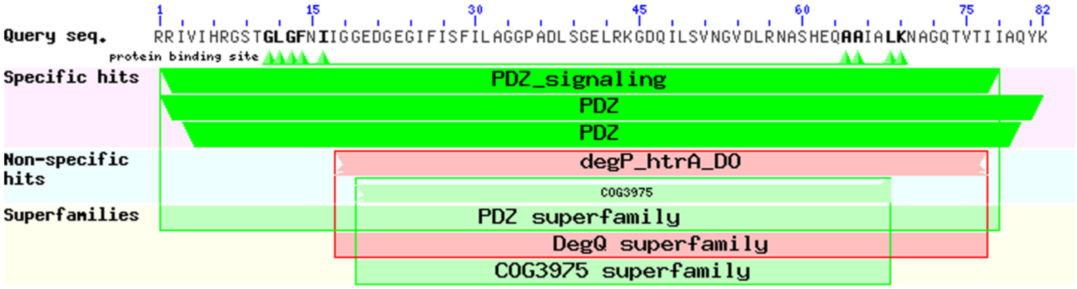
Protein Family	PFAM Number	Number of Sequences	Length of Sequences	Reference Protein Name	Reference Sequence PDB ID
Cadherin	PF00028	6210	93	human protocadherin 9	2EE0 (Sato et al., n.d.)
PDZ	PF00595	12886	81	PDZ domain from PSD-95	1BE9 (Doyle et al., 1996)
	N/A	7517	81	PDZ domain from PSD-95	1BE9 (Doyle et al., 1996)
PhyB	N/A	5333	300	*Arabidopsis thaliana* phytochrome B photosensory module	4OUR (Burgie et al., 2014)
HSP70	PF00012	6223	381	Human Hsp70 ATPase domain	1S3X (Sriram, Osipiuk, Freeman, Morimoto, & Joachimiak, 1997)

Appendix F: Annotated binding site as shown in NCBI, RCSB and UniProt.

A RCSB Annotations for PDZ (PDB ID: 1BE9)

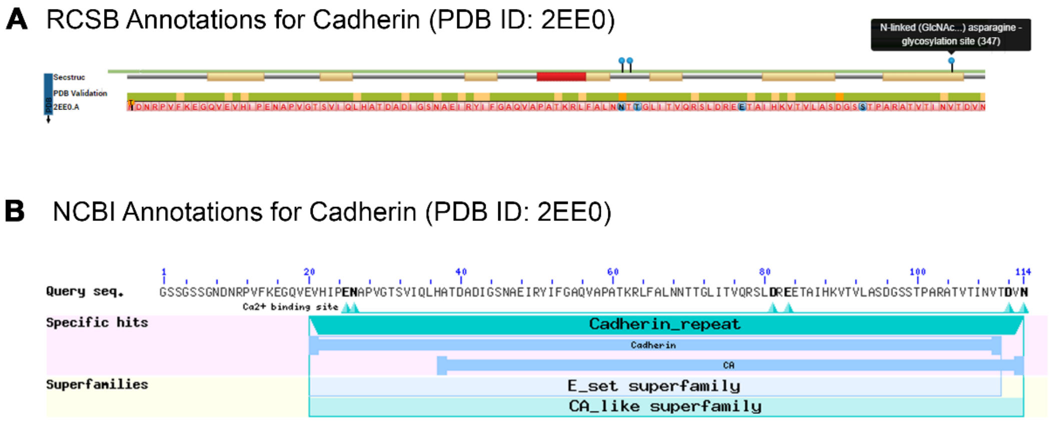

SCOP domains
Secstruc
PDB Validation
1BE9.A
PDZ domain-like
LGEEDIPREPRRIVIHRGSTGLGFNIVGGEDGEGIFISFILAGGPADLSGELRKGDQILSVNGVDLRNASHEQAAIALKNAGQTVTIIAQYKPEEYSRFEA

B NCBI Annotations for PDZ (PDB ID: 1BE9)



No UniProt Annotations for PDZ (PDB ID: 1BE9) available

Figure 35: PDZ (PDB ID: 1BE9) Annotated binding sites (Doyle et al., 1996)(A) Annotations screenshot from RCSB (Doyle et al., 1996). (B) Annotation screenshot from NCBI (National Center for Biotechnology Information, 1988). No relevant annotations were available in UniProt.



A RCSB Annotations for Cadherin (PDB ID: 2EE0)

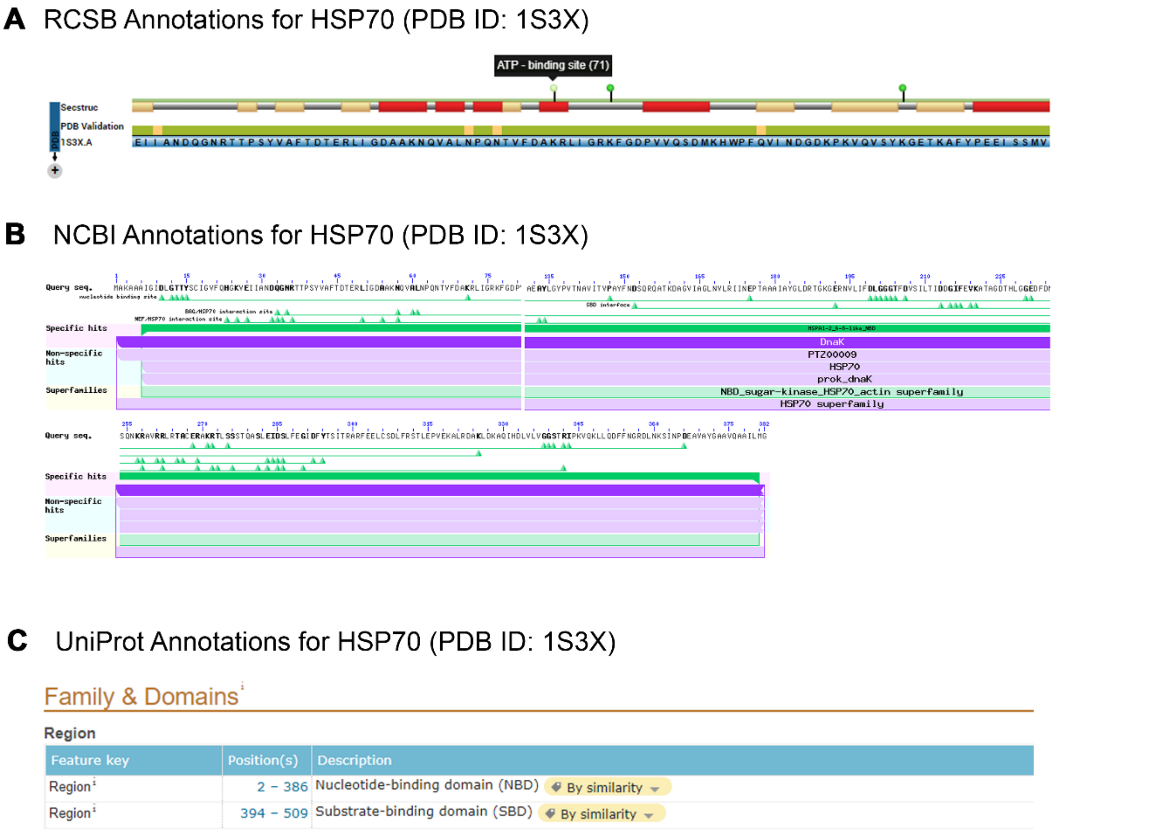
N-linked (GlcNAc...) asparagine - glycosylation site (347)
Secstruc
PDB Validation
2EE0.A
PDB
TDNRPVFKEGQVEVHIPENAPVGTSVIQLHATDADIGSNAEIRYIFGAQVAPATKRLFALNNTTGLITVQRSLDREETAIHKVTVLASDGSSTPARATVTINVTDVN

B NCBI Annotations for Cadherin (PDB ID: 2EE0)

1
20
40
60
80
100
114
Query seq.
GSSGSSGNDNRPVFKEGQVEVHIPENAPVGTSVIQLHATDADIGSNAEIRYIFGAQVAPATKRLFALNNTTGLITVQRSLDREETAIHKVTVLASDGSSTPARATVTINVTDVN
Ca2+ binding site
Specific hits
Cadherin_repeat
Cadherin
CA
Superfamilies
E_set superfamily
CA_like superfamily

No UniProt Annotations for Cadherin (PDB ID: 2EE0)

Figure 36: Cadherin (PDB ID: 2EE0) Annotated binding sites.(Sato et al., n.d.) (A) Annotations screenshot from RCSB (Sato et al., n.d.). (B) Annotation screenshot from NCBI (National Center for Biotechnology Information, 1988). No relevant annotations were available in UniProt (The UniProt Consortium, 2017).



A RCSB Annotations for HSP70 (PDB ID: 1S3X)

ATP - binding site (71)
Secstruc
PDB Validation
1S3X.A
E I I A N D Q G N R T T P S Y V A F T D T E R L I G D A A K N Q V A L N P Q N T V F D A K R L I G R K F G D P V V Q S D M K H W P F Q V I N D G D K P K V Q V S Y K G E T K A F Y P E E I S S M V

B NCBI Annotations for HSP70 (PDB ID: 1S3X)

Query seq.
nucleotide binding site
BAG/HSP70 interaction site
NEF/HSP70 interaction site
SBD interface
Specific hits
HSPA1-2_6-8-like_NBD
Non-specific hits
DnaK
PTZ00009
HSP70
prok_dnaK
Superfamilies
NBD_sugar-kinase_HSP70_actin superfamily
HSP70 superfamily
Query seq.
Specific hits
Non-specific hits
Superfamilies

C UniProt Annotations for HSP70 (PDB ID: 1S3X)

Family & Domains

Region

Feature key	Position(s)	Description
Region	2 – 386	Nucleotide-binding domain (NBD) By similarity
Region	394 – 509	Substrate-binding domain (SBD) By similarity

Figure 37: HSP70 (PDB ID: 1S3X) Annotated binding sites.(Sriram et al., 1997) (A) Annotations screenshot from RCSB (Sriram et al., 1997). (B) Annotation screenshot from NCBI (National Center for Biotechnology Information, 1988). (C) Relevant annotations were available in UniProt.

Appendix G: Table 6 - AUC and F1 scores for all sequences used in this work (full sequence input for training)

Highest values for each protein are shown in bold.

Protein Family	AUC (sequences)	F1 (sequences)	SCA AUC score	SCA F1 score
Cadherin	**0.568**	**0.817**	0.546	0.670
PDZ (NCBI)	**0.715**	**0.840**	0.520	0.753
PDZ (PFAM)	**0.660**	**0.827**	0.520	0.753
PhyB	**0.718**	**0.957**	0.562	0.620
HSP70	0.510	**0.771**	**0.553**	0.709

¥ Result shown for window size 9 with overlap of 1.

Appendix H: Table 7 - AUC window method scores

Highest values for overlap condition are shown in bold.

	Overlap							
Window Size (number of residues)	**1 residue**		**2 residues**		**3 residues**		**4 residues**	
	AUC	**F1**	**AUC**	**F1**	**AUC**	**F1**	**AUC**	**F1**
5	0.375	0.667	0.465	0.827	0.5	0.716	0.583	0.779
6	0.465	0.740	0.520	0.830	**0.611**	0.740	0.437	0.691
7	0.403	0.716	0.451	0.803	0.493	0.790	0.403	0.716
8	0.667	0.840	0.521	0.840	0.520	0.753	0.437	0.778
9	**0.799**	0.815	0.445	0.790	0.500	0.802	0.534	0.778
10	0.340	0.605	0.423	0.753	0.430	0.765	0.410	0.728
11	0.438	0.778	0.410	0.728	0.430	0.765	0.416	0.740
12	0.389	0.691	**0.632**	0.778	0.430	0.765	**0.631**	0.864

Appendix I: Table 8 - AUC *vs*. mean HSW, showing the distribution and variability in the datasets used. Higher AUC is associated with a higher average HSW.

Protein Family	Sequences	AUC	Mean HSW
PhyB (NCBI)	5333	0.778	0.056254
Cadherin (PFAM)	6210	0.568	0.014976
HSP70 (PFAM)	6223	0.510	0.058141
PDZ (NCBI)	7517	0.715	0.010776
PDZ (PFAM)	12886	0.660	0.006286

BIBLIOGRAPHY

Adrian, M., Nijenhuis, W., Hoogstraaten, R. I., Willems, J., & Kapitein, L. C. (2017). A Phytochrome-Derived Photoswitch for Intracellular Transport. *ACS Synthetic Biology [electronic Resource]*, *6*(7), 1248–1256. doi:10.1021/acssynbio.6b00333

Adzhubei, I. A., Schmidt, S., Peshkin, L., Ramensky, V. E., Gerasimova, A., Bork, P., … Sunyaev, S. R. (2010). A method and server for predicting damaging missense mutations. *Nature Methods*, *7*(4), 248–249. doi:10.1038/nmeth0410-248

Aggio, R., Villas-Bôas, S. G., & Ruggiero, K. (2011). Metab: an R package for high-throughput analysis of metabolomics data generated by GC-MS. *Bioinformatics*, *27*(16), 2316–2318. doi:10.1093/bioinformatics/btr379

Ahmed, Z. (2017, June 29). How to Visualize Your Recurrent Neural Network with Attention in Keras. Retrieved February 1, 2018, from https://medium.com/datalogue/attention-in-keras-1892773a4f22

Aliverti, A., Pandini, V., Pennati, A., de Rosa, M., & Zanetti, G. (2008). Structural and functional diversity of ferredoxin-NADP(+) reductases. *Archives of Biochemistry and Biophysics*, *474*(2), 283–291. doi:10.1016/j.abb.2008.02.014

Anfinsen, C. B. (1973). Principles that govern the folding of protein chains. *Science*, *181*(4096), 223–230.

Angermueller, C., Pärnamaa, T., Parts, L., & Stegle, O. (2016). Deep learning for computational biology. *Molecular Systems Biology*, *12*(7), 878. doi:10.15252/msb.20156651

Auldridge, M. E., & Forest, K. T. (2011). Bacterial phytochromes: more than meets the light. *Critical Reviews in Biochemistry and Molecular Biology*, *46*(1), 67–88. doi:10.3109/10409238.2010.546389

Baker, D., & Sali, A. (2001). Protein structure prediction and structural genomics. *Science*, *294*(5540), 93–96. doi:10.1126/science.1065659

Batie, C. J., & Kamin, H. (1984). Electron transfer by ferredoxin:NADP+ reductase. Rapid-reaction evidence for participation of a ternary complex. *The Journal of Biological Chemistry*, *259*(19), 11976–11985.

Beale, S. I. (1993). Biosynthesis of phycobilins. *Chemical Reviews*, *93*(2), 785–802. doi:10.1021/cr00018a008

Beyer, H. M., Juillot, S., Herbst, K., Samodelov, S. L., Müller, K., Schamel, W. W., … Zurbriggen, M. D. (2015). Red Light-Regulated Reversible Nuclear Localization of Proteins in Mammalian Cells and Zebrafish. *ACS Synthetic Biology [electronic Resource]*, *4*(9), 951–958. doi:10.1021/acssynbio.5b00004

Binkowski, T. A., Adamian, L., & Liang, J. (2003). Inferring functional relationships of proteins from local sequence and spatial surface patterns. *Journal of Molecular Biology*, *332*(2), 505–526. doi:10.1016/S0022-2836(03)00882-9

Bochkov, Y., & Palmenberg, A. (2006). Translational efficiency of EMCV IRES in bicistronic vectors is dependent upon IRES sequence and gene location. *Biotechniques*, *41*(3), 283–292. doi:10.2144/000112243

Boyden, E. S., Zhang, F., Bamberg, E., Nagel, G., & Deisseroth, K. (2005). Millisecond-timescale, genetically targeted optical control of neural activity. *Nature Neuroscience*, *8*(9), 1263–1268. doi:10.1038/nn1525

Burén, S., Young, E. M., Sweeny, E. A., Lopez-Torrejón, G., Veldhuizen, M., Voigt, C. A., & Rubio, L. M. (2017). Formation of Nitrogenase NifDK Tetramers in the Mitochondria of Saccharomyces cerevisiae. *ACS Synthetic Biology [electronic Resource]*, *6*(6), 1043–1055. doi:10.1021/acssynbio.6b00371

Burgie, E. S., Bussell, A. N., Walker, J. M., Dubiel, K., & Vierstra, R. D. (2014). Crystal structure of the photosensing module from a red/far-red light-absorbing plant phytochrome. *Proceedings of the National Academy of Sciences of the United States of America*, *111*(28), 10179–10184. doi:10.1073/pnas.1403096111

Burgie, E. S., & Vierstra, R. D. (2014). Phytochromes: an atomic perspective on photoactivation and signaling. *The Plant Cell*, *26*(12), 4568–4583. doi:10.1105/tpc.114.131623

Cahoon, E. B., & Shanklin, J. (2000). Substrate-dependent mutant complementation to select fatty acid desaturase variants for metabolic engineering of plant seed oils. *Proceedings of the National Academy of Sciences of the United States of America*, *97*(22), 12350–12355. doi:10.1073/pnas.210276297

Capra, J. A., & Singh, M. (2007). Predicting functionally important residues from sequence conservation. *Bioinformatics*, *23*(15), 1875–1882. doi:10.1093/bioinformatics/btm270

Cerdán, P. D., & Chory, J. (2003). Regulation of flowering time by light quality. *Nature*, *423*(6942), 881–885. doi:10.1038/nature01636

Chakrabarti, S., Bryant, S. H., & Panchenko, A. R. (2007). Functional specificity lies within the properties and evolutionary changes of amino acids. *Journal of Molecular Biology*, *373*(3), 801–810. doi:10.1016/j.jmb.2007.08.036

Chavez, A., Scheiman, J., Vora, S., Pruitt, B. W., Tuttle, M., P R Iyer, E., … Church, G. M. (2015). Highly efficient Cas9-mediated transcriptional programming. *Nature Methods*, *12*(4), 326–328. doi:10.1038/nmeth.3312

Chellaboina, V., Bhat, S., Haddad, W., & Bernstein, D. (2009). Modeling and analysis of mass-action kinetics. *IEEE Control Systems Magazine*, *29*(4), 60–78. doi:10.1109/MCS.2009.932926

Chen, D., Gibson, E. S., & Kennedy, M. J. (2013). A light-triggered protein secretion system. *The Journal of Cell Biology*, *201*(4), 631–640. doi:10.1083/jcb.201210119

Chen, M., Tao, Y., Lim, J., Shaw, A., & Chory, J. (2005). Regulation of phytochrome B nuclear localization through light-dependent unmasking of nuclear-localization signals. *Current Biology*, *15*(7), 637–642. doi:10.1016/j.cub.2005.02.028

Chenna, R., Sugawara, H., Koike, T., Lopez, R., Gibson, T. J., Higgins, D. G., & Thompson, J. D. (2003). Multiple sequence alignment with the Clustal series of programs. *Nucleic Acids Research*, *31*(13), 3497–3500. doi:10.1093/nar/gkg500

Chiu, F.-Y., Chen, Y.-R., & Tu, S.-L. (2010). Electrostatic interaction of phytochromobilin synthase and ferredoxin for biosynthesis of phytochrome chromophore. *The Journal of Biological Chemistry*, *285*(7), 5056–5065. doi:10.1074/jbc.M109.075747

Cuff, J. A., & Barton, G. J. (2000). Application of multiple sequence alignment profiles to improve protein secondary structure prediction. *Proteins: Structure, Function, and Bioinformatics*. Retrieved from http://onlinelibrary.wiley.com/doi/10.1002/1097-0134(20000815)40:3%3C502::AID-PROT170%3E3.0.CO;2-Q/full

Curatti, L., & Rubio, L. M. (2014). Challenges to develop nitrogen-fixing cereals by direct nif-gene transfer. *Plant Science*, *225*, 130–137. doi:10.1016/j.plantsci.2014.06.003

De Juan, D., Pazos, F., & Valencia, A. (2013). Emerging methods in protein co-evolution. *Nature Reviews. Genetics*, *14*(4), 249–261. doi:10.1038/nrg3414

Di Lena, P., Nagata, K., & Baldi, P. (2012). Deep architectures for protein contact map prediction. *Bioinformatics*, *28*(19), 2449–2457. doi:10.1093/bioinformatics/bts475

Doyle, D. A., Lee, A., Lewis, J., Kim, E., Sheng, M., & MacKinnon, R. (1996). Crystal structures of a complexed and peptide-free membrane protein-binding domain: molecular basis of peptide recognition by PDZ. *Cell*, *85*(7), 1067–1076. doi:10.1016/S0092-8674(00)81307-0

Du, Q.-S., Meng, J.-Z., Wang, C.-H., Long, S.-Y., & Huang, R.-B. (2011). Structural position correlation analysis (SPCA) for protein family. *Plos One*, *6*(12), e28206. doi:10.1371/journal.pone.0028206

Elcock, A. H. (2001). Prediction of functionally important residues based solely on the computed energetics of protein structure. *Journal of Molecular Biology*, *312*(4), 885–896. doi:10.1006/jmbi.2001.5009

Elich, T. D., & Chory, J. (1997). Biochemical characterization of Arabidopsis wild-type and mutant phytochrome B holoproteins. *The Plant Cell*, *9*(12), 2271–2280. doi:10.1105/tpc.9.12.2271

Elloumi, M., Iliopoulos, C. S., Wang, J. T., & Zomaya, A. Y. (2015). *Pattern recognition in computational molecular biology: techniques and approaches*. (M. Elloumi, C. S. Iliopoulos, J. T. L. Wang, & A. Y. Zomaya, Eds.). Hoboken, NJ, USA: John Wiley & Sons, Inc. doi:10.1002/9781119078845

Feng, D.-F., & Doolittle, R. F. (1987). Progressive sequence alignment as a prerequisitetto correct phylogenetic trees. *Journal of Molecular Evolution*, *25*(4), 351–360. doi:10.1007/BF02603120

Fischer, J. D., Mayer, C. E., & Söding, J. (2008). Prediction of protein functional residues from sequence by probability density estimation. *Bioinformatics*, *24*(5), 613–620. doi:10.1093/bioinformatics/btm626

Folcher, M., Oesterle, S., Zwicky, K., Thekkottil, T., Heymoz, J., Hohmann, M., … Fussenegger, M. (2014). Mind-controlled transgene expression by a wireless-powered optogenetic designer cell implant. *Nature Communications*, *5*, 5392. doi:10.1038/ncomms6392

Frankenberg, N, Mukougawa, K., Kohchi, T., & Lagarias, J. C. (2001). Functional genomic analysis of the HY2 family of ferredoxin-dependent bilin reductases from oxygenic photosynthetic organisms. *The Plant Cell*, *13*(4), 965–978.

Frankenberg, Nicole, & Lagarias, J. C. (2003). Phycocyanobilin:ferredoxin oxidoreductase of Anabaena sp. PCC 7120. Biochemical and spectroscopic. *The Journal of Biological Chemistry*, *278*(11), 9219–9226. doi:10.1074/jbc.M211643200

Gambetta, G. A., & Lagarias, J. C. (2001). Genetic engineering of phytochrome biosynthesis in bacteria. *Proceedings of the National Academy of Sciences of the United States of America*, *98*(19), 10566–10571. doi:10.1073/pnas.191375198

Gaspar, H. B., Cooray, S., Gilmour, K. C., Parsley, K. L., Zhang, F., Adams, S., … Thrasher, A. J. (2011). Hematopoietic stem cell gene therapy for adenosine deaminase-deficient severe combined immunodeficiency leads to long-term immunological recovery and metabolic correction. *Science Translational Medicine*, *3*(97), 97ra80. doi:10.1126/scitranslmed.3002716

Graves, A., Mohamed, A., & Hinton, G. (2013). Speech recognition with deep recurrent neural networks. In *2013 IEEE International Conference on Acoustics, Speech and Signal Processing* (pp. 6645–6649). IEEE. doi:10.1109/ICASSP.2013.6638947

Halabi, N., Rivoire, O., Leibler, S., & Ranganathan, R. (2009). Protein sectors: evolutionary units of three-dimensional structure. *Cell*, *138*(4), 774–786. doi:10.1016/j.cell.2009.07.038

Hamilton, N., Burrage, K., Ragan, M. A., & Huber, T. (2004). Protein contact prediction using patterns of correlation. *Proteins*, *56*(4), 679–684. doi:10.1002/prot.20160

Hanke, G., & Mulo, P. (2013). Plant type ferredoxins and ferredoxin-dependent metabolism. *Plant, Cell & Environment*, *36*(6), 1071–1084. doi:10.1111/pce.12046

Hanke, G. T., Kurisu, G., Kusunoki, M., & Hase, T. (2004). Fd : FNR Electron Transfer Complexes: Evolutionary Refinement of Structural Interactions. *Photosynthesis Research*, *81*(3), 317–327. doi:10.1023/B:PRES.0000036885.01534.b8

Henikoff, S., & Henikoff, J. G. (1994). Position-based sequence weights. *Journal of Molecular Biology*, *243*(4), 574–578. doi:10.1016/0022-2836(94)90032-9

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, *9*(8), 1735–1780. doi:10.1162/neco.1997.9.8.1735

Hübschmann, T., Börner, T., Hartmann, E., & Lamparter, T. (2001). Characterization of the Cph1 holo-phytochrome from*Synechocystis* sp. PCC 6803. *European Journal of Biochemistry*, *268*(7), 2055–2063. doi:10.1046/j.1432-1327.2001.02083.x

Hughes, R. M., Bolger, S., Tapadia, H., & Tucker, C. L. (2012). Light-mediated control of DNA transcription in yeast. *Methods*, *58*(4), 385–391. doi:10.1016/j.ymeth.2012.08.004

Jo, T., Hou, J., Eickholt, J., & Cheng, J. (2015). Improving protein fold recognition by deep learning networks. *Scientific Reports*, *5*(1), 17573. doi:10.1038/srep17573

Jones, D. T., Buchan, D. W. A., Cozzetto, D., & Pontil, M. (2012). PSICOV: precise structural contact prediction using sparse inverse covariance estimation on large multiple sequence alignments. *Bioinformatics*, *28*(2), 184–190. doi:10.1093/bioinformatics/btr638

Kaberniuk, A. A., Shemetov, A. A., & Verkhusha, V. V. (2016). A bacterial phytochrome-based optogenetic system controllable with near-infrared light. *Nature Methods*, *13*(7), 591–597. doi:10.1038/nmeth.3864

Karniol, B., Wagner, J. R., Walker, J. M., & Vierstra, R. D. (2005). Phylogenetic analysis of the phytochrome superfamily reveals distinct microbial subfamilies of photoreceptors. *The Biochemical Journal*, *392*(Pt 1), 103–116. doi:10.1042/BJ20050826

Kawano, F., Suzuki, H., Furuya, A., & Sato, M. (2015). Engineered pairs of distinct photoswitches for optogenetic control of cellular proteins. *Nature Communications*, *6*, 6256. doi:10.1038/ncomms7256

Kikis, E. A., Oka, Y., Hudson, M. E., Nagatani, A., & Quail, P. H. (2009). Residues clustered in the light-sensing knot of phytochrome B are necessary for conformerspecific binding to signaling partner PIF3. *PLoS Genetics*, *5*(1), e1000352. doi:10.1371/journal.pgen.1000352

Konermann, S., Brigham, M. D., Trevino, A. E., Hsu, P. D., Heidenreich, M., Cong, L., … Zhang, F. (2013). Optical control of mammalian endogenous transcription and epigenetic states. *Nature*, *500*(7463), 472–476. doi:10.1038/nature12466

Kunkel, T., Speth, V., Büche, C., & Schäfer, E. (1995). In vivo characterization of phytochrome-phycocyanobilin adducts in yeast. *The Journal of Biological Chemistry*, *270*(34), 20193–20200.

Kyriakakis, P., Catanho, M., Hoffner, N., Thavarajah, W., Jian-Yu, V., Chao, S.-S., … Coleman, T. (2018). Biosynthesis of Orthogonal Molecules Using Ferredoxin and Ferredoxin-NADP+ Reductase Systems Enables Genetically Encoded PhyB Optogenetics. *ACS Synthetic Biology [electronic Resource]*. doi:10.1021/acssynbio.7b00413

Lamb, J., Crawford, E. D., Peck, D., Modell, J. W., Blat, I. C., Wrobel, M. J., … Golub, T. R. (2006). The Connectivity Map: using gene-expression signatures to connect small molecules, genes, and disease. *Science*, *313*(5795), 1929–1935. doi:10.1126/science.1132939

Landgraf, F. T., Forreiter, C., Hurtado Picó, A., Lamparter, T., & Hughes, J. (2001). Recombinant holophytochrome in*Escherichia coli*. *FEBS Letters*, *508*(3), 459–462. doi:10.1016/S0014-5793(01)02988-X

Lee, H.-J., & Zheng, J. J. (2010). PDZ domains and their binding partners: structure, specificity, and modification. *Cell Communication and Signaling*, *8*, 8. doi:10.1186/1478-811X-8-8

Levskaya, A., Weiner, O. D., Lim, W. A., & Voigt, C. A. (2009). Spatiotemporal control of cell signalling using a light-switchable protein interaction. *Nature*, *461*(7266), 997–1001. doi:10.1038/nature08446

Li, J., Li, G., Wang, H., & Wang Deng, X. (2011). Phytochrome signaling mechanisms. *The Arabidopsis Book / American Society of Plant Biologists*, *9*, e0148. doi:10.1199/tab.0148

Licursi, M., Christian, S. L., Pongnopparat, T., & Hirasawa, K. (2011). In vitro and in vivo comparison of viral and cellular internal ribosome entry sites for bicistronic vector expression. *Gene Therapy*, *18*(6), 631–636. doi:10.1038/gt.2011.11

Lin, J Y, Lin, M. Z., Steinbach, P., & Tsien, R. Y. (2009). Characterization of engineered channelrhodopsin variants with improved properties and kinetics. *Biophysical Journal*, *96*(5), 1803–1814. doi:10.1016/j.bpj.2008.11.034

Lin, John Y, Knutsen, P. M., Muller, A., Kleinfeld, D., & Tsien, R. Y. (2013). ReaChR: a red-shifted variant of channelrhodopsin enables deep transcranial optogenetic excitation. *Nature Neuroscience*, *16*(10), 1499–1508. doi:10.1038/nn.3502

Lockless, S. W., & Ranganathan, R. (1999). Evolutionarily conserved pathways of energetic connectivity in protein families. *Science*, *286*(5438), 295–299. doi:10.1126/science.286.5438.295

Lopez, C. F., Muhlich, J. L., Bachman, J. A., & Sorger, P. K. (2013). Programming biological models in Python using PySB. *Molecular Systems Biology*, *9*, 646. doi:10.1038/msb.2013.1

Magnan, C. N., & Baldi, P. (2014). SSpro/ACCpro 5: almost perfect prediction of protein secondary structure and relative solvent accessibility using profiles, machine learning and structural similarity. *Bioinformatics*, *30*(18), 2592–2597. doi:10.1093/bioinformatics/btu352

Marsella, L., Sirocco, F., Trovato, A., Seno, F., & Tosatto, S. C. E. (2009). REPETITA: detection and discrimination of the periodicity of protein solenoid repeats by discrete Fourier transform. *Bioinformatics*, *25*(12), i289–95. doi:10.1093/bioinformatics/btp232

Matsubara, H., & Saeki, K. (1992). Structural and functional diversity of ferredoxins and related proteins (pp. 223–280). Elsevier. doi:10.1016/S0898-8838(08)60065-3

Mattis, J., Tye, K. M., Ferenczi, E. A., Ramakrishnan, C., O'Shea, D. J., Prakash, R., … Deisseroth, K. (2011). Principles for applying optogenetic tools derived from direct comparative analysis of microbial opsins. *Nature Methods*, *9*(2), 159–172. doi:10.1038/nmeth.1808

Mihalek, I., Res, I., & Lichtarge, O. (2004). A family of evolution-entropy hybrid methods for ranking protein residues by importance. *Journal of Molecular Biology*, *336*(5), 1265–1282. doi:10.1016/j.jmb.2003.12.078

Milias-Argeitis, A., Summers, S., Stewart-Ornstein, J., Zuleta, I., Pincus, D., El-Samad, H., … Lygeros, J. (2011). In silico feedback for in vivo regulation of a gene expression circuit. *Nature Biotechnology*, *29*(12), 1114–1116. doi:10.1038/nbt.2018

Min, S., Lee, B., & Yoon, S. (2017). Deep learning in bioinformatics. *Briefings in Bioinformatics*, *18*(5), 851–869. doi:10.1093/bib/bbw068

Mizuguchi, H., Xu, Z., Ishii-Watabe, A., Uchida, E., & Hayakawa, T. (2000). IRES-dependent second gene expression is significantly lower than cap-dependent first gene expression in a bicistronic vector. *Molecular Therapy*, *1*(4), 376–382. doi:10.1006/mthe.2000.0050

Morcos, F., Pagnani, A., Lunt, B., Bertolino, A., Marks, D. S., Sander, C., … Weigt, M. (2011). Direct-coupling analysis of residue coevolution captures native contacts across many protein families. *Proceedings of the National Academy of Sciences of the United States of America*, *108*(49), E1293–301. doi:10.1073/pnas.1111471108

Mukougawa, K., Kanamoto, H., Kobayashi, T., Yokota, A., & Kohchi, T. (2006). Metabolic engineering to produce phytochromes with phytochromobilin, phycocyanobilin, or phycoerythrobilin chromophore in Escherichia coli. *FEBS Letters*, *580*(5), 1333–1338. doi:10.1016/j.febslet.2006.01.051

Müller, K., Engesser, R., Metzger, S., Schulz, S., Kämpf, M. M., Busacker, M., … Weber, W. (2013). A red/far-red light-responsive bi-stable toggle switch to control gene expression in mammalian cells. *Nucleic Acids Research*, *41*(7), e77. doi:10.1093/nar/gkt002

Müller, K., Engesser, R., Timmer, J., Nagy, F., Zurbriggen, M. D., & Weber, W. (2013). Synthesis of phycocyanobilin in mammalian cells. *Chemical Communications*, *49*(79), 8970–8972. doi:10.1039/c3cc45065a

Müller, K., Zurbriggen, M. D., & Weber, W. (2014). Control of gene expression using a red- and far-red light-responsive bi-stable toggle switch. *Nature Protocols*, *9*(3), 622–632. doi:10.1038/nprot.2014.038

Murphy, L. R., Wallqvist, A., & Levy, R. M. (2000). Simplified amino acid alphabets for protein fold recognition and implications for folding. *Protein Engineering, Design and Selection*, *13*(3), 149–152. doi:10.1093/protein/13.3.149

Nagao, C., Nagano, N., & Mizuguchi, K. (2014). Prediction of detailed enzyme functions and identification of specificity determining residues by random forests. *Plos One*, *9*(1), e84623. doi:10.1371/journal.pone.0084623

National Center for Biotechnology Information, N. L. of M. (US). (1988). National Center for Biotechnology Information (NCBI). Retrieved December 23, 2017, from https://www.nlm.nih.gov/

Ofran, Y., & Rost, B. (2003). Predicted protein-protein interaction sites from local sequence information. *FEBS Letters*, *544*(1-3), 236–239. doi:10.1016/S0014-5793(03)00456-3

Okada, K. (2009). HO1 and PcyA proteins involved in phycobilin biosynthesis form a 1:2 complex with ferredoxin-1 required for photosynthesis. *FEBS Letters*, *583*(8), 1251–1256. doi:10.1016/j.febslet.2009.03.052

Ouzounis, C., Pérez-Irratxeta, C., Sander, C., & Valencia, A. (1998). Are binding residues conserved? *Pacific Symposium on Biocomputing*, 401–412.

Park, Y., & Kellis, M. (2015). Deep learning for regulatory genomics. *Nature Biotechnology*, *33*(8), 825–826. doi:10.1038/nbt.3313

Pathak, G. P., Strickland, D., Vrana, J. D., & Tucker, C. L. (2014). Benchmarking of optical dimerizer systems. *ACS Synthetic Biology [electronic Resource]*, *3*(11), 832–838. doi:10.1021/sb500291r

Peifer, M., & Timmer, J. (2007). Parameter estimation in ordinary differential equations for biochemical processes using the method of multiple shooting. *IET Systems Biology*, *1*(2), 78–88. doi:10.1049/iet-syb:20060067

Pinto, R., Harrison, J. S., Hsu, T., Jacobs, W. R., & Leyh, T. S. (2007). Sulfite reduction in mycobacteria. *Journal of Bacteriology*, *189*(18), 6714–6722. doi:10.1128/JB.00487-07

Pugalenthi, G., Kumar, K. K., Suganthan, P. N., & Gangal, R. (2008). Identification of catalytic residues from protein structure using support vector machine with sequence and structural features. *Biochemical and Biophysical Research Communications*, *367*(3), 630–634. doi:10.1016/j.bbrc.2008.01.038

Qin, J. Y., Zhang, L., Clift, K. L., Hulur, I., Xiang, A. P., Ren, B.-Z., & Lahn, B. T. (2010). Systematic comparison of constitutive promoters and the doxycycline-inducible promoter. *Plos One*, *5*(5), e10611. doi:10.1371/journal.pone.0010611

Quail, P. H., Boylan, M. T., Parks, B. M., Short, T. W., Xu, Y., & Wagner, D. (1995). Phytochromes: photosensory perception and signal transduction. *Science*, *268*(5211), 675–680.

Quang, D., Chen, Y., & Xie, X. (2015). DANN: a deep learning approach for annotating the pathogenicity of genetic variants. *Bioinformatics*, *31*(5), 761–763. doi:10.1093/bioinformatics/btu703

Ranganathan, R., & Ross, E. M. (1997). PDZ domain proteins: scaffolds for signaling complexes. *Current Biology*, *7*(12), R770–3. doi:10.1016/S0960-9822(06)00401-5

Rekittke, I., Olkhova, E., Wiesner, J., Demmer, U., Warkentin, E., Jomaa, H., & Ermler, U. (2013). Structure of the (E)-4-hydroxy-3-methyl-but-2-enyl-diphosphate reductase from Plasmodium falciparum. *FEBS Letters*, *587*(24), 3968–3972. doi:10.1016/j.febslet.2013.10.029

Remberg, A., Ruddat, A., Braslavsky, S. E., Gärtner, W., & Schaffner, K. (1998). Chromophore incorporation, Pr to Pfr kinetics, and Pfr thermal reversion of recombinant N-terminal fragments of phytochrome A and B chromoproteins. *Biochemistry*, *37*(28), 9983–9990. doi:10.1021/bi980575x

Ribeiro, A. A. S. T., & Ortiz, V. (2015). Energy propagation and network energetic coupling in proteins. *The Journal of Physical Chemistry. B*, *119*(5), 1835–1846. doi:10.1021/jp509906m

Rockwell, N. C., Su, Y.-S., & Lagarias, J. C. (2006). Phytochrome structure and signaling mechanisms. *Annual Review of Plant Biology*, *57*, 837–858. doi:10.1146/annurev.arplant.56.032604.144208

Rodriguez-Romero, J., Hedtke, M., Kastner, C., Müller, S., & Fischer, R. (2010). Fungi, hidden in soil or up in the air: light makes a difference. *Annual Review of Microbiology*, *64*, 585–610. doi:10.1146/annurev.micro.112408.134000

Rost, B., & Sander, C. (1993). Improved prediction of protein secondary structure by use of sequence profiles and neural networks. *Proceedings of the National Academy of Sciences of the United States of America*, *90*(16), 7558–7562.

Rost, B., & Sander, C. (1994). Combining evolutionary information and neural networks to predict protein secondary structure. *Proteins*, *19*(1), 55–72. doi:10.1002/prot.340190108

Sadowski, M. I., & Jones, D. T. (2009). The sequence-structure relationship and protein function prediction. *Current Opinion in Structural Biology*, *19*(3), 357–362. doi:10.1016/j.sbi.2009.03.008

Sakamoto, K., & Nagatani, A. (1996). Nuclear localization activity of phytochrome B. *The Plant Journal: For Cell and Molecular Biology*, *10*(5), 859–868.

Sato, M., Koshiba, S., Watanabe, S., Harada, T., Kigawa, T., & Yokoyama, S. (n.d.). Solution structures of the CA domain of human protocadherin 9. *To Be Published*. Retrieved from https://www.rcsb.org/pdb/explore.do?structureId=2ee0

Schindelin, J., Arganda-Carreras, I., Frise, E., Kaynig, V., Longair, M., Pietzsch, T., … Cardona, A. (2012). Fiji: an open-source platform for biological-image analysis. *Nature Methods*, *9*(7), 676–682. doi:10.1038/nmeth.2019

Sheftel, A. D., Stehling, O., Pierik, A. J., Elsässer, H.-P., Mühlenhoff, U., Webert, H., … Lill, R. (2010). Humans possess two mitochondrial ferredoxins, Fdx1 and Fdx2, with distinct roles in steroidogenesis, heme, and Fe/S cluster biosynthesis. *Proceedings of the National Academy of Sciences of the United States of America*, *107*(26), 11775–11780. doi:10.1073/pnas.1004250107

Shimizu-Sato, S., Huq, E., Tepperman, J. M., & Quail, P. H. (2002). A light-switchable gene promoter system. *Nature Biotechnology*, *20*(10), 1041–1044. doi:10.1038/nbt734

Shintani, D., & DellaPenna, D. (1998). Elevating the vitamin E content of plants through metabolic engineering. *Science*, *282*(5396), 2098–2100.

Smith, R. W., Helwig, B., Westphal, A. H., Pel, E., Hörner, M., Beyer, H. M., … Fleck, C. (2016). Unearthing the transition rates between photoreceptor conformers. *BMC Systems Biology*, *10*(1), 110. doi:10.1186/s12918-016-0368-y

Somarowthu, S., & Ondrechen, M. J. (2012). POOL server: machine learning application for functional site prediction in proteins. *Bioinformatics*, *28*(15), 2078–2079. doi:10.1093/bioinformatics/bts321

Somarowthu, S., Yang, H., Hildebrand, D. G. C., & Ondrechen, M. J. (2011). High-performance prediction of functional residues in proteins with machine learning and computed input features. *Biopolymers*, *95*(6), 390–400. doi:10.1002/bip.21589

Sønderby, S. K., Sønderby, C. K., Nielsen, H., & Winther, O. (2015). Convolutional LSTM networks for subcellular localization of proteins. In A.-H. Dediu, F. Hernández-Quiroz, C. Martín-Vide, & D. A. Rosenblueth (Eds.), *Algorithms for computational biology* (pp. 68–80). Cham: Springer International Publishing. doi:10.1007/978-3-319-21233-3_6

Spencer, M., Eickholt, J., & Jianlin Cheng. (2015). A Deep Learning Network Approach to ab initio Protein Secondary Structure Prediction. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, *12*(1), 103–112. doi:10.1109/TCBB.2014.2343960

Spiltoir, J. I., Strickland, D., Glotzer, M., & Tucker, C. L. (2016). Optical control of peroxisomal trafficking. *ACS Synthetic Biology [electronic Resource]*, *5*(7), 554–560. doi:10.1021/acssynbio.5b00144

Sriram, M., Osipiuk, J., Freeman, B., Morimoto, R., & Joachimiak, A. (1997). Human Hsp70 molecular chaperone binds two calcium ions within the ATPase domain. *Structure*, *5*(3), 403–414.

Strimmer, K., & von Haeseler, A. (1997). Likelihood-mapping: a simple method to visualize phylogenetic content of a sequence alignment. *Proceedings of the National Academy of Sciences of the United States of America*, *94*(13), 6815–6819.

Szymczak, A. L., Workman, C. J., Wang, Y., Vignali, K. M., Dilioglou, S., Vanin, E. F., & Vignali, D. A. (2004). Correction of multi-gene deficiency in vivo using a single “self-cleaving” 2A peptide-based retroviral vector. *Nature Biotechnology*, *22*(5), 589–594. doi:10.1038/nbt957

Tachikawa, K., Schröder, O., Frey, G., Briggs, S. P., & Sera, T. (2004). Regulation of the endogenous VEGF-A gene by exogenous designed regulatory proteins. *Proceedings of the National Academy of Sciences of the United States of America*, *101*(42), 15225–15230. doi:10.1073/pnas.0406473101

Terry, M. J., McDowell, M. T., & Lagarias, J. C. (1995). (3Z)- and (3E)-phytochromobilin are intermediates in the biosynthesis of the phytochrome chromophore. *The Journal of Biological Chemistry*, *270*(19), 11111–11118.

The UniProt Consortium. (2017). UniProt: the universal protein knowledgebase. *Nucleic Acids Research*, *45*(D1), D158–D169. doi:10.1093/nar/gkw1099

Tieleman, T., & Hinton, G. (2012). RMSprop. *COURSERA: Neural Networks for Machine Learning*.

Toettcher, J. E., Gong, D., Lim, W. A., & Weiner, O. D. (2011). Light-based feedback for controlling intracellular signaling dynamics. *Nature Methods*, *8*(10), 837–839. doi:10.1038/nmeth.1700

Tooley, A. J., Cai, Y. A., & Glazer, A. N. (2001). Biosynthesis of a fluorescent cyanobacterial C-phycocyanin holo-alpha subunit in a heterologous host. *Proceedings of the National Academy of Sciences of the United States of America*, *98*(19), 10560–10565. doi:10.1073/pnas.181340998

Tu, S.-L., Gunn, A., Toney, M. D., Britt, R. D., & Lagarias, J. C. (2004). Biliverdin reduction by cyanobacterial phycocyanobilin:ferredoxin oxidoreductase (PcyA) proceeds via linear tetrapyrrole radical intermediates. *Journal of the American Chemical Society*, *126*(28), 8682–8693. doi:10.1021/ja049280z

Tung, C.-W., & Ho, S.-Y. (2007). POPI: predicting immunogenicity of MHC class I binding peptides by mining informative physicochemical properties. *Bioinformatics*, *23*(8), 942–949. doi:10.1093/bioinformatics/btm061

Tyszkiewicz, A. B., & Muir, T. W. (2008). Activation of protein splicing with light in yeast. *Nature Methods*, *5*(4), 303–305. doi:10.1038/nmeth.1189

Velázquez Escobar, F., Buhrke, D., Fernandez Lopez, M., Shenkutie, S. M., von Horsten, S., Essen, L.-O., … Hildebrandt, P. (2017). Structural communication between the chromophore-binding pocket and the N-terminal extension in plant phytochrome phyB. *FEBS Letters*, *591*(9), 1258–1265. doi:10.1002/1873-3468.12642

Von Horsten, S., Straß, S., Hellwig, N., Gruth, V., Klasen, R., Mielcarek, A., … Essen, L.-O. (2016). Mapping light-driven conformational changes within the photosensory module of plant phytochrome B. *Scientific Reports*, *6*, 34366. doi:10.1038/srep34366

Walsh, I., Martin, A. J. M., Di Domenico, T., & Tosatto, S. C. E. (2012). ESpritz: accurate and fast prediction of protein disorder. *Bioinformatics*, *28*(4), 503–509. doi:10.1093/bioinformatics/btr682

Wang, D., & Nyberg, E. (2015). A Long Short-Term Memory Model for Answer Sentence Selection in Question Answering . *ACL*. Retrieved from https://www.semanticscholar.org/paper/A-Long-Short-Term-Memory-Model-for-Answer-Sentence-Wang-Nyberg/828dbeb7cf922dc9b6657dd169b8d26d2b58eedb

Wang, S., Peng, J., Ma, J., & Xu, J. (2016). Protein secondary structure prediction using deep convolutional neural fields. *Scientific Reports*, *6*, 18962. doi:10.1038/srep18962

Wang, X., Chen, X., & Yang, Y. (2012). Spatiotemporal control of gene expression by a light-switchable transgene system. *Nature Methods*, *9*(3), 266–269. doi:10.1038/nmeth.1892

Waterhouse, A. M., Procter, J. B., Martin, D. M. A., Clamp, M., & Barton, G. J. (2009). Jalview Version 2--a multiple sequence alignment editor and analysis workbench. *Bioinformatics*, *25*(9), 1189–1191. doi:10.1093/bioinformatics/btp033

Watson, J. D., Laskowski, R. A., & Thornton, J. M. (2005). Predicting protein function from sequence and structural data. *Current Opinion in Structural Biology*, *15*(3), 275–284. doi:10.1016/j.sbi.2005.04.003

Yan, C., Dobbs, D., & Honavar, V. (2004). A two-stage classifier for identification of protein-protein interface residues. *Bioinformatics*, *20 Suppl 1*, i371–8. doi:10.1093/bioinformatics/bth920

Yang, Z., Yang, D., Dyer, C., He, X., Smola, A., & Hovy, E. (2016). Hierarchical attention networks for document classification. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (pp. 1480–1489). Stroudsburg, PA, USA: Association for Computational Linguistics. doi:10.18653/v1/N16-1174

Yanovsky, M. J., & Kay, S. A. (2003). Living by the calendar: how plants know when to flower. *Nature Reviews. Molecular Cell Biology*, *4*(4), 265–275. doi:10.1038/nrm1077

Yizhar, O., Fenno, L. E., Davidson, T. J., Mogri, M., & Deisseroth, K. (2011). Optogenetics in neural systems. *Neuron*, *71*(1), 9–34. doi:10.1016/j.neuron.2011.06.004

Yonekura-Sakakibara, K., Onda, Y., Ashikari, T., Tanaka, Y., Kusumi, T., & Hase, T. (2000). Analysis of reductant supply systems for ferredoxin-dependent sulfite reductase in photosynthetic and nonphotosynthetic organs of maize. *Plant Physiology*, *122*(3), 887–894.

Zhang, K., & Cui, B. (2015). Optogenetic control of intracellular signaling pathways. *Trends in Biotechnology*, *33*(2), 92–100. doi:10.1016/j.tibtech.2014.11.007

Zhou, X. X., Chung, H. K., Lam, A. J., & Lin, M. Z. (2012). Optical control of protein activity by fluorescent protein domains. *Science (New York)*, *338*(6108), 810–814. doi:10.1126/science.1226854

Zhou, X. Y., Morreau, H., Rottier, R., Davis, D., Bonten, E., Gillemans, N., … d Azzo, A. (1995). Mouse model for the lysosomal disorder galactosialidosis and correction of the phenotype with overexpressing erythroid precursor cells. *Genes & Development*, *9*(21), 2623–2634.

Zvelebil, M. J., & Baum, J. O. (2008). *Understanding Bioinformatics* (illustrated.). Garland Science.