# Engineering Molecules, Mineralization and Magnetism in Biology by Directed Evolution and Computation

Dissertation Advisor: Pamela A. Silver Xueliang Leon Liu

**Engineering Molecules, Mineralization and Magnetism in Biology by Directed Evolution and Computation**

## ABSTRACT

The intersection of synthetic biology with physics and computer science generates rich opportunities for both advancing our understanding of biological entities and systems as well as engineering biology to address a variety of scientific or societal challenges. In this thesis, I describe my efforts, along with my colleagues across disciplines, to explore the boundaries of synthetic biology with physics and computer science to engineer molecules, materials, and magnetism in the biological context. In Chapter 2, I will demonstrate applying structural insights of a key metabolic protein in microbes toward targeted mutagenesis, enabling tailoring of the production of microbial biofuel molecules applicable as renewable energy supplies. In Chapter 3, I will discuss the application of directed evolution toward engineering inorganic nanomaterial synthesis, as well as their applications with cells. Specifically I will show that using the ubiquitous and important iron storage enzyme ferritin as template, random mutagenesis and magnetic selection could lead to more efficient iron sequestration and magnetic phenotype for cells with potential applications in noninvasive magnetic imaging and manipulation in biology. The directed evolution approach here is suitable due to lack of predictive understanding between properties of the protein complex such as its molecular composition and structure, with the properties of the functional, inorganic nanoparticle products. Lastly in Chapter 4, I will introduce a computational approach toward engineering with limited prior

knowledge by employing deep, artificial neural networks to learn directly from abundant protein sequence data, making accurate property predictions on new, unannotated proteins that can be validated by experiment. In the future, combining tools and ideas across disciplines as well as the core strategies of rational design, directed evolution, and computational prediction could accelerate our progress in engineering and understanding of biology.

**Acknowledgement**

The last few years have been a fun and challenging adventure, intellectually and emotionally. There has been a lot of up and downs, and I have been fortunate to have the support of a large and diverse group of people in the Silver Lab and at Harvard more generally.

First I want to acknowledge my PhD advisor Professor Pamela Silver, particularly for her confidence in my abilities and openness to my pursuits. I was able to join her lab to explore synthetic biology initially as someone without almost any relevant biology experience or knowledge going into my third year of PhD in applied physics. This opportunity has allowed me to both learn the new field quickly as well as to explore its boundaries with my previous domains of expertise in the physical sciences, doing truly inter-disciplinary work. Furthermore, beyond keeping me on track from a high perspective, Pam has granted me almost complete intellectual freedom to explore and execute my own ideas in the lab. Not only does this satiate my innate curiosity and drive, but this experience has given me valuable training as an independent investigator and critical thinker.

Next I want to acknowledge my advisor Dr. Jeff Way at the Wyss Institute. I joined the lab as a young, naïve grad student without much idea about biology initially, and Jeff was instrumental in getting me started quickly through sharing his vast knowledge and insights. As I developed quickly as a synthetic biologist, Jeff similarly allowed me to fully explore my own ideas and pursuits without hindrance. This has similarly helped me mature as an independent researcher.

I want to also acknowledge my colleagues with whom I have worked closely, including Dr. Wade Hicks who worked with me initially on biofuels and from whom I acquired most of my biological knowledge and expertise when I started. I want to

thank the undergraduates who have worked closely with me during the summers including Paola Lopez and Michael Giles. Their work ethics and intellectual curiosity contributed greatly to the research progress. I also want to thank Dr. David Riglar and Alexander (Sasha) Naydich for collaborations. More generally, I have constantly received great advice from almost all members of the Silver Lab. I want to thank in particular Joe Torella, Cameron Myhrvold, Tyler Ford, Steph Hays, Ryan Richardson, Gairik Sachdeva and Tobias Giessen for our frequent discussions and their support.

Outside of the Silver Lab, I want to acknowledge the professors and collaborators who have taught me a lot and given me valuable advice. I want to thank my thesis defense committee members including Professors Robert Westervelt, Sean Eddy, Joanna Aizenberg and Shmuel Rubinstein for their time and feedback. In particular, Professor Sean Eddy has provided many valuable suggestions throughout this thesis particularly for the work related to protein sequence analysis using neural networks. I also want to thank Professor Ronald Walsworth in physics, Debora Marks in computational systems biology and Aaron Grant in magnetic resonance imaging. The members of their groups, particularly Pauli Kehayias, Chenchen Luo, Gopal Varma who have devoted significant time and effort working together with me exploring challenging science. I also want to thank the technical staff members of the Wyss Institute, Harvard Medical School Electron Microscopy Facility, and Center for Nanoscale System who have taught me a lot and have always been helpful in resolving experimental issues.

Being in Silver lab is fun not only for research but also for its social atmosphere with its diverse group of people from different backgrounds with different interests. Over the years we have done many hikes, including one on summer

solstice in New Hampshire every year and one every fall to admire the New England foliage. Together, we have done many camping trips, game nights, Christmas parties, restaurant trips and ice-cream eating competitions, of which I am proud to be the winner for the last two years and also the current record holder as of 2016 with 52 scoops. Much thanks need to go to the tireless organizers for these events including Steph Hays, Tyler Ford, Marika Ziesack, Cameron Myhrvold, Alina Chan, Anna Chen, Matt Mattozzi, Gairik Sachdeva, Andy Shumaker, Avi Robinson-Mosher, Finn Stirling and Elena Schäfer. The lab would not have been such a fun place without their participation and hard-work.

Lastly I want to acknowledge my family and my friends outside of the Silver Lab for their support over the years.

# Table of Contents

## CHAPTER 1

# INTRODUCTION

**The Problems of Sustainability and Complexity**

Since the dawn of life on earth, Homo sapiens, out of all species of life, have excelled at discovering and changing the environment around us. We made tools. We built cities and empires. We even departed this planet and reached into outer space. Through science we have increased our knowledge of the laws that govern Nature, and through technology we have applied that understanding to better serve the needs of our species.

With the ever expansions of our population, now at 7.4 billion, and our appetite for improving our living standards through increased utilization of resources, we have caused increasing burden and disturbance to the environment in which we live. The question of sustainability naturally emerges: how can we continue at this rate of growth, in population, economy, or resource utilization, until it has to stop or even decrease because we will have exceeded our environment's "carrying capacity" of our species? Even more appalling to consider, could we push our environment far out of equilibrium so as to drastically and irreversibly decrease our habitability at least in the shorter term future and trigger an implosion of our civilization from its inherent instabilities?

Now is not the first time humanity has been faced with this daunting question. At the dawn of the Industrial Revolution, economist Thomas Malthus of England presented the idea of "Malthusian Trap": increased resource utilization would not result in increased welfare per capita due to increased pressures from population growth. Malthus wrote in "An Essay on the Principle of Population" in 1798 (in the midst of the first major upheaval in Europe, the French Revolution):[1]

*The power of population is so superior to the power of the earth to produce subsistence for man, that premature death must in some shape or other visit the*

*human race. The vices of mankind are active and able ministers of depopulation. They are the precursors in the great army of destruction, and often finish the dreadful work themselves.*

Fortunately, such dire predictions have not panned out entirely. Since then, despite catastrophes such as the World Wars, major famines and epidemics throughout parts of the world, overall the global population has increased dramatically as more and more people are being brought out of poverty to enjoy higher standards of life. This in large parts is credited to education (of women in particular), and to advances in science and technology that have enabled us to overcome barriers to better utilize the resources available to us and expand our "carrying capacity". In particular, the "Green Revolution" led by Norman Borlaug around the middle of the 20$^{th}$ century, which combined the application of advances in high-yield crop varieties, fertilizers, irrigation and mechanization in agriculture, resulted in dramatic increases in both yield and overall production outpacing population increases[2]. Thus the Malthusian trap had been averted.

As human beings improve beyond mere subsistence and attain mobility, entertainment and other enjoyments of life afforded by modern technology, the concern for sustainability expands beyond that of just food security. Energy, for example, is a common denominator in many of these human activities and pursuits. Even for the Green Revolution, many of the technologies such as fertilizers and mechanization have led to increased energy input. Technologies like synthetic biology can potentially offer solutions as will be explored in later sections. And this could impact not only food and energy but other resources that could become scarce, such as clean water, air and certain raw materials.

As science itself has evolved over the centuries with its world changing discoveries, many of the new or remaining problems that we are faced with both in science and society today are marked by increasing complexity. This is taking the premise that our understanding of the world, both physical and social, can be formulated and expressed in the language of logic, mathematics, as Sir Isaac Newton, one of the most influential figures of the “Scientific Revolution”, exemplified in his work *Philosophiæ Naturalis Principia Mathematica* (Mathematical Principles of Natural Philosophy). In physics, Newton first applied equations of motion and calculus to describe the dynamics of macroscopic objects. Centuries after Newton, mathematical theories such as statistical thermodynamics and electromagnetism have been developed to model the microscopic and the “intangible”. In the 20$^{th}$ century, two pillars of physics were erected with advances in relativity and quantum mechanics. In particular, quantum mechanics, unintuitive as it is, is the most precise theory that describes the physical world down to the microscopic level of the atom and the subatomic particles, where classical mechanics developed by Newton fails. Theoretically in our quantum mechanical world, since everything we know is composed of atoms or the fundamental particles, the theory of quantum mechanics could naively enable us to model, predict, and engineer everything, from molecules to materials to ever more advanced machines and systems that pervade our society.

However, such optimism had been met with practical difficulties, in particular with our limitations to compute. The core issue is complexity. As one composes atoms to make molecules or materials, the number of variables and degrees of freedom quickly increases in the model to deter clean, analytical solutions useful for prediction and engineering. Exceptions do arise, for example in cases such as the elegant arrangement of atoms in solid crystals, the symmetries afford great reduction

in complexity to give rise to highly accurate models that describe the properties of crystalline materials such as metals and semiconductors vital to our society and industries today[3]. However in most other areas complexity seems an insurmountable barrier to rapid progress. This is particularly true in biology, where the often lack of symmetries in the delicate arrangement of atoms into small molecules or the arrangement of building blocks such as nucleotides or amino acids into polymers and macromolecules uniquely convey information and function. Furthermore, interactions among constituents of a system, often approximated only to the lowest orders for theoretical feasibility in applying equations of physics, are prevalent in biology at all levels (molecules, cells, organisms), and could give rise to the "emergent" properties of the system as a whole. Modelling such complex, asymmetrical systems and interactions, even for a single protein with a few thousand atoms, with the precise laws of quantum mechanics is a daunting task even for the best computers today. Hence approximations have been made employing classical theories, such as in the approach to apply classical force fields in molecular dynamics. Dynamical equations with increasing levels of approximation are used for higher level systems and their interactions including molecules in a cell, cells in an organism, or individuals and populations interacting in a society. While making these approximations to details is sufficient for predictive modelling of certain systems, often the most difficult problems in the physical or social sciences do not offer deep insights when complexities are glossed over. Prominent examples include the general prediction of the fold and function of proteins, the interaction among molecules and cells in development and disease, the origin of human intelligence arising from the brain with its billions of neurons with their trillions of connections, and at a higher level the behavior of human agents and the interactions among

themselves and with the outside world to affect markets or environments, and consequently the financial or physical climate.

In recent years, rapid advances in our power and capacity to compute and the massively growing rate of data gathering across domains start to offer new approaches for tackling the challenges of complexity, in particular building intelligent machines and systems exceeding certain human capabilities to solve complex problems across fields. Such advances will be explored in the context of their application particularly to biology in the following sections.

**The Promise of Synthetic Biology and Artificial Intelligence**

Synthetic biology applies engineering principles and approaches to the study and application of biology. Biology is viewed as a system, analogous to a computer or a legal framework, that can be modified from its existing form or designed de novo following a set of principles. There are several factors that led to the rapid growth in the synthetic biology discipline at the dawn of this century[4,5], two of the most important are the proliferation of computers in society and the rapid development of genomics and molecular biology in science. This coincidence of events led naturally toward the view that the underlying design principles of a manmade system like a computer and a natural biological system are highly analogous: each is composed of software carrying information and hardware carrying out tasks. Hence like the computer, biological systems can be designed according to similar principles. Two core principles or concepts shared between synthetic biology and other engineering disciplines (e.g. electrical, chemical or social) are modularity and networks. These concepts will be explored in greater detail in a later section.

Since its beginning, synthetic biology has found application across a diverse array of industries. Synthetic biology approaches have been applied toward engineering microbes for drug[6] or renewable energy[7] production, plants for better crops[8], insects for disease control[9], and even human cell for drug production or disease treatment[10]. Some of these applications will be explored in detail in later chapters.

Around the same time as biologists and engineers were trying to create artificial cells[11] with the tools of synthetic biology, computer scientists have experienced another revolution in the field of Artificial Intelligence (AI) and machine learning. The term “artificial intelligence” was formally introduced at the Dartmouth

Summer Research Project on Artificial Intelligence" in 1956. In one definition by Nils J. Nilsson,[12]

*"Artificial intelligence is that activity devoted to making machines intelligent, and intelligence is that quality that enables an entity to function appropriately and with foresight in its environment"*

Work contributing to this field started long before 1956, in particular in Alan Turing's classic "Computing Machinery and Intelligence" where he first introduced the "Turing test" as a way to test if a computer can imitate human intelligence[12]. Later between the 1950s and 1970s, many fields blossomed within the artificial intelligence community such as computer vision, natural language processing, and importantly artificial neural networks starting from the invention of Rosenblatt's Perceptron inspired by the model of the biological neuron. However in the 1980s, the field of artificial intelligence could not demonstrate significant practical successes and lost interest and funding, experiencing an "AI Winter"[12]. Later in the 1990s and the start of the 21st century, driven by technological progress in data gathering (e.g. Internet, sensors) and computational power (e.g. processing, storage), AI started to demonstrate significant progress learning from data. Particularly in the last few years, deep-learning AI systems built upon large neural networks have managed to exceed human cognitive abilities in several areas, most notably in machine vision (e.g. models inspired by AlexNet[13]), knowledge representation (e.g. IBM Watson) and search and planning (e.g. Google DeepMind AlphaGo[14]).

As data and computational power continue to expand at accelerating pace, the data-driven artificial intelligence models continue to improve and expand their applications into domains outside of computer science. Fortunately biology has coincidentally experienced a data revolution from sequencing and high-throughput

experimentation particularly at single cell resolution[15], producing not only large volumes of data for gathering potential insights toward engineering of complex biological systems, but also toward potential improvement of biologically inspired AI systems such as the artificial neural network based on the biological insights. In later chapters, I will explore one specific area where concept and technologies of artificial intelligence could be applied to a long-standing challenge in biology.

But first, I will explore in the next three sections how the discipline of synthetic biology intersects with three other important fields encompassing the stories of my thesis work: energy, physics and computer science. Synthetic biology and these disciplines have been and will be ever more mutually reliant, and these beneficial interactions will provide many rich opportunities for rapid progress in science and technology into the future.

**Synthetic Biology and Energy**

Energy is the most fundamental unit of life. Without it, life cannot exist. And over billions of years of evolution, life of all forms have developed ways to extract and store it in reduced carbons, either directly produced from the physical elements (e.g. autotrophs) or extracted from other life (e.g. heterotrophs). By far, the sun has been the dominant energy source for life on Earth. Much of global terrestrial biomass production is attributable to the sun through photosynthesis, the conversion of sunlight and carbon dioxide into the energy-storing carbon molecules. This is true even for the human society today, as the currently dominant fossil fuels are derived from the products of photosynthesis millions of years ago. Among the renewables, photovoltaic or solar thermal directly convert sunlight into usable electricity. Wind, tidal, and hydroelectric energy all derive from the potential energy stored in the Earth climate system by solar irradiance. Two hours of sunlight on earth carries more energy than our current human consumption in an entire year[16]. Hence as our world is now trying to provide sustainable supply of energy for our ever increasing needs without potentially disastrous effects on our environment and climate, for example by releasing large quantities of stored carbon into the air through fossil fuel burning, we can look toward the sun with its large potential for renewable, sustainable and clean solutions.

Scientists across disciplines have engineered many solutions for capturing and converting solar energy directly or indirectly into usable and storable forms for humans. In physics, the photovoltaic was first practically developed in the 1950s by the Bell Labs[17], leveraging the photo-excitability of certain semiconducting materials like silicon, to directly convert sunlight into electricity that can be transmitted, used or stored in batteries. Given the relatively low energy density of batteries for storage

compared to chemical fuels[18], chemists have also worked out many catalysts that store the electrical potential energy into the bonds of reduced chemicals[19], most notably hydrogen gas from splitting water[20]. And biologists have engineered plants and microbes for increased production of biomass or other compounds[7,8,21]. With the advent of synthetic biology, these engineering efforts can accelerate to obtain even better technologies for renewable energy production.

One of the key practical challenges for any energy technology, however, is cost. Fossil fuels are energy rich and can be extracted and refined with competitive economics using current technologies (e.g. horizontal drilling, fracking). Even though the costs of renewables like solar photovoltaic have been decreasing dramatically with their increasing adoption in our society, their total contribution to global energy consumption still amounts to much less than the fossil fuels (e.g. coal, oil, natural gas) due to higher costs[22]. Current renewable energy sources that depend indirectly on solar such as wind and hydroelectric benefit from high power density and are already approaching or beating price parity with electricity generation from fossil fuels[23]. However, direct conversion of solar energy is still difficult to apply widely largely due to intermittency issues and the relatively low energy density of solar irradiance, at roughly 1000 Watts per square meter. Commercial photovoltaics can capture and convert 18% of that energy into electricity, which is at least 2 to 3 times better than the best of photosynthesis[24]. In addition to this lower efficiency of conversion, biological systems are much less durable compared to inorganic semiconductors and need to be constantly renewed to remain functional. This is partially due to biology's susceptibility to heat, radicals, radiation (e.g. ultraviolet that causes DNA damage), and other biological factors such as mutations and disease (e.g. viral infection).

Despite the above limitations, biology could offer two key advantages over purely inorganic energy conversion approaches, the natural ability to store energy in high energy density molecules, and the largely unexplored potential of synthetic biology to optimize and program biological entities via evolution or design to create self-replicating and sustainable energy production systems without the capital and energy-intensive industrialized manufacturing processes for inorganic devices like the photovoltaics. Chapter 2 of this thesis will explore how microbes could be engineered by design to tailor production of high energy density fuel molecules. The later chapters will explore other strategies employing sensors, directed evolution, or computational approaches toward better design of biological systems. And as we improve our abilities to program and engineer cells using ever improving tools of synthetic biology, it would then be possible to address all the issues related to efficiency and reliability and optimize biology toward renewable production of energy that would become cost competitive and widely applicable.

## Synthetic Biology and Physics

Synthetic biology as an engineering discipline serves as a vital bridge between several core disciplines, particularly biology and physics. This bridge has enabled mutual trade of information and technologies that helped progress in both disciplines.

Synthetic biology has created an influx of mathematical, computational and physical tools developed first in physics to help accelerate the study and application of biology. For example, the equations that model switches, oscillations, and circuits are frequently utilized in synthetic biology to design and test biological circuits[4,5]. Examples include not only dynamical differential equations describing the interaction among a few molecules as in a sensor or counter, but also large scale modeling of whole cell metabolism, of which the equilibrium model of Flux Balance Analysis (FBA)[25] is a well-known example. Additionally, the study of extremely complex networks such as the cognitive neural network or interactome of biomolecules in development or disease often rely on similar mathematical and computational tools that have been developed first in physics. Zooming into the network at the node level, the study of the behavior of a single cell such as the neuron was first famously modelled as an electrical circuit by Alan Hodgkin and Andrew Huxley[26] who later received the Nobel Prize in Physiology or Medicine in 1963 for this work. On the other hand, atomistic simulations of biomolecules such as individual proteins using Molecular Dynamics (MD)[27] via force-fields and High Performance Computing (HPC) have provided insights into nano-scale interactions of biology at a fundamental scale often inaccessible by direct experimentation. Meanwhile, experimentalists in biology have made great advances into studying biology with high resolution thanks to a wide array of tools and technologies adopted from physics. The application of x-ray

diffraction in physics toward biological crystals has allowed atomistic-scale determination of the structure of important biological molecules such as the DNA or proteins and in turn revolutionized our understanding of molecules, the fundamental units of a biological system for any synthetic biology effort. In recent years these fields are further supplemented by developments in extremely high-resolution electron microscopy (EM) to resolve larger, difficult to crystallize proteins and complexes[28]. The cryogenic and high vacuum conditions of these microscopes as well as their aberration-corrected optics for achieving high resolution were techniques that have been developed for decades in physics to study condensed matter or astronomy. And aside from probing individual molecules in isolated and well-controlled conditions as is often convenient for physics, technologies from physics, particularly optics, electronics and magnetics have also enabled investigation of living systems in their native contexts via interactions with electromagnetic fields. For instance, fluorescence microscopy has revolutionized cell biology and in recent years further advanced the resolution below the light-diffraction limit using a variety of optical or computational tricks[29,30]. Moreover in recent years, opto-genetics has conversely enabled the control of biological signals such as action potentials via light[31,32]. Synthetic biology has played a vital role in these developments through the discovery, engineering and application of light sensitive molecules[33] in various biological contexts as reporters or actuators. Using just the electric field component, biologists have studied and manipulated neurons to understand the function of the brain. On the other hand using the tissue penetrating magnetic field component, magnetic resonance imaging (MRI) has enabled noninvasive studies of the function, or malfunction of critical human organs such as the brain. In all of these endeavors, synthetic biology has already, or could

potentially further enhance the physical techniques by engineering more sensitive, selective and accurate biological constructs toward the study of biology.

In the other direction, synthetic biology has generated new questions, insights, materials and model systems for studies that would expand our knowledge and application of physics. For example, to unravel the high complexity of most non-equilibrium, non-homogeneous biological systems would require advances in theory and modeling, such as in statistical thermodynamics. Synthetic biology tools allow manipulations and control of such systems for study, either by a top-down strategy via genetic alterations (e.g. CRISPR/CAS9[34,35]), or bottom up by building a cell from scratch according to blueprint[11]. On the other hand, many natural or engineered biological systems can serve as inspirations for engineering new physical tools. For example, the coloring of certain butterfly wings due to properties of their photonic crystal-like microstructure that selectively interacts with different wavelengths of light could lead to bioinspired engineering of new photonic materials for applications in optics[36]. Furthermore, synthetic biology applied toward the synthesis of nanoscale materials via biological pathways could also provide not only new types of materials for physical characterizations, but also insights on how such materials interact and could be rendered useful within biological systems. One example is the physics of magnetoreception, or how magnetic field could interact with biological components to enable certain biological species to orient and navigate in their natural environments. Two prominent theories exist, based on either the presence of magnetite particles that could activate by force via magnetic field[37], as was found with magnetotactic bacteria, or radical-based transduction where a pair of radical spins generated within proteins such as a crytochrome when exposed to blue light could be sensitively affected by ambient magnetic field to affect the duration of crytochrome's light-

activated state[38]. Synthetic biology could allow generation of simple model systems to test the plausibility of these hypotheses, for example by synthesizing magnetite particles in vivo and linking them to other cellular components for function. Such efforts, if successful, could open up a new realm of biological control via the noninvasive tissue-penetrating magnetic field. Chapter 3 of the thesis will explore these ideas pertinent to synthesizing nanomaterials and enabling magnetic control in cells in greater detail.

**Synthetic Biology and Computer Science**

Similar to its relationship to physics, synthetic biology's relationship with computer science is also a two-way street, allowing the two disciplines to exchange concepts and tools and to generate insights that mutually benefit.

Like computers, biological systems such as the cell can receive, process, store and output information. Hence much could be borrowed directly from computer science and applied toward the effort to engineer the biological computer. In the most literal sense, synthetic biologists have re-created Boolean-logic circuits that underlie modern computers in the "genetic-wiring" of the cell with a complete set of logic gates, low level hierarchical layering, and memory[39,40]. At a higher level, assembly language Verilog has been adapted to make abstract the design of genetic circuits in cells. Beyond these achievements, Nature itself has written its own programs into its diversity of life-forms over the last 3.8 billion years of evolution. Each cell, as a biological computer, stores its own operating system in its genetic codes. The size of the genetic code for a typical human cell at 3 billion letters still exceeds our current ability to write from scratch[41]. On the other hand, broadly applicable and targeted genome editing technologies like CRISPR/CAS9, similar to the "find and replace" functions in computer programs, have dramatically accelerated the process of biological hacking. Nowadays, hacking in synthetic biology is often accomplished literally via viruses to deliver the extra codes (DNA pieces), either by introducing them in the form of plasmids that could be replicated and maintained as standalone scripts, or by direct mutagenesis or integration into the native host genome. Given the large amount of code that Nature has already written into life and our still very limited understanding of its syntax, meaning and its entire network of interactions and dependencies, one must then appreciate the difficulty faced by synthetic biologists to

hack into the system often without full clarify and still achieve the desired result. One could appreciate this more when considering the difficulty and significant efforts devoted toward debugging large operating systems for physical computers that human have once designed according to our own understood logic starting from a clean slate.

Fortunately, two important concepts widely used in computer science to partially manage and make use of this complexity can be similarly applied to synthetic biology. The first is modularity. An example of this in computer science can be found in the widely-adopted practice of “object-oriented programming”, whereby codes are organized into objects consisted of fields for data and methods for procedures. The objects can be instantiated, passed, inherited and reused. Similar efforts have been undertaken in synthetic biology to create a standardized list of parts that can be combined to form modules. One example is a basic gene-expression module consisting a promoter, ribosome binding site (RBS), gene coding sequence with start and stop codons to inform translation and a transcriptional terminator. Moreover, the diversity of Nature’s existing programs provides a rich resource of parts to be incorporated into the genetic modules. Well known examples of such parts include reporters (e.g. fluorescent proteins), DNA binding proteins that modify other procedures (e.g. Cas nucleases), enzymes that carry out important chemical transformations for metabolism (e.g. cytochrome P450) and more. Nonetheless, our understanding of the function of even just the individual parts today is still very limited due largely to our inability to generally predict protein folding, which is highly correlated to function, directly from genetic sequences. Chapter 4 of this thesis will investigate a solution approach to this by loosely applying the second major concept that can be borrowed from computer science, networks. Networks are composed of the individual nodes represented by the parts or modules, along with

their connectivity and interactions. Important properties of a network include its architecture, robustness as well as its emergent behavior. Taking for instance an artificial neural network for machine learning as will be examined in greater detail in Chapter 4, its architecture describes its layers and connectivity, robustness describes its ability to maintain function with changes in internal or environmental factors such as the loss of a fraction of connections (e.g. “Dropout” in Deep Learning[42]), and the emergent behavior is the ultimate output or purpose arising from the collective action of the parts, for example detecting a cat in an image. These concepts translate directly to biology where network is deeply endogenous especially when considering that of genetic and protein interactions or the cellular network of the neuronal brain. This close analogy would allow much theoretical and modelling developments from math and computer science to be applied toward better understanding of the important biological networks.

In the reverse direction, the richness of complexity and innovations in the important biological networks such as the brain could provide much inspiration for improved or novel computational architectures and algorithms. Treating the human brain as a network, its architecture involving billions of neurons and their trillions of connections is still largely unknown, leading currently to significant efforts for its mapping[43,44]. On the other hand, the human brain is in some ways highly robust, with hemispherectomy as an extreme example where patients with one cerebral hemisphere removed can still maintain cognitive function. Finally, the brain’s emergent behaviors such as cognition, consciousness and intelligence still defy full understanding or even sufficient definitions. Nonetheless, even our little understanding of this biological cellular network has already reaped rich rewards in artificial intelligence in recent years, most notably in the application of artificial neural

networks for machine learning. One particular success story has been the development of deep convolutional neural networks (ConvNet) inspired by organizations of the biological visual cortex[45]. ConvNet processes raw image inputs in a hierarchical fashion across its many layers while employing extensive filtering to eventually derive meaning from the image (e.g. dog). ConvNet has massively reduced the number of parameters of a naïve, fully connected neural network model that is extremely difficult to train and generalize to new data (i.e. “over-fitting”). Meanwhile to achieve greater robustness as in biological systems, the “Dropout” technique which randomly severs neural connections during training has also dramatically improved prediction performance, altogether to exceed human level of accuracy. Despite these achievements, there are still many others human cognitive abilities, notably creativity, that are beyond current physical computers’ reconstitutions. Furthermore, the human brain is orders of magnitude more energy efficient using only around 12 watts of power[46] while passing information at rates far less than the speed of light unlike digital communication. Hence with synthetic biology and its genetic tools for probing the biological cognitive networks, there is hope to gain much more insights about these biological networks’ architecture, robustness and emergent behaviors that could lead to similar or even more powerful and efficient constitutions outside of the biological systems in the future.

## Chapter Outline

Each of the following three chapters will examine closely the relationship between synthetic biology and the fields of energy, physics and computer science with additional data, analysis, and suggestions for future directions. Chapter 2 will explore first how protein engineering could be applied toward the central metabolism of microbes to enable tailored production of precursors for fuel molecules for renewable energy. Chapter 3 will explore the applications of genetic sensors and directed evolution for engineering biological constructs that enable production of inorganic nanomaterials as well as for biological systems to interact with magnetic fields, with potential applications in noninvasive diagnostics and signal transduction. Chapter 4 will explore the use of artificial neural networks for accurate prediction of protein function directly from sequence, particularly in the context of iron mineralizing proteins examined experimentally and several other classes of protein function.

## References


1. Maltus, Thomas Robert. *An essay on the principle of population.* Vol. 1. Cosimo, Inc., 2006.

2. "The Nobel Peace Prize 1970 - Presentation Speech". *Nobelprize.org.* Nobel Media AB 204. Web. 18 Oct 2016. <http://www.nobelprize.org/nobel_prizes/peace/laureaetes/1970/press.html>

3. Ashcroft, N. & Mermin, D. *Solid State Physics*. (Cengage Learning, 1976).

4. Collins, J. J., Gardner, T. S. & Cantor, C. R. Construction of a genetic toggle switch in Escherichia coli. *Nature* **403,** 339–342 (2000).

5. Elowitz, M. B. & Leibler, S. A synthetic oscillatory network of transcriptional regulators. *Nature* **403,** 335–338 (2000).

6. Ro, D. *et al.* Production of the antimalarial drug precursor artemisinic acid in engineered yeast. *Nature* **440,** 3–6 (2006).

7. Peralta-Yahya, P. P., Zhang, F., del Cardayre, S. B. & Keasling, J. D. Microbial engineering for the production of advanced biofuels. *Nature* **488,** 320–8 (2012).

8. Lin, M. T., Occhialini, A., Andralojc, P. J., Parry, M. A. J. & Hanson, M. R. A faster Rubisco with potential to increase photosynthesis in crops. *Nature* **513,** 547–550 (2014).

9. Hammond, A. *et al.* A CRISPR-Cas9 gene drive system targeting female reproduction in the malaria mosquito vector *Anopheles gambiae*. *Nat. Biotechnol.* **34,** 78–83 (2016).

10. Maeder, M. L. & Gersbach, C. A. Genome-editing Technologies for Gene and Cell Therapy. *Mol. Ther.* **24,** 430–446 (2016).

11. Gibson, D. G. *et al.* Creation of a Bacterial Cell Controlled by a Chemically Synthesized Genome. **329,** 1–6 (2010).

12. Stone, P., Brooks, R., Brynjolfsson, E., Calo, R., Etzioni, O., Hager, G., Hirschberg, J., Kalyanakrishnan, J., Kamar, E., Kraus, S., Leyton-Brown, K., Parkes, D., Press, W., Saxenian, A., Shah, J., Tambe, M., & Teller. A. "Artificial Intelligence and Life in 2030." One Hundred Year Study on Artificial Intelligence: Report of the 2015-2016 Study Panel, Stanford University, Stanford, CA, September 2016. Doc: http://ai100.stanford.edu/2016-report. Accessed: September 6, 2016.

13. Krizhevsky, A., Sutskever, I. & Hinton, G. E. ImageNet Classification with Deep Convolutional Neural Networks. *Adv. Neural Inf. Process. Syst.* 1–9 (2012). doi:http://dx.doi.org/10.1016/j.protcy.2014.09.007

14. Silver, D. *et al.* Mastering the game of Go with deep neural networks and tree search. *Nature* **529,** 484–489 (2016).

15. Gawad, C., Koh, W. & Quake, S. R. Single-cell genome sequencing: current state of the science. *Nat Rev Genet* **17,** 175–188 (2016).

16. Moriarty P. & Honnery D. What is the global potential for renewable energy? *Renewable and Sustainable Energy Reviews*. **16**, 244-252, (2012).

17. Chapin, D. M., Fuller, C. S. & Pearson, G. L. A new silicon p-n junction photocell for converting solar radiation into electrical power [3]. *J. Appl. Phys.* **25,** 676–677 (1954).

18. Energy On a Sphere. *National Oceanic and Atmospheric Administration* Available at: https://sos.noaa.gov/Datasets/dataset.php?id=579 (Accessed: 18th October 2016)

19. Lewis, N.S., Nocera, D. G. Powering the planet: Chemical challenges in solar energy utilization. *Pro. Natl. Acad. Sci. U. S. A.* **103,** 15729-35 (2007).

20. Walter, M. G. *et al.* Solar Water Splitting Cells. *Chemical Reviews*. **110(11),** 6446–6473 (2010).

21. Torella, J. P. *et al.* Tailored fatty acid synthesis via dynamic control of fatty acid elongation. *Proc. Natl. Acad. Sci. U. S. A.* **110,** 11290–5 (2013).

22. International Energy Agency. Key world energy statistics. *International Energy Agency*, (2016).

23. EIA, US. Annual Energy Outlook 2016. *US Energy Information Administration, Washington, DC* (2016).

24. Blankenship, R. E. *et al.* Comparing Photosynthetic and Photovoltaic Efficiencies and Recognizing the Potential for Improvement. *Science*. **332,** 805–810 (2011).

25. Orth, J. D., Thiele, I. & Palsson, B. Ø. What is flux balance analysis? *Nat. Biotechnol.* **28,** 245–8 (2010).

26. Hodgkin, A. L. & Huxley, A. F. A quantitative description of membrane current and its application to conduction and excitation in nerve. *J. Physiol.* **117,** 500–544 (1952).

27. Karplus, M. & McCammon, J. A. Molecular dynamics simulations of biomolecules. *Nat. Struct. Biol.* **9,** 646–652 (2002).

28. Callaway, E. The Revolution Will Not Be Crystallized. *Nature* **525,** 172–174 (2015).

29. Rust, M. J., Bates, M. & Zhuang, X. W. Sub-diffraction-limit imaging by stochastic optical reconstruction microscopy (STORM). *Nat Methods* **3,** 793–795 (2006).

30. Betzig, E. *et al.* Imaging Intracellular Fluorescent Proteins at Nanometer Resolution. *Science*. **313,** 1642–1646 (2006).

31. Deisseroth, K. Optogenetics. *Nat. Methods* **8,** 26–29 (2011).

32. Boyden, E. S., Zhang, F., Bamberg, E., Nagel, G. & Deisseroth, K. Millisecond-timescale, genetically targeted optical control of neural activity. *Nat. Neurosci.* **8,** 1263–8 (2005).

33. Heim, R., Cubitt, A. B. & Tsien, R. Y. Improved Green Fluorescence. *Nature* **373,** 663–664 (1995).

34. Cong, L. *et al.* Multiplex Genome Engineering Using CRISPR/Cas Systems. *Science.* **339,** 819–824 (2013).

35. Jinek, M. *et al.* A programmable dual-RNA-guided DNA endonuclease in adaptive bacterial immunity. *Science.* **337,** 816–822 (2012).

36. Vukusic, P. & Sambles, J. R. Photonic structures in biology. *Nature* **424,** 852-855 (2003).

37. Kirschvink, J. L., Walker, M. M. & Diebel, C. E. Magnetite-based magnetoreception. *Curr. Opin. Neurobiol.* **11,** 462–467 (2001).

38. Ritz, T., Adem, S. & Schulten, K. A model for photoreceptor-based magnetoreception in birds. *Biophys. J.* **78,** 707–718 (2000).

39. Bonnet, J., Yin, P., Ortiz, M. E. & Endy, D. Amplifying Genetic Logic Gates. *Science*. **340,** 599-603 (2013)

40. Siuti, P., Yazbek, J. & Lu, T. K. Synthetic circuits integrating logic and memory in living cells. *Nat. Biotechnol.* **31,** 448–52 (2013).

41. Boeke, J. D. *et al.* The Genome Project – Write. *Science*. **353**, 126-7 (2016).

42. Srivastava, N., Hinton, G. E., Krizhevsky, A., Sutskever, I. & Salakhutdinov, R. Dropout : A Simple Way to Prevent Neural Networks from Overfitting. *J. Mach.*

*Learn. Res.* **15,** 1929–1958 (2014).

43. Human Connectome Project. (2016). Available at: http://www.humanconnectomeproject.org/. (Accessed: 25th September 2016)

44. The BRAIN Initiative. (2016). Available at: https://www.braininitiative.nih.gov/. (Accessed: 25th September 2016)

45. LeCun, Y. *et al.* Deep learning. *Nature* **521,** 436–444 (2015).

46. Clarke, D.D. & Sokoloff, L. Regulation of Cerebral Metabolic Rate. In: Siegel, G.J., Agranoff B.W>, Albers, R.W., et al., editors. Basic Neurochemistry: Molecular, Cellular and Medical Aspects. 6th edition. Philadelphia: Lippincott-Raven; 1999. Available from: https://www.ncbi.nlm.nih.gov/books/NBK28194/

## CHAPTER 2

# ENGINEERING ACYL CARRIER PROTEIN TO ENHANCE PRODUCTION OF SHORTENED FATTY ACIDS

## Preface

This was work conducted mainly in collaboration with Dr. Wade Hicks and Jeffrey Way. This work was published in 2016: *Biotechnology for Biofuels* 2016 **9**:24. Supporting materials for this chapter are found in Appendix A.


## Abstract

**Background**: The acyl carrier protein (ACP) is an essential and ubiquitous component of microbial synthesis of fatty acids, the natural precursor to biofuels. Natural fatty acids usually contain long chains of 16 or more carbon atoms. Shorter carbon chains, with increased fuel volatility, are desired for internal combustion engines. Engineering the length specificity of key proteins in fatty acid metabolism, such as ACP, may enable microbial synthesis of these shorter chain fatty acids.

**Results**: We constructed a homology model of the *S. elongatus* ACP, showing a hydrophobic pocket harboring the growing acyl chain. Amino acids within the pocket were mutated to increase steric hindrance to the acyl-chain. Certain mutant ACPs, when over-expressed in *E. coli*, increased the proportion of shorter chain lipids; I75W and I75Y showed the strongest effects. Expression of I75W and I75Y mutant ACPs also increased production of lauric acid in *E. coli* that expressed the C12-specific acyl-ACP thioesterase from *Cuphea palustris*.

**Conclusions**: We engineered the specificity of the ACP, an essential protein of fatty acid metabolism, to alter the *E. coli* lipid pool and enhance production of medium-chain fatty acids as biofuel precursors. These results indicate that modification of ACP itself could be combined with enzymes affecting length specificity in fatty acid synthesis to enhance production of commodity chemicals based on fatty acids.

## Introduction

With the continuous rise in global energy needs and adverse climate changes, development of cleaner and renewable alternatives to fossil fuels has become paramount. Microbial synthesis of biofuels is an attractive, renewable alternative to fossil fuels.[1–3] Organisms naturally synthesize large quantities of fuel-like hydrocarbons in the form of lipids, which are used in cell membranes and other molecules. In microbes, the end products of fatty acid metabolism are long acyl-chains consisting mostly of 16-18 carbons. When extracted for fuels, these long chain carbon molecules remain solid at room temperature and lack favorable physical properties such as higher volatility and lower viscosity. Such properties are characteristic of medium-length (8-12) carbon chains used ubiquitously in fuels for vehicles and jets.

Previous work on the biological synthesis of medium length fuel precursors has employed thioesterase enzymes with medium-length chain specificity to release free fatty acids (FFA) from intermediates in fatty acid synthesis.[4–7] Here we employ a complementary strategy to bias FFA synthesis toward shorter chains by engineering acyl-carrier protein (ACP), an essential protein and key component of fatty acid metabolism. In fatty acid synthesis in bacteria and plants, ACP is attached to the acyl chain and presents it to the other enzymes during successive cycles of elongation and reduction (Figure 1).[8–11] ACP is a small (~9kB), acidic (pI = 4.1) protein abundant in the cytoplasm, constituting about 0.25% of all soluble proteins in *E. coli*.[8] The structure of ACP is highly conserved even among variants with low sequence similarity. Four alpha helices, with the major helices I, II and IV running parallel to each other, enclose a hydrophobic pocket that harbors the acyl-chain; minor helix III runs perpendicular to these (Figure 2). The acyl chain is connected to a 4-phosphopantetheine modification at a conserved serine and enters the hydrophobic cavity between helices II and III.

Roujeinikova et al. solved the structures of *E. coli* ACP attached to C6, C7, and C10 fatty acids.[12] In each case, the distal end of the fatty acid terminates in a deep pocket within the protein near Ile72 (corresponding to Ile75 of the *S. elongatus* ACP), with the phosphopantetheine group also entering the pocket to varying degrees. Acyl chains up to 8 carbons are fully bound within the pocket, with the thioester bond sequestered in the core of the protein.[8,12–14] We therefore hypothesized that the size of ACP's hydrophobic pocket influences the composition of lipid lengths in a cell. As the acyl chain grows to a length of around 16, the thioester bond becomes more fully solvent exposed, which may facilitate cleavage by downstream processing enzymes.

We found that over-expressing certain mutant ACPs altered the composition of the cellular lipid pool and increased production of certain medium chain fatty acids. Our findings could be useful for microbial production of transportation biofuels based on metabolically engineered pathways.

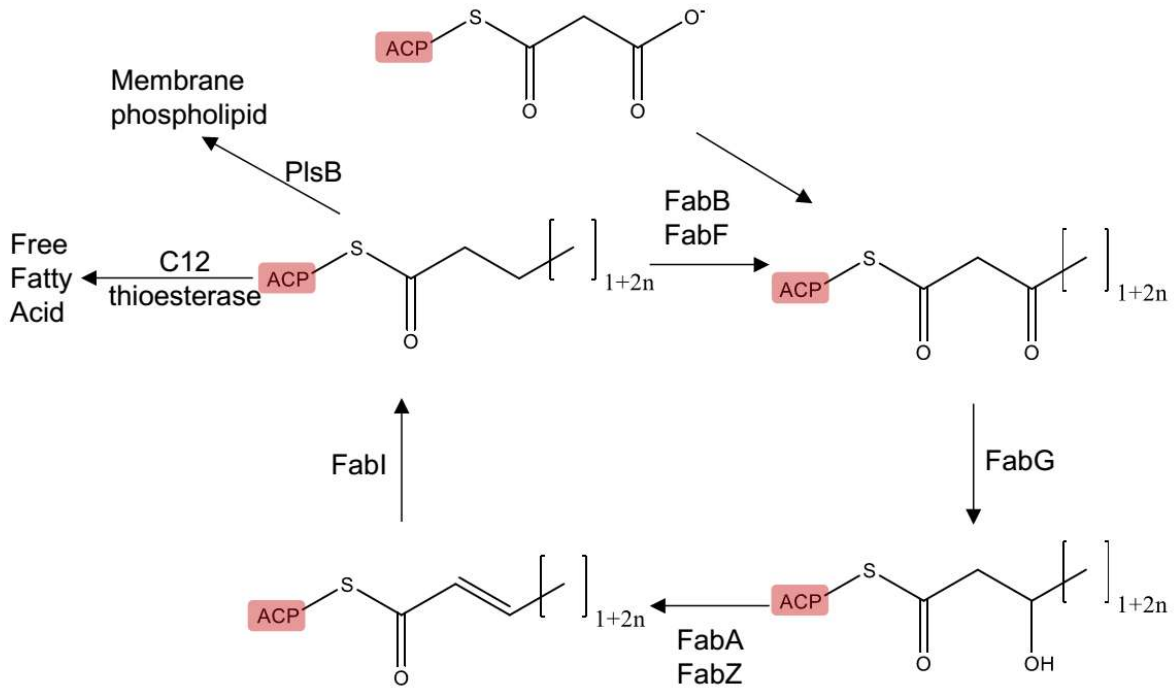



**Figure 1 Overview of Fatty acid Synthesis.** Fatty acid synthesis proceeds through iterative cycles of elongation. In each cycle, the acyl-chain is extended by 2-carbons using a malonyl-ACP as a carbon donor (by FabB or FabF) and subsequently reduced into a saturated chain (by FabG, FabA, FabZ and FabI). From the first 2-carbon malonyl-ACP to the final length fatty acid processed through this cycle, the hydrophobic acyl-chain is attached to and shielded by the ACP instead of existing in a free form.

## Results and Discussions

To enhance production of medium-chain fatty acids, we constructed mutants of ACP designed to decrease the acyl-chain pocket size (Figure 2). Variants of the cyanobacterial (*S. elongatus*) ACP were expressed in an *E. coli* host. We chose *S. elongatus* ACP due to its natural compatibility with recently discovered enzymes of the cyanobacterial alkane biosynthesis pathway[15], which could enable microbial synthesis of fatty alcohol or alkanes via the acyl-acyl carrier protein reductase and aldehyde decarbonylase enzymes native to cyanobacteria. The native *E. coli* ACP gene was left intact, as we found that its knockout could not be rescued by complementation from expression of wild-type *E. coli* ACP encoded on an IPTG inducible plasmid, likely due to sensitivity in timing and level of expression for the essential ACP in the cell during growth and replication (data not shown). To determine which hydrophobic residues of *S. elongatus* ACP lined the inner, acyl-chain pocket, we constructed a structural homology model using SWISS-MODEL server using the published crystal structure of *E. coli* ACP bound to a C10 fatty acyl chain (2FAE, sequence identity of 62.67%) as a template, achieving a good Global Model Quality Estimation (GMQE) score of 0.8 (Figure 2). We constructed nine single amino acid mutants by exchanging small hydrophobic side chain residues, such as isoleucine or leucine, with bulkier hydrophobic side-chains such as phenylalanine, methionine, tyrosine or tryptophan. ACPs initially fold into an inactive apo state. Conversion to the active holo state is achieved through post-translational modification whereby 4'-phosphopantetheine is transferred from co-enzyme A (CoA) to a specific serine residue on the apo-ACP (Ser39 on *S. elongatus* ACP).[8,16] ACP over-expression may reduce the CoA pool and lead to toxic accumulation of apo-ACP, which inhibits sn-glycerol-3-phosphate acyltransferase[16,17] , so as a quick check for functional expression of recombinant

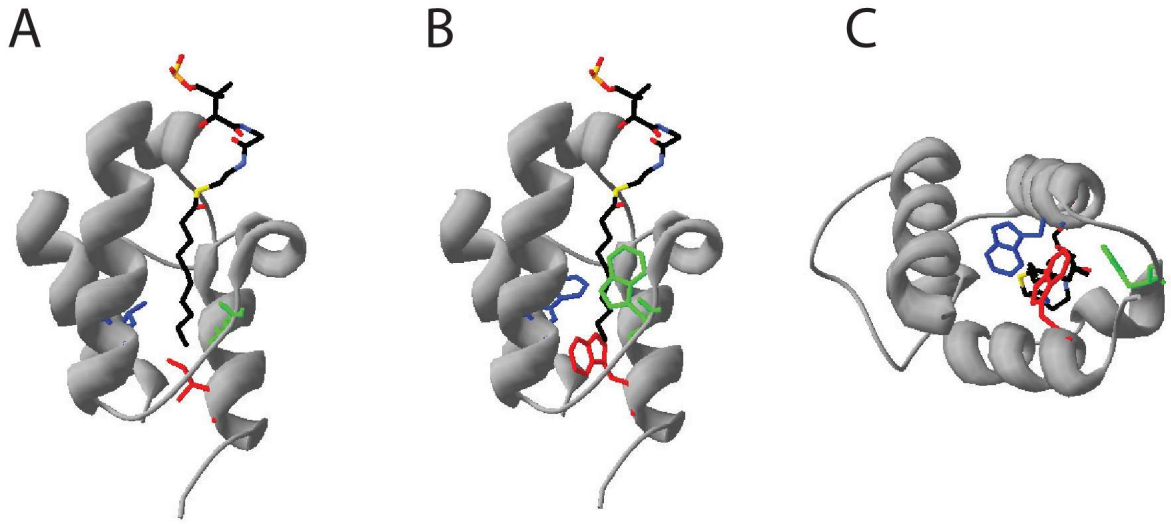



**Figure 2 Se-ACP Structural Homology Models with WT and Mutant Residues. A.** Homology model of Se-ACP bound to a C10 acyl-chain is shown. Highlighted in blue (residue 49), green (residue 57) and red (residue 75) are small hydrophobic amino acids lining the WT ACP pocket, Leu, Ile, and Ile, respectively. Each residue was separately mutated to a bulkier hydrophobic amino acid: methionine, tyrosine or tryptophan in order to induce steric hindrance and favor shorter-chain fatty acid synthesis. **B.** For illustration, a homology model with all three residues of interest mutated to tryptophan shows how each side chain might be positioned when mutated separately. Trp75 (red) extends closest to the acyl-chain terminus. **C.** Looking up through the axis of the acyl chain from the bottom perspective of the ACP pocket, Trp75 (red) is more directly in line with the acyl chain, as compared to the other mutant residues. This substitution appears to introduce direct steric hindrance to the acyl chain, while Trp at position 49 or 57 does not.

ACPs, we measured culture growth kinetics over 15 hours. Compared to no-plasmid or uninduced controls, cells expressing wild-type ('WT') *E. coli* ACP (Ec-ACP), WT *S. elongatus* ACP (Se-ACP), or mutant Se-ACPs all showed suppressed growth at low levels of ACP expression and worsened at higher expression levels (Figure S1), suggesting that these recombinant cyanobacterial ACPs were expressed and properly folded.

To analyze the effect of mutant Se-ACPs on lipid pools, we used gas chromatography mass spec (GC-MS) to characterize fatty acid methyl esters (FAMEs) derived from lipid pools in Se-ACP overexpressing cells. We compared ratios of FAME peak areas for each sample to minimize effects of differences in growth and sample extraction. We detected peaks for FAMEs derived from the naturally most abundant palmitic acid (C16) and the shorter, less abundant myristic acid (C14) and quantified these peaks in all sample spectra and calibrated to molar concentrations based on a standard curve (Figure S6). Together, C14 and C16 accounted for >90% of total fatty acids extracted in all samples (Figure S5). The concentration ratios of C14 to C16 were calculated and compared across controls and cells expressing Se-ACP point mutants. For all uninduced samples, the C14:C16 ratio was around 0.1 (Figure 3A). After induction, only the I75W and I75Y Se-ACP mutants demonstrated a statistically significant increase in the C14:C16 ratio relative to cells expressing WT Se-ACP: the mutants respectively caused 3- and 2.7-fold increases ($p<0.05$, two-tailed student-t test for either the "C14:C16 ratio" of the induced or the "Induced C14:C16 over Uninduced" ratio; Figure 3), indicating that their lipid pools had shifted toward shorter acyl chains. Mutants that replaced Leu49 or Ile57 did not increase the proportions of shorter fatty acids compared to over-expressing WT ACPs. The side chain of isoleucine 75 is positioned in the hydrophobic pocket close to the terminus of

the acyl-chain, more so than residues 49 and 57, which contact the side of the acyl chain (Figure 2A).[12] Mutating Ile75 to phenylalanine or methionine may cause slight shifts in lipid pool chain-length composition (Figure 3). Homology modeling indicated that the Tyr75 and Trp75 sidechains protrude roughly two carbon-carbon bond distances further into the hydrophobic acyl-chain pocket than an isoleucine at this position (Figure 2B and 2C; only I75W shown). Therefore, I75W and I75Y Se-ACP mutants may directly hinder elongation from C14 to C16 in fatty acid synthesis and skew the fatty acid pool towards shorter chain lengths.

To explore the potential to further skew cellular lipids towards short chain lengths, particularly those shorter than 14 carbons long, we introduced secondary point-mutations in addition to the Se-ACP I75W or I75Y mutations based on knowledge of the predicted molecular structure from homology model. Amino acids with small hydrophobic side-chains such as isoleucine, valine or alanine were exchanged for a bulkier methionine, a polar glutamine, or a hydrophilic arginine. Double mutant Se-ACPs did not significantly increase the C14:C16 ratio beyond either single I75W or I75Y mutation alone (Figure S3), and did not cause observable production of chains shorter than C14.

As an additional control, the Se-ACP serine 39 residue, which is post-translationally modified with 4-phosphopantethene, was mutated to alanine (S39A), thereby generating an inactive, obligate apo-ACP. Overexpressing this inactive ACP resulted in similarly low C14:C16 ratio compared to WT (Figure 3). Growth was suppressed by over-expressing this mutant protein, suggesting that the protein was correctly folded[16,17].

These result indicated that expression of mutant ACPs could be used to enhance production of a medium-chain fatty acid. To explore conditions for optimal

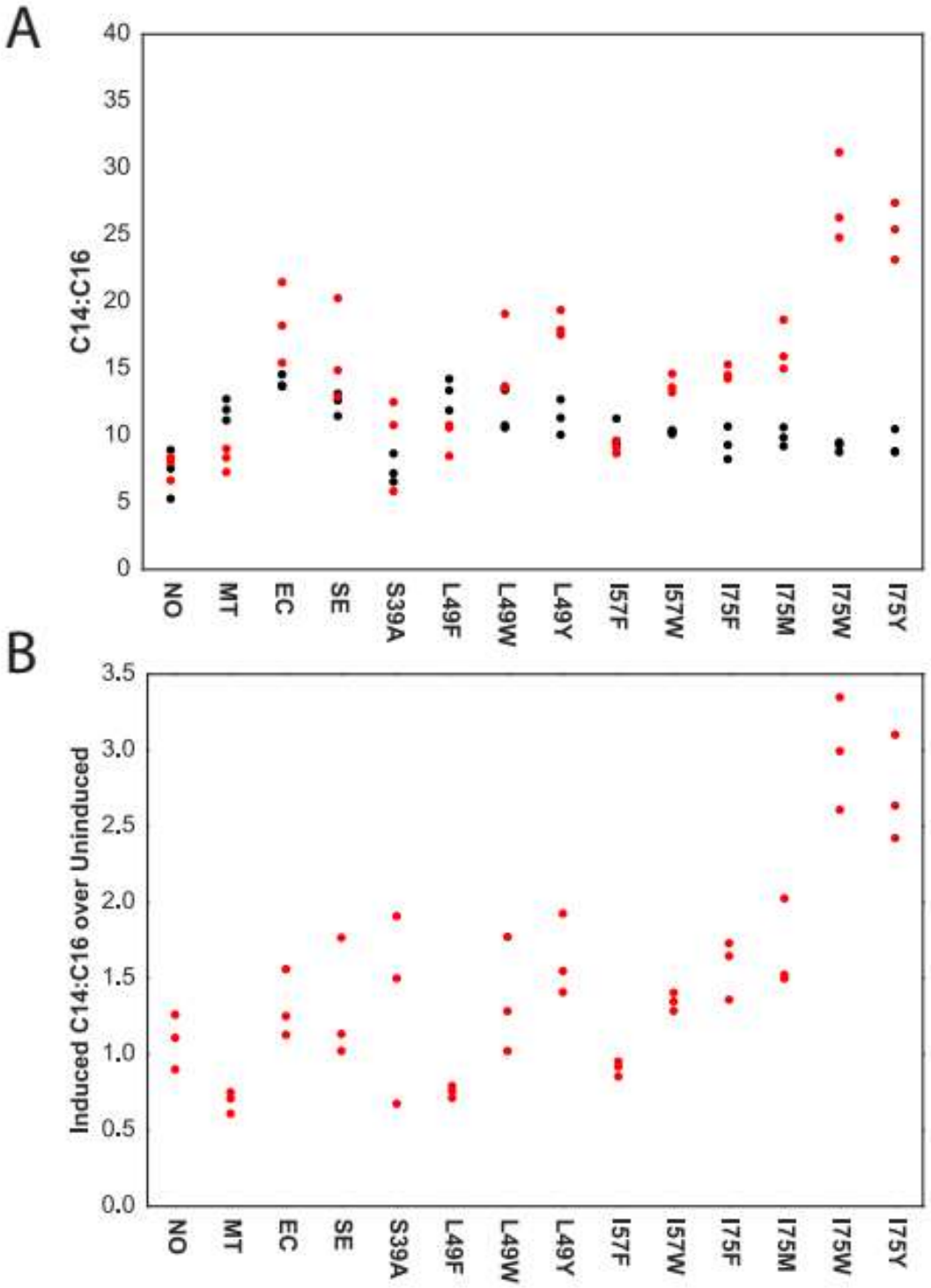



**Figure 3 GC-MS analysis of Cellular Lipids in Single ACP Mutants. A.** Ratios of C14 to C16 molar concentrations for uninduced (black) and induced (red) strains: no vector (NO), empty vector (MT), WT *E. coli* ACP (EC), WT *S. elongatus* ACP (SE). **B.** Fold changes of induced vs. uninduced C14:C16 ratios. The I75W and I75Y mutants have significantly increased C14:C16 ratios as compared to expressing WT Se-ACP ($p<0.05$ for individual two-tailed student-t test on *either* induced "C14:C16" *or* "Induced C14:C16 over Uninduced", not a double/multiple t-test comparison). Data represents triplicate biological measurements.

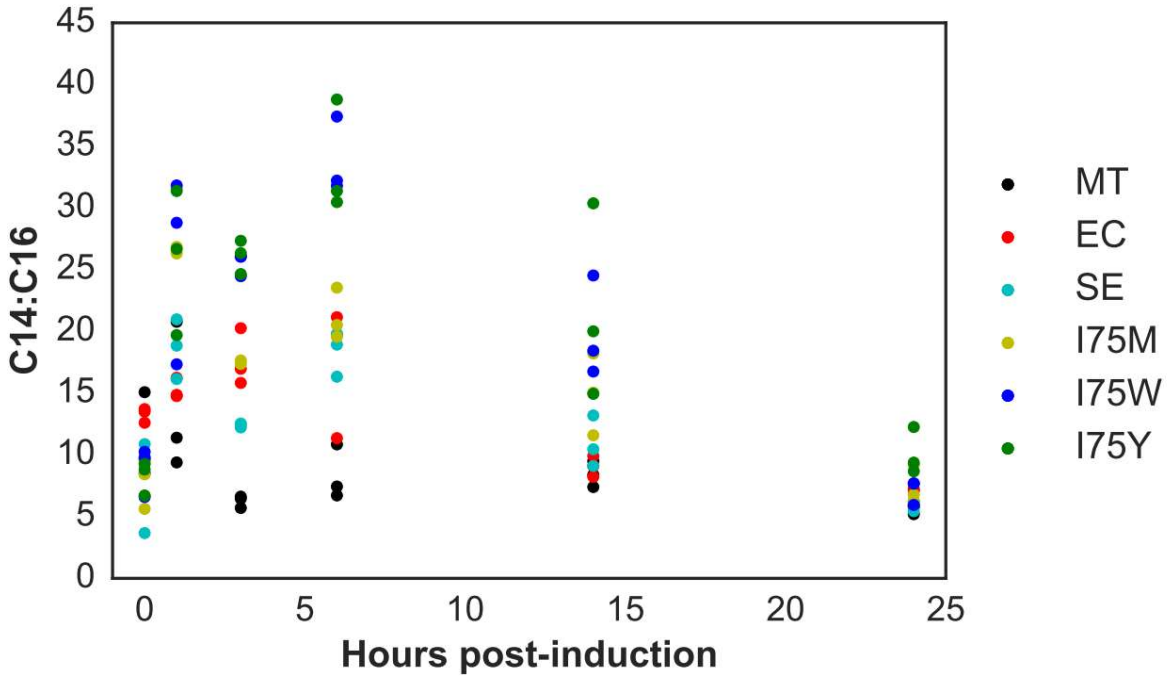



**Figure 4 Time Course of C14:C16 Ratios** Se-ACP I75W and I75Y demonstrate the highest C14:C16 cellular lipid ratio at 5 hours after induction during the growth phase. As the cell cultures saturate past 14 hours, the ratios decrease to the baseline of around 0.05-0.1. Data represent triplicate biological measurements.

production, we characterized C14:C16 ratios over a 24 hour time-course. The lipid pool composition shows that the highest C14:C16 ratio occurs around 5 hours post-induction (Figure 4). Longer induction times resulted in a decreased C14:C16 ratio for all strains, particularly for Se-ACP I75W and I75Y mutants, which fell and became indistinguishable from controls by 24 hours. This highlights the importance of growth phase on lipid composition. During exponential growth, when cells are actively dividing and building new membranes, fatty acid metabolism is highly active, and an abundance of mutated ACPs with reduce pocket sizes likely biases the fatty acid pool toward shorter acyl-chains[18]. It may be that membrane synthesis proceeds with greater fidelity as cell growth slows. Alternatively, short-chain fatty acids may be actively replaced with fatty acids of the correct length, which would be more apparent in stationary phase when new C14 fatty acids are not being added to membrane lipids.

We next tested the effect of mutant ACPs on production of lauric acid (C12). A thioesterase that specifically produces 12-carbon chains (*UcFatB1* from *Cuphea palustris*)[6] was co-expressed with wild-type and mutant *Se*-ACPs and FFA production was measured by GC-MS analysis of fatty acid ethyl esters (FAEE) derived from produced FFAs (Figure 5). We hypothesized that increased levels of shorter chain acyl-ACPs would serve as substrates to the medium chain-specific thioesterase and further increase the yield of medium chain FFAs. In conjunction with expressing the C12 thioesterase, strains over-expressing I75W or I75Y mutant ACPs significantly increased medium chain FFA yields (Figure 5); all controls produced less FFA than the I75W or I75Y mutants. There were significant differences between the various controls, such as the wildtype ACP-expressing strains producing less FFA than the empty vector strain, presumably reflecting the fact that overproducing various forms of ACP can divert carbon and energy flux and may also affect fatty acid metabolism

by, for example, depleting CoA or non-productively interacting with other enzymes.[16,17] FFA yields were uncorrelated to differences in growth rates among all the strains (Figure S4).

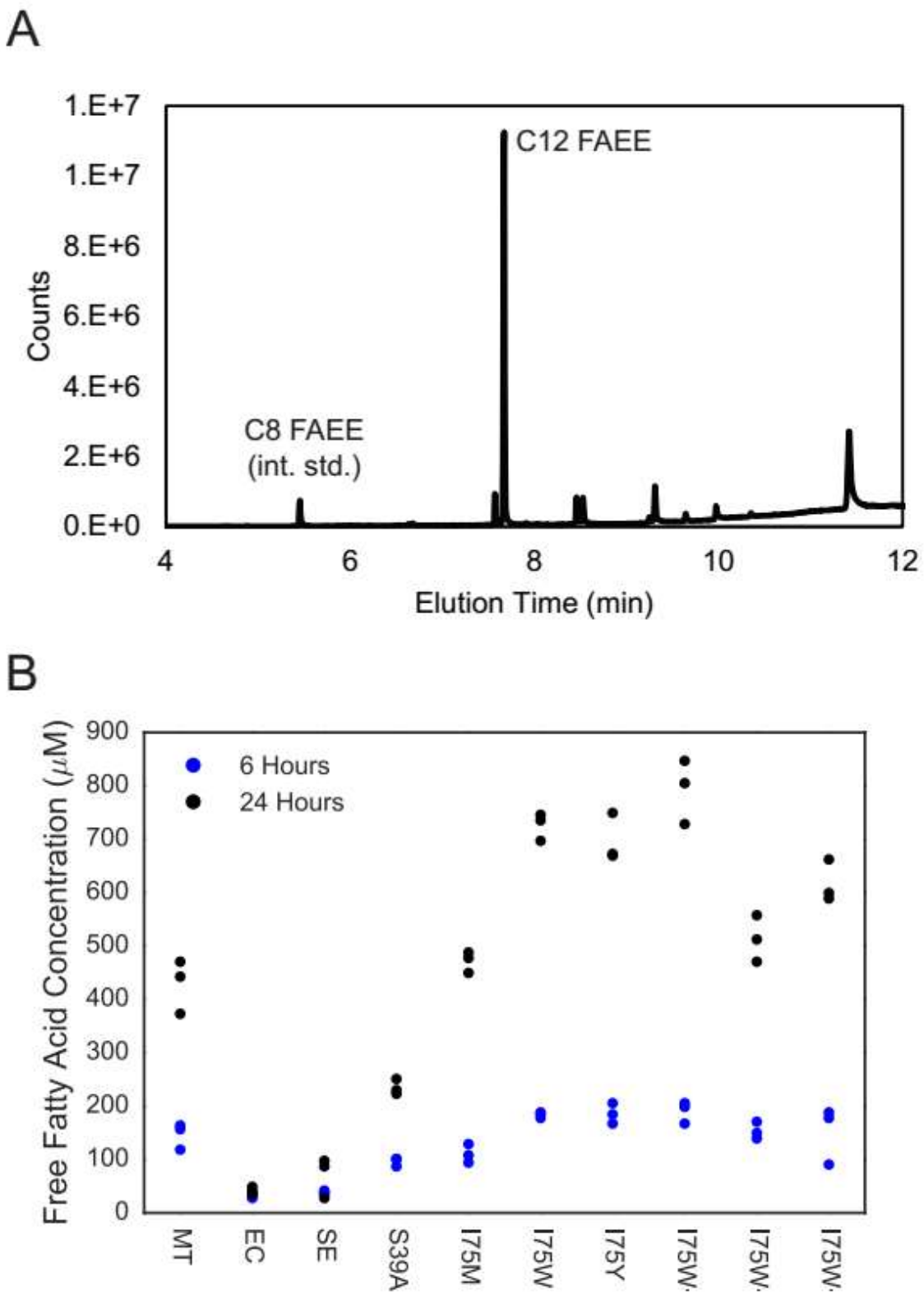



**Figure 5 Free Fatty Acid Production by C12 thioesterase.** **A.** Representative GC-MS trace of FAEEs derived from cell cultures shows thioesterase specificity toward 12-carbon acyl-chains. **B.** FFA concentrations measured from cell cultures at 6 hours (blue) and 24 hours (black) post-induction of both the C12 thioesterase and the indicated ACP. The Se-ACP I75W and I75Y mutants and their derivatives yield more FFA than controls. Data represent triplicate biological measurements.

## Conclusions

In sum, we have shown that ACP, an essential protein in fatty acid metabolism, can be modified by site-directed mutagenesis to skew cellular lipid pools towards smaller acyl-chain lengths. Specifically, expressing certain mutant ACPs enhanced the level of C14 fatty acids in membrane lipids, and by co-expressing mutant ACPs with a chain-length specific thioesterase production of a medium-chain free fatty acid (lauric acid) was enhanced. These results are consistent with a hypothesis that bacterial ACPs influence lipid chain-length during fatty acid synthesis. Other enzymes involved in fatty acid synthesis also likely affect chain-length, and engineering modified acyl-chain specificity has been similarly achieved. For example, FabB and FabF catalyze elongation of fatty acid chains (Figure 1), and have a clearly defined pocket that should accommodate carbon chains up to about 18.[19] Val et al. engineered the FabF pocket to accommodate a maximum of 6 carbons.[20] Similarly, the cyanobacterial aldehyde decarbonylase solved structure[21,22] contains electron density corresponding to a C18 fatty acid or aldehyde; Khara et al. modified this enzyme to have specificity for medium-chain substrates.[22] The C8, C12 and C14-specific plant-derived acyl-ACP thioesterases apparently also control length of fatty acid products, although the underlying structural mechanisms have not been identified. Since FFAs contain the hydrophilic carboxylic acid functional group, they have lower energy density, higher boiling points than hydrophobic molecules like alkanes of same carbon-chain length and are also potentially corrosive to engines, not ideal as fuel molecules. Instead, FFAs can act as precursors to further enzymatic modification for transformation into highly desired fuel molecules such as fatty alcohols and alkanes. Engineering such enzymes (e.g. aldehyde decarbonylases, acyl-ACP reductases, and carboxylic acid reductases) towards shorter carbon chain substrate recognition will

likely be key to tailoring biofuel formulations. To achieve the ultimate goal of efficient biofuel synthesis, it may be necessary to engineer the length specificity of several enzymes – most such enzymes have evolved to handle chains of 16-18 carbons, but shorter chains are desired in fuels. This technology could help to optimize biofuel yield and molecular makeup, which would benefit the goal of developing energy sources alternative to fossil fuels.

## Materials and Methods

### Homology Modelling

The structural model of *Se*-ACP harboring a decanoyl-chain was obtained by homology to the published x-ray crystal structure of the *E. coli* decanoyl-ACP (2FAE) using SWISS-MODEL.[12]

### Strain Construction

Double stranded DNA encoding *E. coli* and *S. elongatus* ACP genes were synthesized as gBlocks (Integrated DNA Technologies) and cloned into the pCDF-Duet vector by Gibson Assembly.[23] Single and double amino acid mutations of the *Se*-ACP gene were incorporated during DNA synthesis. An empty pCDF-Duet-1 vector (Millipore) without the ACP gene was included as control. Plasmids were sequence-verified and transformed into *E. coli* BL21(DE3). For free fatty acid (FFA) production, the C12 thioesterase gene (*UcFatB1* from *Cuphea palustris*) was cloned into pET-Duet-1 vector (Millipore) and transformed into strains harboring the plasmids carrying the ACP variants. DNA sequences of the relevant genes and constructs can be found in Table S1 in Appendix A.

### Growth Kinetics Assay

ACP expressing strains in triplicates were inoculated from single colonies representing independent transformants into LB medium, grown overnight to saturation and back-diluted into M9 minimal media containing 0.4% glucose. The cultures were grown to mid-exponential phase (OD~0.4), dispersed into 96-well plates, induced with various concentrations of IPTG and left to grow shaking at 37˚C in a plate reader (BioTek NEO). The optical densities (OD) of the cultures were recorded every 5 minutes over 15 hours by the plate reader. The growth curves, as well as the final OD after 15 hours

were compared among the strains to quantify growth suppression by ACP over-expression.

**Analysis of Cellular Lipid Composition**

ACP expressing strains in triplicates were inoculated in LB, grown overnight and back-diluted into M9 minimal media containing 3% glucose. The cultures were grown to an optical density of 0.4, induced with 1mM IPTG, and grown for 6 more hours at 37˚C. For the time course experiment (Figure 4), the cultures were left to grow for up to 24 hours. After growth, 10ml of cell culture was used for extraction and analysis, corresponding to wet biomass weights (pellet) of around 5mg (ACP over-expressing, growth defect) to 10mg (not inducing ACP). The cells were pelleted and resuspended in 1:1 methanol:chloroform with 2% glacial acetic acid for lysis, hydrolysis of membrane lipids, and solubilization of fatty acids into the organic phase. Octanoate (C8 fatty acid) was added into the mixture as an internal standard. After vigorous mixing by vortexing, the organic phase was transferred by glass pipettes into glass vials, and the chloroform solvent was evaporated by nitrogen. The vials were then treated with methanol containing 1.25M HCl at 50˚C for 15 hours to catalyze methylation of the fatty acids. The reaction was quenched by adding 5ml of 100mg/ml sodium bicarbonate. 0.5ml hexane was added and the mixture was vortexed vigorously before the hexane phase containing the fatty acid methyl esters (FAME) was extracted and subsequently analyzed on a GC-MS (Agilent 6890/5975).[24] First a standard set of FAMEs with varying chain lengths was run on the GC-MS in scan mode to determine the identity of each fatty acid peak based on the elution time for each fatty acid and comparison of its fragment profile to those in the NIST database (via Agilent ChemStation software). Fatty acid peaks from the extracted cell samples were also identified using scan mode. To quantify peak areas, the background was

minimized by using Selective Ion Mode (SIM) whereby the elution times were used to determine fatty acid identity and only the most dominant mass peaks pertaining to each fatty acid methyl ester were counted. For calibration of concentrations, standard curves for C14 and C16 FAMEs dissolved in hexane were taken in the range of 0.1 to 400mg/L. A linear fit of hexane background-subtracted peak area to known concentration was extracted in the 0.1 to 6.215 mg/L range to cover the range of concentrations seen in the cell samples. Molar concentration was determined by dividing mass concentration (mg/L) by the molecular weight of C14 FAME (242 g/mol) or C16 FAME (270.4 g/mol). To compare the proportions of different chain lengths in each sample, the molar concentration ratio of C14 to C16 FAME was taken.

**Analysis of Free Fatty Acid (FFA)**

ACP and C12 thioesterase expressing strains in triplicates were grown in M9 minimal media containing 3% glucose and induced with IPTG as described above. After 6 or 24 hours of growth, 5 microliters of each culture (cells and media, as medium chain FFA may be secreted) was transferred to wells of a new 96-well plate for high-throughput spectrometric determination of FFA concentration using the Roche Free Fatty Acid Kit (Product Number 11383175001). The free fatty acid is first converted via acyl-CoA synthetase into acyl-CoA, which is then oxidized in the presence of acyl-CoA oxidase to enoyl-CoA, releasing $H_2O_2$ in the process that converts 2,4,6-tribromo-3-hydroxy-benzoic acid (TBHB) and 4-aminoantipyrine (4-AA) to a red dye detectable by spectrometer at 546nm. To specifically detect lauric acid, cultures of ACP plus thioesterase-expressing cells were lysed and extracted with chloroform. The FFA was ethylated and run on the GC-MS to determine the spectrum of chain lengths.

## References


1. Lennen, R. M. & Pfleger, B. F. Microbial production of fatty acid-derived fuels and chemicals. *Curr. Opin. Biotechnol.* **24(6)**, 1044-1053 (2013). doi:10.1016/j.copbio.2013.02.028

2. Choi, Y. J. & Lee, S. Y. Microbial production of short-chain alkanes. *Nature* **502,** 571-574 (2013). doi:10.1038/nature12536

3. Dellomonaco, C., Clomburg, J. M., Miller, E. N. & Gonzalez, R. Engineered reversal of the β-oxidation cycle for the synthesis of fuels and chemicals. *Nature* **476,** 355–9 (2011).

4. Zheng, Y.-N. *et al.* Optimization of fatty alcohol biosynthesis pathway for selectively enhanced production of C12/14 and C16/18 fatty alcohols in engineered Escherichia coli. *Microb. Cell Fact.* **11,** 65 (2012).

5. Lennen, R. M. & Pfleger, B. F. Engineering Escherichia coli to synthesize free fatty acids. *Trends Biotechnol.* **30,** 659–67 (2012).

6. Torella, J. P. *et al.* Tailored fatty acid synthesis via dynamic control of fatty acid elongation. *Proc. Natl. Acad. Sci. U. S. A.* **110,** 11290–5 (2013).

7. Voelker, T. A. & Davies, H. M. Alteration of the specificity and regulation of Alteration of the Specificity and Regulation of Fatty Acid Synthesis of Escherichia coli by Expression of a Plant Medium- Chain Acyl-Acyl Carrier Protein Thioesterase. *Journal of Bacteriology* **176(23),** 7320-7327 (1994).

8. Chan, D. I. & Vogel, H. J. Current understanding of fatty acid biosynthesis and the acyl carrier protein. *Biochem. J.* **430,** 1–19 (2010).

9. Crosby, J. & Crump, M. P. The structural role of the carrier protein–active controller or passive carrier. *Natural product reports* **29(10)**, 1111–1137 (2012).

10. Nguyen, C. *et al.* Trapping the dynamic acyl carrier protein in fatty acid biosynthesis. *Nature* **505,** 427–31 (2014).

11. Masoudi, A., Raetz, C. R. H., Zhou, P. & Pemble IV, C. W. Chasing acyl carrier protein through a catalytic cycle of lipid A production. *Nature* **505,** 422–426 (2013).

12. Roujeinikova, A. *et al.* Structural studies of fatty acyl-(acyl carrier protein) thioesters reveal a hydrophobic binding cavity that can expand to fit longer substrates. *J. Mol. Biol.* **365,** 135–45 (2007).

13. Chan, D. I., Stockner, T., Tieleman, D. P. & Vogel, H. J. Molecular dynamics simulations of the Apo-, Holo-, and acyl-forms of Escherichia coli acyl carrier protein. *J. Biol. Chem.* **283,** 33620–9 (2008).

14. Roujeinikova, A. *et al.* X-Ray Crystallographic Studies on Butyryl-ACP Reveal Flexibility of the Structure around a Putative Acyl Chain Binding Site. *Structure* **10,** 825–835 (2002).

15. Schirmer, A., Rude, M. a, Li, X., Popova, E. & del Cardayre, S. B. Microbial biosynthesis of alkanes. *Science* **329,** 559–62 (2010).

16. Keating, D. H., Carey, M. R. & Cronan Jr., J. E. The Unmodified (Apo) Form of Escherichia coli Acyl Carrier Protein Is a Potent Inhibitor of Cell Growth. *J. Biol. Chem.* **270,** 22229–35 (1995).

17. Magnuson, K., Jackowski, S., Rock, C. O. & Cronan, J. E. Regulation of fatty acid biosynthesis in Escherichia coli. *Microbiol. Rev.* **57,** 522–42 (1993).

18. Knivett, V. a & Cullen, J. Fatty acid synthesis in Escherichia coli. *Biochem. J.* **103,** 299–306 (1967).

19. White, S. W., Zheng, J., Zhang, Y. M. & Rock, C. O. The structural biology of type II fatty acid biosynthesis. *Annu. Rev. Biochem.* **74,** 791–831 (2005).

20. Val, D., Banu, G., Seshadri, K., Lindqvist, Y. & Dehesh, K. Re-engineering ketoacyl synthase specificity. *Structure* **8,** 565–566 (2000).

21. Jia, C. *et al.* Structural insights into the catalytic mechanism of aldehyde-deformylating oxygenases. *Protein Cell* **6,** 55–67 (2015).

22. Khara, B. *et al.* Production of propane and other short-chain alkanes by structure-based engineering of ligand specificity in aldehyde-deformylating oxygenase. *Chembiochem* **14,** 1204–8 (2013).

23. Gibson, D. G. *et al.* Enzymatic assembly of DNA molecules up to several hundred kilobases. *Nat. Methods* **6,** 343–5 (2009).

24. Lennen, R. M., Braden, D. J., West, R. a, Dumesic, J. a & Pfleger, B. F. A process for microbial hydrocarbon synthesis: Overproduction of fatty acids in Escherichia coli and catalytic conversion to alkanes. *Biotechnol. Bioeng.* **106,** 193–202 (2010).

CHAPTER 3

# ARTIFICIAL MINERALIZATION, MAGNETISM AND METAL SENSING IN CELLS

## Preface

In this Chapter I introduce the application of synthetic biology toward inorganic nanomaterial synthesis inside cells and some their functional applications. Supporting materials for this chapter are found in Appendix B.

## Abstract

Genetically encoding the synthesis of functional nanomaterials such as magnetic nanoparticles enables sensitive and non-invasive biological sensing and control. Via directed evolution of the natural iron-sequestering ferritin protein, we discovered key mutations that lead to significantly enhanced cellular magnetism, resulting in increased physical attraction of ferritin-expressing cells to magnets and increased contrast for cellular magnetic resonance imaging (MRI). The magnetic mutants further demonstrate increased iron biomineralization measured by a novel fluorescent genetic sensor for intracellular free iron. In addition, we engineered *Escherichia coli* cells with multiple genomic knockouts to increase cellular accumulation of various metals towards potential applications in bioremediation and mining. Lastly to explore further protein candidates for biomagnetism, we characterized members of the DUF892 family using the iron sensor and magnetic columns, confirming their intracellular iron sequestration that results in increased cellular magnetization.

## Introduction

Inorganic nanomaterials have been used in a wide range of biological applications including fluorescent or plasmonic labelling for imaging, magnetic labelling for extraction and high throughput sequencing and drug-delivery[1–3]. However, unlike genetically-encoded labels such as green fluorescent protein (GFP), chemically synthesized inorganic nanomaterials, despite their versatile physical and chemical properties, are ultimately limited in their biological application by their lack of integration with the genetic circuitry of the cell. Synthetic biology can bridge this gap by programming cells to controllably synthesize their own nanomaterials in response to biological signals. Those nanomaterials can be further tailored within cells to interact with other components and transduce biological signals downstream.

There are few examples of bio-synthesized inorganic nanomaterials in Nature. Certain species of bacteria and archaea can mineralize nanoparticles via proteins or metabolites that reduce toxic metal cations[1,4]. Notably, magnetotactic bacteria of the genus *Magnetospirillum* naturally synthesize crystalline magnetite nanoparticles and align them as a passive navigation compass for the cell in its natural environment[5–7]. Despite speculation on the presence of similar inorganic magnetic nanoparticles in animals such as fish and humans, no such biomineralization pathways have been confirmed so far[8–11]. However, all cells do use inorganic bio-mineralization to maintain near constant concentrations of essential trace metals via high affinity chelators and storage proteins for times of excess. One prominent example is the ferritin, a ubiquitous class of proteins found in all domains of life that play a crucial role in iron homeostasis[2,12–18]. Ferritins form shells composed of 24 monomers each, creating an inner cavity in order to store iron in a hydrated, amorphous form of iron

oxide similar to the mineral ferrihydrite. (Figure 1a, b) Iron oxide is biocompatible and magnetic

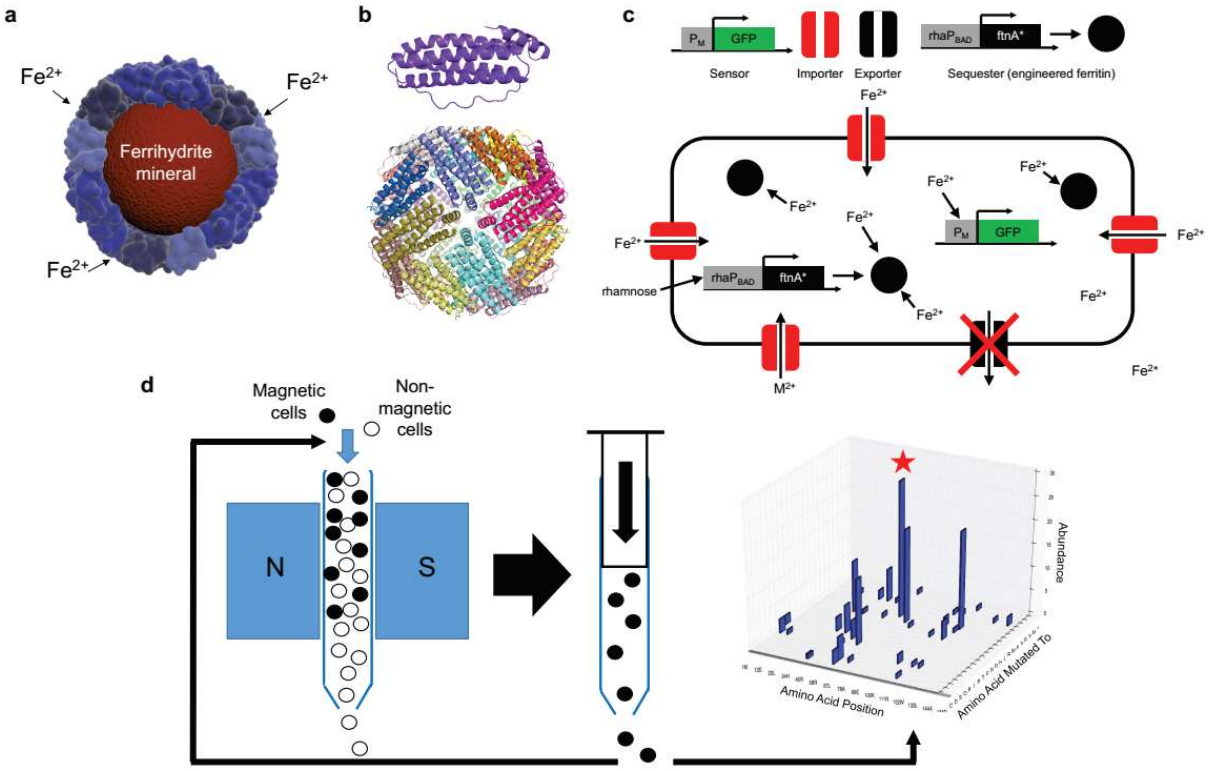



**Figure 1 Engineering cellular magnetism (a)** Schematic of the ferritin protein shell encaging a ferrihydrite nanoparticle **(b)** crystal structure of the ferritin monomer (top) and the self-assembled, 24-homomers cage (bottom). **(c)** The cell engineered to accumulate metals by knockout of genomic exporters (black) and expression of importers (red). Mutant ferritin particles (black spheres) induced by a rhamnose promoter biomineralize iron into intracellular magnetic particles. A genetic fluorescence sensor monitors intracellular free $Fe^{2+}$ level. **(d)** directed evolution for increased magnetism: iterative selection by high gradient magnetic column of a library of cells expressing randomly mutated ferritins was carried out over 10 days (1 cycle/day). Subsequent sequencing analysis of the magnetically retained mutants enabled discovery of mutations in key residues (e.g. red star: T64) that enhanced cellular magnetism.

depending on its crystal structure. However, the mineralized iron stored inside natural ferritins exhibits poor crystallinity which facilitates iron release in times of need but also results in a very modest inherent magnetic moment[19–21]. Even though the factors that control crystallization and hence the properties of the magnetic nanoparticles inside ferritin cages are not completely clear, natural ferritins still represent an excellent starting point for protein engineering aimed at increasing the inherent magnetism of ferritin particles. Engineering increased bio-magnetism would open up the way for advanced applications in non-invasive biological sensing, imaging and actuation[22–34]. Here, we employ directed evolution of ferritin to enhance the magnetism and biomineralization capability of engineered *E. coli*. In addition, a novel genetic biosensor for cytoplasmic free iron was developed to easily measure biomineralization efficiency *in vivo*, combined with genetic manipulations of metal transporters to optimize intracellular metal levels[35–39]. (Figure 1c)

## Results

We discovered novel mutations in ferritin that increase cellular magnetization via “magnetic evolution”, combining random mutagenesis with iterative magnetic column selection (Figure 1d). Specifically, a randomized library of *E. coli* ferritin (*ftnA*) mutants was cloned and transformed into *E. coli*. Magnetic cells from this population were enriched through ten iterative rounds of growth followed by passage through high-gradient magnetic columns. Eventually, the selected population of cells was plated and sequenced, and the most frequent mutations were identified and validated in freshly cloned and transformed cells (Figure 2a). We further combined and tested additional mutants informed by results of the initial random mutagenesis experiment. In total, we tested 38 mutants of *ftnA* via overexpression in *E. coli*. (Table S1) The native ferritin-clan genes (*ftnA, bfr, dps*) were knocked-out to prevent interference with the magnetic evolution approach. Changes in cellular magnetism were compared via measurement of the retention fraction of cells in high-gradient magnetic columns. The results show that the majority of mutants, compared to wild-type ferritin, significantly increase levels of magnetism of cells (Figure 2b). The top mutant (m1) contains the double point mutations H34L and T64I at the 2-fold symmetrical “B-type channel” (Figure 2c). These point mutations result in smaller or less polar residues occupying the entrance to the “B-type channel”. This would likely increase the size of the channel encouraging ion diffusion into the core[40]. The two sites were individually mutated to alanine and demonstrated increased magnetism compared to wild-type while still being less magnetic than the best double mutant. Furthermore, a BSA-calibrated SDS-PAGE gel of the cells shows that the mutants, in particular the most magnetic ones, exhibited reduced protein expression compared to wild-type, confirming that the increased cellular magnetism is not caused by

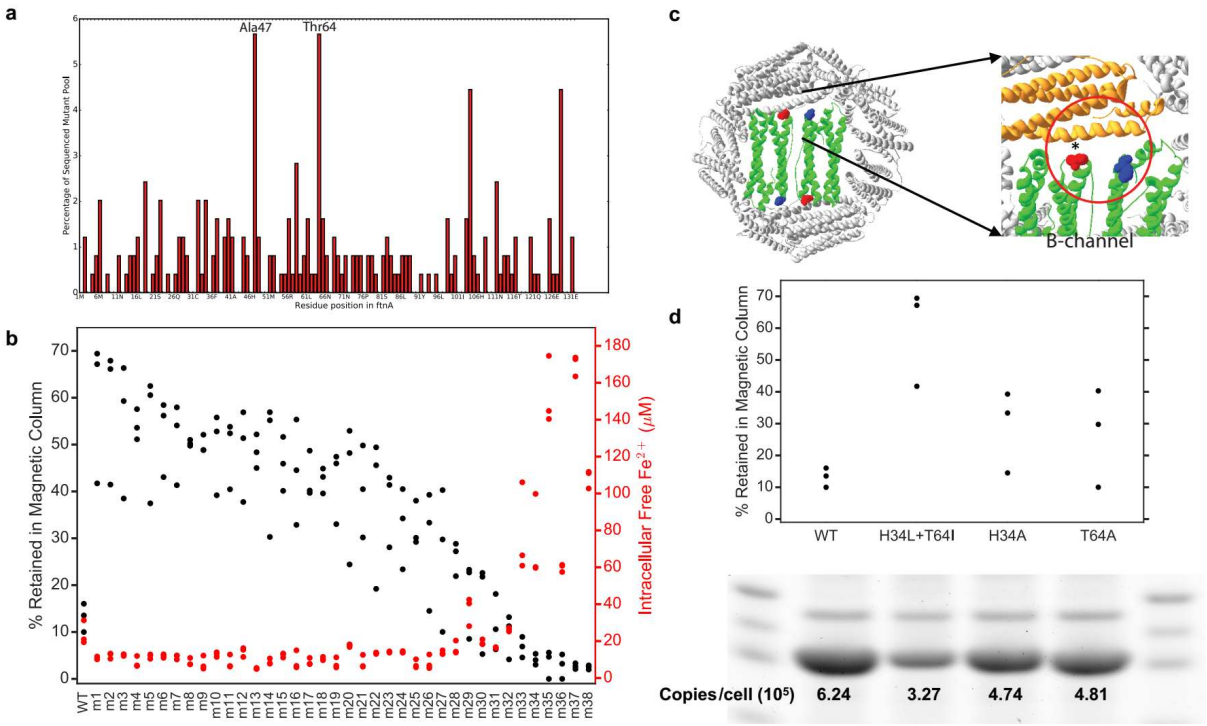



**Figure 2 Mutations in ferritin proteins enhance magnetism. (a)** Histogram of mutations from sequencing samples of the magnetically-evolved culture reveals mutations of key residues (e.g. Thr64, Ala47) that enhance biomagnetism. **(b)** Magnetic column retention for all engineered ferritins (Table S1) including H34L+T64I (m1), sorted in descending order, most showing significant increase over wildtype (m1-m21, $p<0.05$ by two-tailed t-test, N=3). The intracellular free iron concentrations inferred from genetic iron sensor measurements are anti-correlated to magnetic column retention levels. **(c)** The most magnetic ferritin mutant (H34L+T64I) is doubly mutated at the “B-channel” (red circle) that transports iron. **(d)** The importance of the His34 (blue) and Thr64 (red) sites are individually validated by single mutants, demonstrating increased magnetic retention levels over the wildtype (WT). Furthermore, quantitative SDS-PAGE gel analysis (cropped to show ftnA band) demonstrates weaker expression for the mutants compared to the wildtype, confirming that increased magnetism is not caused by mere increases in protein and nanoparticle quantities.

increased levels of ferritin expression. (Figure 2d, S1)

We further developed a genetic fluorescent iron sensor that demonstrates an increase in cytosolic iron sequestration in strains expressing mutant ferritins with increased magnetism. The iron sensor employs the *E. coli* promoter *fiu*, which contains four overlapping binding sites for the transcription factor *fur* (ferric uptake regulator). *Fur* bound to $Fe^{2+}$ represses downstream expression of GFP on a low-copy plasmid (p15A origin) (Figure 3a). Hence the *fiu*-based sensor reports depletion of free iron in the cytoplasm. Further calibration using a range of concentrations of the cell-permeable ferrous iron chelator bipyridine (bpd) allows conversion of culture density-normalized fluorescence to free iron concentrations per cell (Figure S2). This converted concentration was monitored over time for *E. coli* growing from exponential to stationary phase in LB medium supplemented with different Fe(II) sulfate concentrations (0 to 5 mM) over a range of ferritin induction levels (0% to 0.2% rhamnose). The time courses show rapid sequestration of intracellular iron under high ferritin inductions and decreased time to reach peak iron concentrations (Figure 3, b-g). Without induction, significantly higher intracellular Fe levels were observed when iron was supplemented into the medium at 5 mM, close to the observed viability limit for *E. coli* (~10mM). Comparing the final readout at 15 h between cells expressing wild-type and the most magnetic mutant (m1) shows that the mutant sequesters more iron, especially at low induction levels below 0.01% rhamnose (Figure 3, h-i, replicated in Figure S3, S4). We also found that for the ferritin mutants, the intracellular iron concentration in stationary phase are generally anti-correlated with magnetic column retention (Figure 2c). Considering the lower protein expression levels of the mutant as estimated quantitatively from SDS-PAGE gel analysis of whole-cell lysates (Figure 2d), iron sequestration (i.e. decrease of

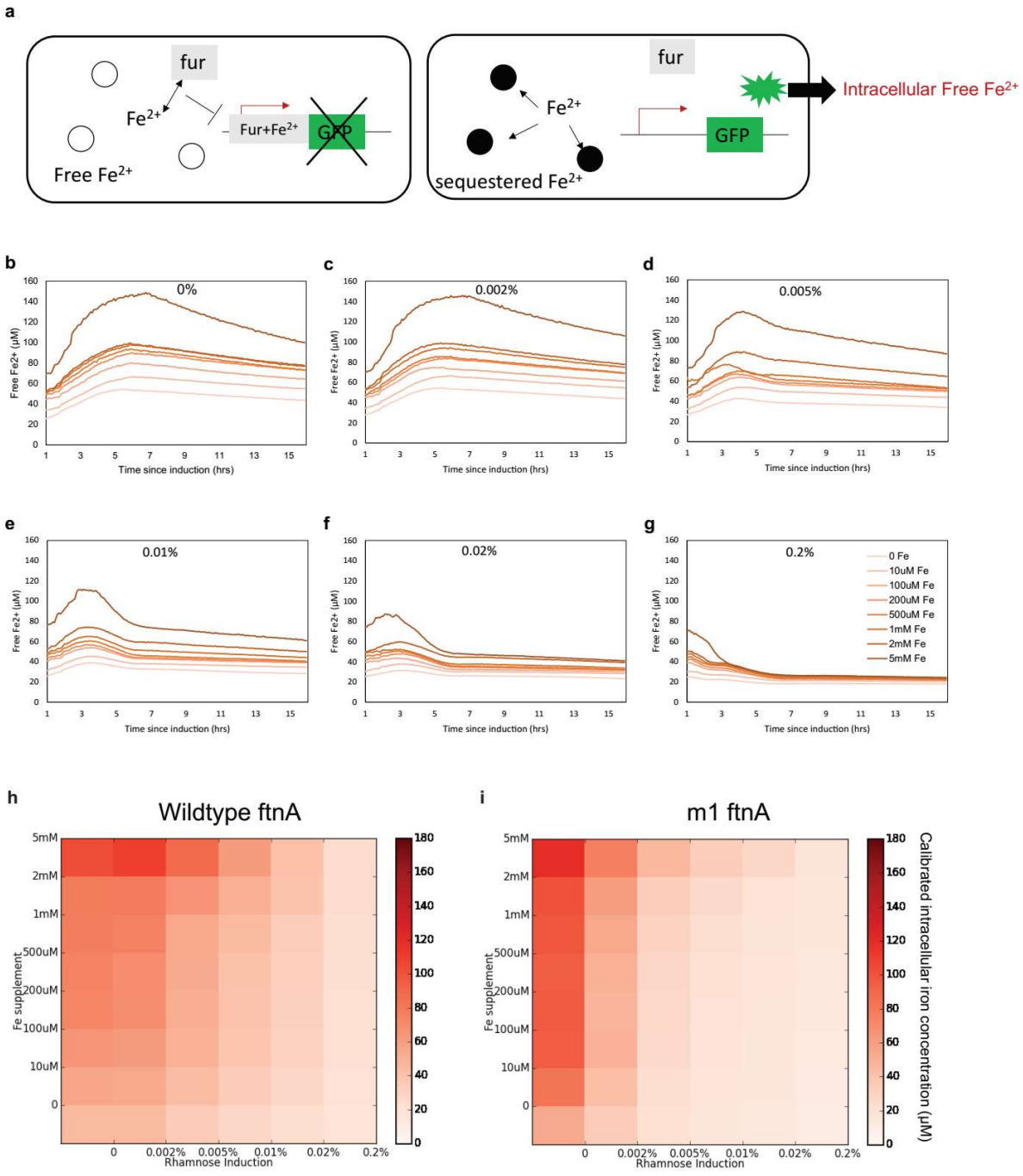

a
fur
Fe2+
Fur+Fe2+
GFP
Free Fe2+
fur
Fe2+
GFP
sequestered Fe2+
Intracellular Free Fe2+
b
0%
c
0.002%
d
0.005%
e
0.01%
f
0.02%
g
0.2%
0 Fe
10uM Fe
100uM Fe
200uM Fe
500uM Fe
1mM Fe
2mM Fe
5mM Fe
Free Fe2+ (μM)
Time since induction (hrs)
h
Wildtype ftnA
i
m1 ftnA
Fe supplement
5mM
2mM
1mM
500uM
200uM
100uM
10uM
0
Rhamnose Induction
0
0.002%
0.005%
0.01%
0.02%
0.2%
Calibrated intracellular iron concentration (μM)

**Figure 3 Genetic fluorescent sensor monitors cellular $Fe^{2+}$ sequestration in WT and mutant ferritin expressing cells (a)** Free $Fe^{2+}$ binds to the ferric uptake regulator (apo-fur). The Fe-bound fur binds the fiu promoter sequence to repress transcription of GFP (right). Sequestration of free $Fe^{2+}$ by ferritins increases sensor fluorescence which with calibration is converted to the intracellular free $Fe^{2+}$ concentration. **(b-g)** Calibrated free $Fe^{2+}$ concentrations in *E. coli* expressing wildtype ferritin from up to 15 hours after induction by 0% (b), 0.002% (c), 0.005% (g), 0.01% (e), 0.02% (f), and 0.2% (g) rhamnose with media $Fe^{2+}$ supplement of 0 μM, 10 μM, 100 μM, 200 μM, 500 μM, 1 mM, 2 mM, 5 mM. Without ferritins high media supplement at 5mM can dramatically alter the intracellular iron homeostatic setpoint. At high induction levels free $Fe^{2+}$ is efficiently sequestered up to the highest media supplement concentration. The heatmaps with color saturation proportional to calibrated intracellular free iron levels show that compared to the wildtype **(h)**, the best ferritin mutant **(i)** is much more effective at sequestering iron at lower protein levels (0.01% rhamnose induction) and at high environmental iron concentration (up to 2 mM). This is consistent with their greater magnetism despite lower protein expression.

iron concentration) generally correlates with greater magnetic retention.

Cellular magnetism and mineralization were further confirmed by a range of characterization methods. The increased magnetic moment of the mutant cells was visualized directly by growing cells in LB rich medium in titer plate over ring-shaped permanent magnets[21]. Over-expression of mutant ferritins produced a sharper "image" of the magnets compared to wild-type expression, due to greater cellular magnetic moment and hence force on the cells from the surface field gradient of the permanent magnets (Figure 4a). Iron-loaded ferritins could be visualized in some cell cross-sections as aggregates of electron-dense puncta using transmission electron microscopy (TEM) (Figure 4b). Due to their paramagnetism, the local field inhomogeneity around the particles in the presence of an applied magnetic field increases spin-spin relaxation of the spins of protons in surrounding water molecules, producing contrast for T2* based MRI using a gradient-echo pulse sequence. The significant reduction in T2* was only observed for over-expression of the mutant m1, attributable to the much greater moment of the individual particles (Figure 4c). Lastly, Superconducting Quantum Interference Device (SQUID) measurement of cellular magnetic moment shows that only overexpression of the mutant ferritins leads to a positive paramagnetic contribution over the negative diamagnetic contribution from cells (Figure 4d).

In addition to iron-sequestration, the ferritin mutants have potential to co-sequester other elements. We first engineered *E. coli* with knockouts in three divalent cation exporters (*rcnA, fieF, zntA*) and also the iron master regulator *fur*, causing increased metal accumulation in the cells, particularly for iron, zinc and cobalt (Figure 5). Knockout of *zntA* alone seemed sufficient for increasing zinc concentrations, particularly when ferritins were over-expressed. However, due

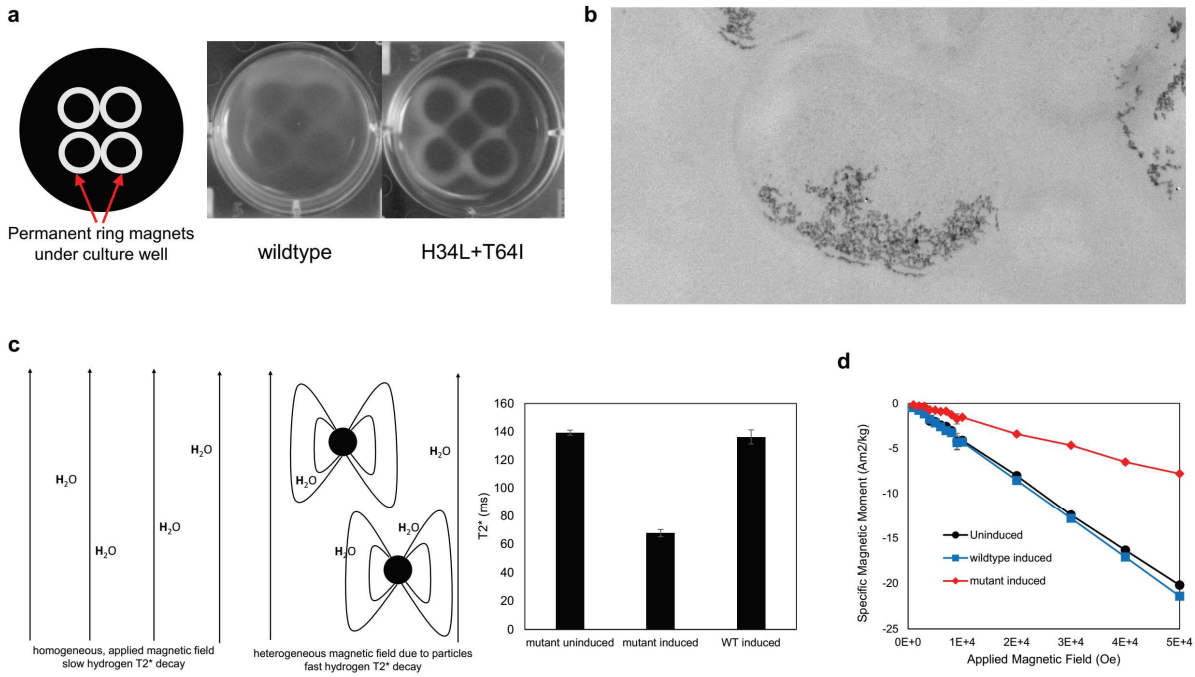



**Figure 4 Characterizing cellular magnetism and biomineralization. (a)** In liquid culture, *E. coli* expressing the mutant are attracted to ring-shaped magnets much more than the wildtype, due to stronger magnetic moments. **(b)** TEM images show electron-dense magnetic particles in a cross section of fixed and embedded cells that expressed engineered ferritins. **(c)** In MRI imaging, the presence of magnetic ferritin particles creates local magnetic field inhomogeneities that variably alter nearby proton relaxation rates and accelerate T2* decay. Triplet cultures of *E. coli* expressing ferritin mutants more significantly decreased T2* relaxation time compared to both no expression or overexpressing the wildtype ($p<0.05$ by two-tailed t-test, N=3), which creates imaging contrast. **(d)** SQUID magnetometry shows that only expressing the mutant ferritin led to positive paramagnetic increases in the cellular magnetic moment relative to the uninduced control. (Error bars are standard deviations from sequential measurements on the same sample)

to compensatory effects among the exporters and also *fur* regulation, only knocking out *fur* along with the three exporters (*rcnA, fieF, zntA*) significantly increased cellular iron levels as measured by inductively coupled plasma mass spectrometry (ICP-MS). We also found co-expressing importers could increase iron levels and cellular magnetism (data not shown). We then tested several elements known to absorb to or alloy with iron oxide: cobalt, nickel, cadmium and arsenic. Cells were exposed to varying concentrations of these elements. Only strains expressing magnetic mutant ferritins exhibited a decreased growth defect, suggestive of toxic metal sequestration by over-expressed ferritin mutants (Figure S5). The benefit of expressing the mutants over the wild-type ferritins can be explained by the increased biomineralization rate of the mutant ferritins which aids in co-sequestration of other elements into the growing iron oxide core. Hence cells over-expressing the metal sequestering ferritins could be applied toward biological mining or remediation of precious or toxic metals. And highly sensitive and selective intracellular sensors for these metals similar to the iron sensor demonstrated here would be desirable to directly monitor the effect of ferritin sequestration.

Lastly to explore constructs other than the well-known ferritins as candidate starting points to modulate biomagnetism, we studied the iron sequestration and magnetic properties of Domain of Unknown Function 892 (DUF892) proteins, which fold into a similar 4-helix bundle as ferritins and are thus distantly related in structure-based phylogeny[41] (Figure 6a). Three predicted, uncharacterized proteins in the DUF892 family (A0A072C8A3 from *Sinorhizobium americanum*, F7XJ33 from *Sinorhizobium meliloti* and Q1QGY5 from *Nitrobacter hamburgensis*) were cloned and over-expressed in wildtype *E. coli* containing the fluorescent iron sensor. The converted intracellular iron levels based on fluorescence measurements were

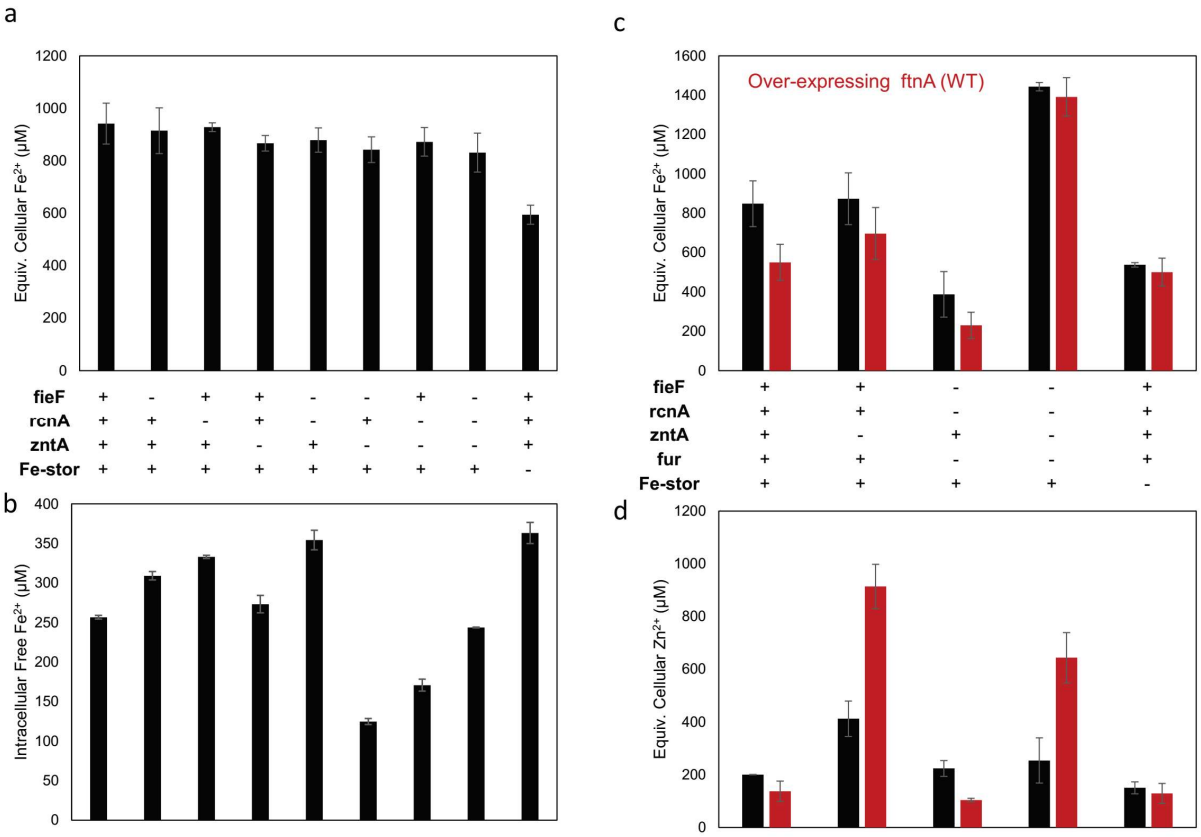



**Figure 5 Effect of genomic knockouts on cellular metal concentrations (a)** ICP-MS measurement of Fe concentration for combinations of ferritin or metal exporter knockouts (-). With the master regulator fur protein present, iron homeostasis is maintained unless multiple ferritin homologs are knocked out (Fe-stor = *ftnA*, *bfr* and *dps*). **(b)** Genetic Fe-sensor measurement of intracellular free $Fe^{2+}$ concentration shows effect of knockouts in the same strains despite homeostasis of total Fe in A. **(c)** ICP-MS of total Fe concentration for additional mutants with *fur* KO, which increases total Fe level. Over-expressing ftnA (red bars), however, did not increase iron level. **(d)** ICP-MS of zinc concentration for additional mutants with *fur* KO. Knockout of *zntA* alone dramatically increases cellular Zn levels, especially when expressing ferritins capable of binding zinc. *zntA* plays a crucial role in Fe export in the absence of the iron exporters and regulator *fur*, resulting in dramatically decreased Fe levels in contrast to its knockout. *zntA* knockout in combination with either *fieF* or *rcnA* KO further increases intracellular Fe sequestration. Comparison of total Fe content (a) vs. free $Fe^{2+}$ via the genetic sensor (b) illustrates the interplay between metal transport and storage in cells. (Error bars are standard deviations of repeated ICP-MS measurements on the same sample)

markedly decreased compared to the no-expression control, with a concomitant increase in the percentage of cells retained in magnetic columns suggesting biomineralization (Figure 6b). These results confirm that the function of DUF892 proteins is likely similar to that of ferritins, namely to sequester free iron in times of excess or stress.

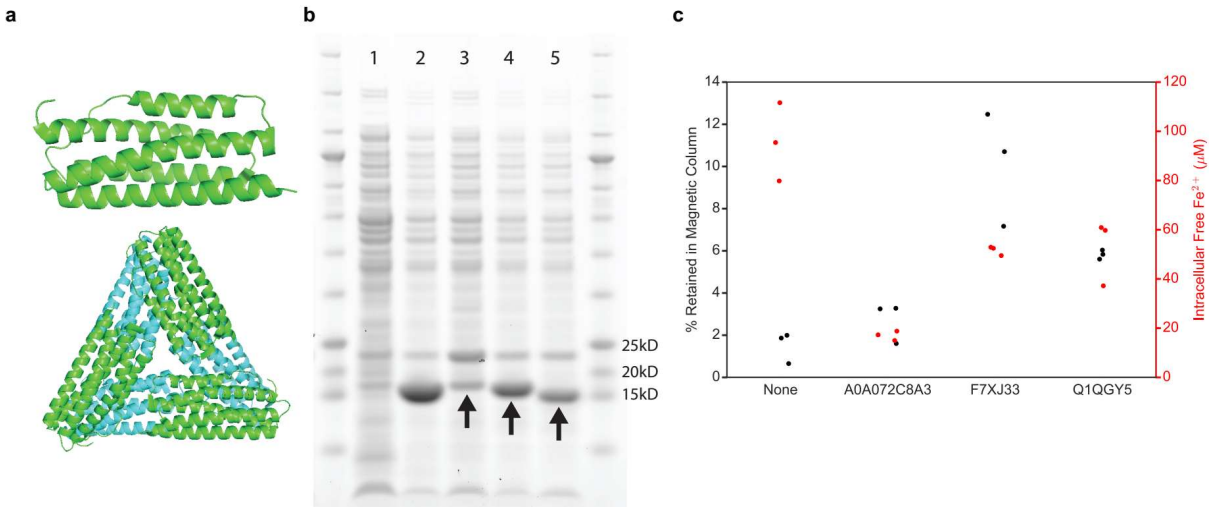



**Figure 6 DUF892 proteins sequester iron and increase cellular biomagnetism.** **(a)** The crystal structure of a DUF892 family protein monomer (*E. coli* YciF; PDB ID: 2GS4) shows a similar 4-helix bundle as ferritin (top), but with different predicted quaternary structure (bottom). **(b)** Verification by SDS-PAGE of cell lysates for the over-expression in *E. coli* of three predicted proteins in the DUF892 family: A0A072C8A3 (lane 3), F7XJ33 (lane 4), Q1QGY5 (lane 5). Lane 1 is no-expression control. Lane 2 is over-expression of ftnA. The over-expressed proteins have similar theoretical molecular weights with bands running close to 15kD mark (arrows). The common band close to 25kD mark represents GFP output of the iron sensor integrated into the *E. coli* **(c)** The new DUF892 proteins demonstrate significantly decreased intracellular free iron levels ($p<0.05$ by two-tailed t-test, N=3) and increased level of cellular magnetism by magnetic column retention measurement compared to the no-expression control.

**Discussions**

In this study, we used directed evolution by magnetic selection to discover point mutations in ferritin that enhance iron sequestration and the magnetic phenotype of cells expressing ferritin mutants. Increased cellular magnetism was shown to lead to physical attraction to magnetic field gradients and improved MRI contrast. We further developed a new genetic iron sensor in *E. coli* that fluorescently detects variations in intracellular free iron levels to validate iron sequestration by ferritin mutants and other previously uncharacterized iron-sequestering proteins.

The genetic iron sensor was used to demonstrate the ability of magnetic ferritin mutants to more efficiently sequester iron. In addition, it revealed several features of cellular iron homeostasis. First, intracellular free iron concentration of *E. coli* increases sub-linearly with increasing iron supplement in culture media until around 2 mM, above which the sensor displays a sudden increase in intracellular free iron levels, coincident with significant growth defect and death (Figure 3b). However, over-expression of ferritins rapidly sequesters extra iron, independent of the range of supplement concentrations tested, to yield a low, steady iron level in the cells. The lower free iron level under ferritin expression despite high supplement levels for iron indicates that the cells are not able to efficiently compensate for lost iron by increasing import. This is due to the active nature of iron import relying largely on the synthesis of iron siderophores which may be the bottleneck in maintaining iron homeostasis during rapid iron starvation.

The directed evolution and magnetic selection of ferritins is an efficient strategy to obtain magnetic protein constructs. The mutations found in the most magnetic ferritin at the ferritin B-type channel directly affect iron transit into the protein shell. It is commonly recognized that the 3-fold symmetrical channels of the

ferritin particle are most important for iron transit. It has also been assumed that iron entry requires Fe(II) to first localize to and oxidize at the ferroxidase active site close to the center of the 4-helix bundle. The role of the B-type channel has not been highlighted until recently[40]. Here, we show mutations around the B-type channel directly affect iron sequestration rates *in vivo*. In particular, the H34L and T64I mutations likely enlarged the B-type channel to increase the influx of un-oxidized Fe(II) ions. Histidine has larger side chain compared to leucine. On the other hand, threonine at position 64 could serve as a C-terminal cap of the second major helix via hydrogen bonding, hence its replacement could change pore structure to favor ion entry. This in turn may affect not only the concentration but also the balance of Fe(II) and oxidized Fe(III) species in the core. Natural ferritin mineralizes mostly Fe(III) ions into $Fe_2O_3$, which may exhibit antiferromagnetic coupling of iron atom spins resulting in a very low magnetic moment. However, magnetite ($Fe_3O_4$), the most magnetic form of iron oxide, contains mixed valence atoms of Fe(II) and Fe(III). Hence the influx of Fe(II) into the ferritin core without oxidation could change the stoichiometry of iron valencies inside the nano-reactor to favor mineralization of the more magnetic magnetite. The increased iron-sequestration and cellular magnetism of mutants resulting from the directed mutation of residues His34 and Thr64 at the B-channel to alanine supports this hypothesis. Interestingly, we found that certain mutants with N-terminal “SpyTag” also increased magnetization (Fig. 2, Table S1). The SpyTag were introduced to allow potential covalent attachment of ferritin particles onto other targets possessing the corresponding “SpyCatcher” tag. Unlike the H34L and T64I double mutant, however, the SpyTagged ferritin displayed comparable or stronger protein band intensity on SDS-PAGE gel relative to wildtype ferritin in similarly grown and induced cells (data not shown), suggesting high protein

expression levels that can increase the number of ferritin particles per cell contributing to higher total iron sequestration and magnetic response. Furthermore, the extra peptide at the ferritin N-terminus, close to the 2-fold symmetric B-type channel in the self-assembled structure, could alter local structures around the pore to similarly affect its properties and potentially increase iron flux as some of the other beneficial mutations.

The magnetic particles from ferritin mutants could serve as noninvasive, genetically-encoded reporters of biological activity in deep tissues transparent to magnetic fields, as demonstrated by their enhanced MRI relaxation rate in T2* MRI. T2 imaging using spin-echo sequence is more commonly reported for magnetic particles. However, spin-echo sequences are designed to cancel the dispersion in transverse relaxation due to magnetic field inhomogeneity produced by the ferritin iron-oxide particles in the static limit. Hence the T2 signal change is dependent not only on the magnetic moment but also on change in the particles' local magnetic environment due to diffusion during measurement. Hence T2 measurement is likely dependent on measurement parameters like sequence duration, whereas T2* gradient echo is not. For *in vivo* imaging, both T2 and T2* imaging modes can be used to detect contrast generated from endogenous magnetic particles. The ferritin mutants could allow noninvasive reporting of *in vivo* biological signals from engineered cells such as bacteria in the gut, or immune cells targeted to cancer tumors. There are further potential application of engineered intracellular magnetic particles as magnetic force or magneto-thermal actuators to manipulate proteins such as ion channels or subcellular compartments to control cellular behavior. Such applications include control of calcium-based signaling in mammalian cells, particularly neurons to induce or inhibit action potentials (e.g. "magneto-

genetics")[33,34], and manipulations of subcellular organelles and structures (e.g. magnetic tweezer). However, the natural size of the ferritin cage, at 8 nm in core diameter, is too small to create sufficient magnetic moment from the encaged nanoparticle to achieve desired levels of sensitivity and reliability in most actuation applications[19]. Larger protein shell templates are desired, and Nature may contain several such templates such as encapsulins[42]. The approach developed here of using a fluorescent metal sensor and magnetic directed evolution could be similarly used, such as done to characterize the function of DUF892 proteins here, to discover additional novel candidates that would enable enhanced biomagnetism applications in mammalian biology.

In another application realm, ferritins have been mutated or modified with a variety of biological and chemical tags to confer biological properties such as targeting to specific cells[43] to be used as potential drug-delivery vehicles and imaging agents *in vivo*. Douglas *et al.* first modified the exterior N-terminus of human heavy chain ferritin (HFn) with the RGD-4C peptide (CDCRGDCFC) that selectively binds integrins $\alpha_v\beta_3$ and $\alpha_v\beta_5$ for cancer cell targeting[44]. They showed that the peptide-modified ferritins can similarly mineralize iron oxide *in vitro* under non-physiological conditions (65°C, pH 8.5 with $H_2O_2$ addition) as wildtype ferritins and demonstrate expected size-dependent magnetic properties. Furthermore, the peptide-modified ferritins exhibited enhanced targeting to cancer cells (C32 melanoma) and macrophages (enriched in diseased tissue such as atherosclerotic plaques) as well as better cellular internalization[45], inspiring later works that demonstrated favorable anti-tumor activities from cell-targeted ferritins loaded with the drug doxorubicin (Dox). In addition to RGD-4C, other functionalities have been attached to the *in vitro* mineralized ferritin particles including α-melanocyte-

stimulating hormone to improve cancer targeting specificity[46], polyethylene glycol (PEG, 4kDa) for immune masking[46], epidermal growth factor (N-terminal fusion) for breast cancer cell targeting[47], and β-cyclodextrins attached onto a cysteine substituting a native serine residue for complexion with and slow release of hydrophobic molecules[48]. Moreover, Kang et al. have inserted either the Fc-binding peptide (GGGGGGDCAWHLGELVWCTGGG-GGAS)[49] or the thrombin cleavage peptide (GGLVPRGSGAS)[50] into the flexible loop between helices D and E on the *Pyrococcus furiosus* ferritin for binding of cell-targeting antibodies or thrombin-mediated release of cargo, respectively. Most of the work modifying ferritins have focused on characterization of the additional functionality (e.g. cell targeting) while demonstrating normal or similar morphology and characteristics of the modified particles compared to unmodified wildtype *in vitro*. These works have not directly compared the *in vivo* iron sequestration or magnetic properties of these modified protein cages to those of the wildtype, which have been the focus in my work toward the goal of maximizing *in vivo*, genetically-encoded iron sequestration and magnetization. However, there are some similarities in the modifications. For example in my work, the N-terminal addition of a SpyTag peptide (AHIVMVDAYKPTK) to the *E. coli* ftnA ferritin increased iron sequestration and magnetism *in vivo* through potentially enlarging the ion transit channels of the cage via creating steric hindrance between monomers in the cage. Hence it is possible that for the *in vitro* prepared ferritins used for cell-targeting or drug delivery, the addition of bulky peptides or functional groups near the ion channels of the ferritin, or potentially also the point mutations discovered to enhance mineralization and magnetism in this study, could improve iron mineralization and magnetism via similar mechanism. Conversely, the targeting peptides, if added to the ferritins of this study

expressed *in vivo*, could endow them with both enhanced magnetism useful for noninvasive reporting as well as specific biological targeting abilities. However, due to the drastically different reaction conditions for iron mineralization *in vivo* versus *in vitro* (e.g. temperature, pH, reagents and their concentrations), the potential benefits of specific protein modifications or mutations toward mineralization and magnetism of ferritins in a different environment would need to be confirmed via direct comparisons of the physical and chemical properties between the modified and wildtype ferritins in that same environment. If verified, these results could lead to development of multi-modal, genetically-encoded particles useful for cell-specific imaging and therapeutic applications.

Lastly, results of two preliminary experiments exploring the mammalian applications of engineered magnetic ferritins specifically for *in vivo* MRI imaging (in mouse gut) and actuation (of ferritin-attached TRPV1 channels in HEK293) are included in Appendix B. The feasibility of magnetic actuation of protein particles (i.e. “magneto-genetics”) was analyzed with first-principles physics by a recent publication by Professor Markus Meister[19]. Meister showed that for the experiments in animals published this year that demonstrated magnetic control of neuronal ion channels via ferritins[33,34], the force or energy that can be generated by ferritin particles in those setups are at least 4 to 10 orders of magnitude lower than required to affect the ion channels, or even compared to the energy of thermal fluctuation (kT=4.11E-21 J at room temperature of 298K). This back-of-envelop result suggests not only that the published results demonstrating magneto-genetic remote-control of neurons *in vivo* might be due to other non-magnetic mechanisms, but also that the barrier for even future engineering or optimization of the ferritin constructs toward the goal of “magneto-genetics” could be very substantial. Taking these conclusions at

face value, how could one explain that the mutated ferritins expressed in the *E. coli* cells here overcome thermal fluctuation to exert force great enough to move the cells toward permanent magnets (Figure 4a) or retain them in high-gradient magnetic columns against gravitational flow (Figure 2)? There are certain assumptions leading to Meister's final numerical results that are worth exploring here especially pertaining to the experimental setup in this chapter, and Meister has similarly acknowledged some of these toward the end of the Discussions section of his paper.

The total magnetic energy or force in the system depends on the total sample magnetic moment and the strength of the applied magnetic field and its gradient. First the magnetic field in this study was applied via N52 grade NdFeB permanent magnet (K&J Magnetics, Inc.) in close proximity to the cells (i.e. cells culture dish placed right on top of the magnets, or the cells flow through microporous column charged by the magnets). The magnetic field strength near the surface of the N52 permanent magnet employed is close to half a Tesla, at least 10 times stronger than that reported in the magneto-genetic animal studies analyzed by Meister. More importantly, the gradient of the magnetic field, to which the magnetic force is proportional, decays rapidly as the fourth power of the distance from the magnetic dipole source. Such distance between the magnetic field source and the ferritins is necessarily large for noninvasive magneto-genetic control of the ferritin-containing cells through the head of the mouse applying a magnetic field from the outside, but much smaller for cells cultured directly above permanent magnets (around one millimeter) and much less inside the microporous magnetic column. One order of magnitude difference in this distance is amplified by four orders in the magnetic field gradient or correspondingly in the force. Therefore, the stronger source field combined with the smaller distance from the dipole source can partially make up for

the theoretically weak forces that Meister calculated in the case of the magneto-genetic animal experiments by potentially several orders of magnitude. However for non-invasive applications, large separation distance between the sample and magnet is desired, and strong magnetic field gradients may be difficult to achieve in practice.

On the other hand, the magnetic energy or force is related to the total magnetic moment of the ferritins, which depends on the number, size and magnetic moment of the particles. In the *E. coli* and likely also the HEK293 cells, ferritins were over-expressed in significant amounts (Figure 2d), leading to possible aggregation of the particles (Figure 4b). Depending on the separation of the particles (e.g. twice the thickness of the protein coat around 3-4nm), the aggregated particles could align their moments and behave as a larger, more powerful magnetic dipole. This would be similar to but much less powerful than the aligned chain of desirably large, single domain magnetite particles synthesized by magnetotactic bacteria as a compass powerful enough to move the cell even under the weak magnetic field of the earth (tens of microTeslas)[6]. The effect of protein and particle aggregation, which can be a common phenomenon in protein over-expressing cells, was not explicitly taken into account by Meister except in a 2-dimensional membrane surface monolayer model, possibly because the neuronal cells in the animal studies could not typically express proteins in similarly high amounts as *E. coli* or HEK293.

Besides the number of particles, the size of the typical ferritin particles is constrained by the self-assembled protein shell with an inner diameter around 8-9 nanometers filling up to 4500 iron atoms[15]. However, natural ferritins are often incompletely filled, resulting in a smaller 5-6nm diameter core cited by Meister while assuming a lower fill of 2400 iron atoms per particle. Results in this chapter show

that the best ferritin mutants sequester more iron from the cells without greater protein expression (Figure 2), translating to more iron atoms per particle or greater fill ratio.

Finally for the magnetic moment per particle, the animal studies cited by Meister utilized natural ferritins, which Meister calculated consistently using experimental data from different sources to yield a very modest value of around 2.4E-22 $J/T^2$ for their magnetic susceptibility (assuming paramagnetism). Under a field of 0.05 T in Meister's calculations, this yields a magnetic moment per natural ferritin particle of around 1.2E-23 J/T. In my study, mutagenesis and directed evolution were applied, aiming to improve the ferritins' magnetic susceptibility through potentially altering iron packing and increasing the more magnetic magnetite content from the natural ferrihydrite. I have attempted collaborations with the research group of Ronald Walsworth at Harvard to use the high spatial-resolution nitrogen-vacancy diamond magnetometer[51] to measure the magnetic moment of individual cells with mutant ferritin particles *in vivo*, but so far without definitive results due to poor signal over noise. Without such experimental data on the *in vivo* magnetic moment of the mutant ferritin particles, one could still estimate an upper limit. If all 4500 iron atoms maximally packed into the ferritin particle were aligned (e.g. superparamagnetism at below the Curie temperature), their individual magnetic moments due to their unpaired electrons' quantum spins sum up to yield around 2E-19 J/T per ferritin particle. This upper limit is about 4 orders of magnitude greater than what Meister calculated based on experimental data for natural ferritin particle, which are known to possess anti-ferromagnetic domains in ferrihydrite with anti-aligned spins that cancel out their magnetic moment contributions[2]. However, also due to the small size of the ferritin particle, even an aligned single magnetic domain

would have its total moment flipped by random thermal fluctuation unless aligned by an external magnetic field. In comparison, the magnetite particle in magnetotactic bacteria, which is about 5 times larger than a ferritin core in diameter, remains single-domain with a ferromagnetic moment of around 1.6E-17 J/T per particle[6], which is 2 orders of magnitude larger than the theoretical maximum and 6 orders of magnitude larger than the observed (natural) magnetic moment for a ferritin particle. It is no wonder then that just a dozen of these particles, when aligned, could move entire cells even under the weak magnetic field of the earth. Therefore there is still much room for potentially improving the magnetic moment for biologically synthesized iron oxide particles by both increasing the magnetite content and size of the particle, for example through mutagenesis as employed here upon protein templates like certain viral capsids or encapsulins that make larger particles. This strategy combined with aggregation or physical alignment of the particles as in magnetotactic bacteria (enforced by mamK protein structure[6]), and application of stronger magnetic fields (e.g. stronger source dipole) or gradients (e.g. shorter sample distance from source) could potentially make up for the 5 to 10 orders of magnitude shortfall that Meister carefully calculated for the natural ferritin particles and relatively weak magnetic field gradients in the published animal studies. This combined strategy could allow magnetism to move whole cells (as demonstrated in this chapter, and is naturally accomplished with ease for magnetotactic bacteria) and also potentially move the concept of non-invasive magneto-genetics control of biology closer to reality.

## Materials and Methods

### Construction of *E. coli* knock-out strains

We chose *E. coli* BW25113 as the background to introduce plasmids expressing recombinant proteins or conducting sequential knock-out of genes via P1 transduction from knock-out strains in the KEIO collection of non-essential *E. coli* knock-out strains. Plasmid pCP20 was electroporated into the transduced cells and grown at 37°C on LB agar plate to induce expression of recombinase to flip out the Kanamycin resistance cassette used for transduction selection, as well as loss of the pCP20 plasmid. The genes knocked out include genes that serve as endogenous iron storage proteins (ftnA, bfr, dps and their combinations), genes that serve as exporters of metal cations (fieF, rcnA, zntA and their combinations) and the iron master regulator fur. BW25113 with all ftnA, bfr and dps knocked out served as the background strain for expressing recombinant ferritins so as to minimize the small background of endogenous iron storage proteins. BW25113 with all fief, rcnA, zntA and fur knockous served as the metal over-accumulation strain to minimize metal export and down-regulation of import by fur in a metal-rich environment.

### Construction of recombinant ferritin

*E. coli* ferritin gene ftnA was cloned into a high copy-number plasmid (pUC origin of replication) with rhamnose inducible promoter ($rhaP_{BAD}$, with native *E. coli* transcription factors RhaS and RhaR) and kanamycin resistance cassette via Gibson Assembly. The DNA plasmid was verified by Sanger Sequencing (Genewiz) and transformed into *E. coli* BW25113 cells via electroporation. Expression of ftnA was induced in cells by adding rhamnose to cell culture (maximum 0.2%) during log-

phase growth (OD600~0.4). DNA sequences of the most relevant genes and constructs can be found in Table S2 in Appendix B.

**Directed evolution of magnetic ferritin mutants**

Error-prone PCR (Agilent GeneMorph II Random Mutagenesis Kit) was used to generate random mutants of the *E. coli* ferritin (ftnA) sequence with on average one mutation per copy. The PCR products were cloned into the vector with rhamnose promoter and transformed via electroporation into *E. coli* strain containing the GFP-based genetic iron sensor and triple knockout of endogenous iron sequesterers (ftnA, bfr, dps). This library of cells each expressing a particular ftnA mutant was grown and induced in log-phase with 0.2% rhamnose to initiate high level expression of ferritins and simultaneously exposed to 100uM iron (II) sulfate as supplement. The cells were cultured with good oxygenation in deep-well plates shaken at 900rpm. In saturation phase (~20 hours post metal-exposure), the cells were filtered through the magnetic column (Miltenyi LD column) placed between neodymium permanent magnets (K&J Magnetics Inc., BX8C4-N52) to create a high magnetic field gradient that help to retain cells that are more magnetic in the column whereas the cells expressing less magnetic ferritin mutants are flushed out. Subsequently, the permanent magnets are removed and the previously retained magnetic cells were eluted separately, grown again to log-phase, and re-induced and exposed to iron in fresh media to iterate another day of growth and magnetic selection. In total, 10 days or iterations of magnetic selection were conducted before the final eluted magnetic cells were plated. Around 100 colonies were picked and sequenced by Sanger sequencing (Genewiz). A Python script was used to compare the colonies' ferritin sequences to the wildtype ftnA sequence, translate the DNA codons into amino

acids *in silico*, and list the most frequent or representative amino acid mutations. The top mutations were re-constituted in wildtype ftnA plasmid (NEB Q5 Site-Directed Mutagenesis Kit) and verified for increased iron sequestration (via GFP-based genetic iron sensor) and magnetic column retention relative to overexpressing the wildtype ftnA.

**Magnetic Column Retention characterization**

A high-gradient magnetic column (Miltenyi LD columns) was sandwiched between two neodymium permanent magnets (K&J Magnetics Inc., BX8C4-N52) to create high magnetic field gradients inside the column. The column is first wetted by passage of 2ml of PBS 1X buffer. Then 500µl of cells re-suspended in PBS 1X buffer were added and flowed through by gravity into the elution tube, followed by addition of 3ml of PBS 1X buffer to wash through any unbound cells into the elution tube. Once dry, the column is removed from the permanent magnets, and 3ml of PBS 1X buffer is pushed through the column to extract the magnetically retained cells into a separate retention tube. Measuring OD600 of the elution and retention tubes allow estimation of cell counts and the percentage of total cells retained by the magnetic column.

**Iron level characterization by genetic sensor**

For the genetic iron sensor, the *E. coli* fiu promoter was cloned along with a super-folder GFP (sfGFP) reporter via Gibson Assembly into a low copy (p15A origin), chloramphenicol-resistance plasmid compatible with the ferritin-expressing plasmid. Iron levels were measured for cells containing the ferritin-expression and iron sensor plasmids by taking the GFP fluorescence of the culture of cells (488nm excitation by

laser, 512nm emission) in 96-well plate format using the BioTek NEO plate-reader. For calibration, known concentrations of iron sequesterer bi-pyridine were added to cell cultures. The fluorescence measured were normalized to culture density by dividing by OD600 measured by the same plate-reader. The increase in normalized fluorescence of the cells was plotted against the increase in bipyridine (or consequent decrease in free iron) and modeled (Fig S2) to determine the conversion between fluorescence reading and free iron concentration.

**Iron level characterization by Inductively-Coupled Plasma Mass Spectrometry (ICP-MS)**

Cells to be measured were resuspended in 20% ultrapure nitric acid and mixed overnight (>12hrs) to lyse and digest organic material. The solution is then fed to Perkin Elmer 6100 ICP-MS machine (Trace Metal Lab, Harvard School of Public Health) to determine the concentration of a variety of elements (Fe, Co, Ni, Zn, Cd, As, Mg, Mn, U, Li) Prior to sample measurement, machine calibration was performed using solutions containing a linear range of known concentrations of the measured metals.

**Magnetic Resonance Imaging (MRI)**

MRI characterization was performed on a Brucker 9.4T MRI system at the MRI preclinical core of the Beth Israel Deaconess Medical Center. Cells were first washed of metal in culture and resuspended in metal-free PBS 1X buffer and subsequently loaded into NMR tubes. The NMR tubes were further immersed in DI water in a container to prevent artefacts due to susceptibility mismatch at the glass/air interface. Once samples were loaded into the MRI system, T2* weighted

relaxation times were obtained from fits of the signal decay after a gradient-echo pulse sequence and compared among samples. MRI imaging of mice *in vivo* was performed using the same instrument. T2-weighted slices were obtained using spin-echo sequences.

**Superconducting QUantum Interference Device (SQUID) Magnetometry**

Cells were first washed by DI water three times to remove excess iron, then pelleted by high-speed centrifuge (4000g). The cell pellets were flash-frozen by dipping into liquid nitrogen, followed by sublimation in a lyophilizer for one hour to remove water. The desiccated pellets were weighed and loaded into the Quantum Design MPMS SQUID and the sample magnetization (moment) was measured as magnetic field was varied in steps.

**Imaging *E. coli* culture over permanent magnets**

*E. coli* cells induced to express ferritins were grown to saturation in LB media supplemented with 100μM iron and back diluted 1:5 with fresh LB media with 20% OptiPrep Density Gradient Medium to help suspend the cells. The cultures were dispensed into wells of a 6-well titer plate, where ring magnets were placed directly under the bottom of the wells, and left undisturbed. Plates were imaged by a digital camera after overnight growth to capture visible aggregation of cells toward the magnets.

**TEM of *E. coli* cross sections**

*E. coli* cells in culture were resuspended in PBS 1X and fixed with glutaraldehyde overnight. The samples were subsequently dried, embedded in resin, and sectioned

without additional heavy metal staining. The sections were imaged using the Tecnai $G^2$ Spirit BioTWN microscope at variable acceleration voltage and magnification to maximize contrast of the electron dense nanoparticles in the cell cross-sections.

**SDS gel analysis of protein expression levels**

*E. coli* cells were resuspended in SDS Buffer (NuPAGE LDS Buffer) with reducing agent, followed by two cycles of boiling at 95°C for 5 minutes and vigorous vortexing to lyse cells and denature proteins. The lysate was centrifuged to pellet cell debris, and the protein suspension was diluted and added to NuPAGE 4-12% Bis-Tris gel with MES buffer. For calibration of protein concentration via densitometry analysis, known dilutions of BSA protein in the same buffer and similarly denatured were added to the same gel containing the samples. Empty lanes in the gel were filled with equal volume of SDS buffer. After running at 200V for 35 minutes, the gel was removed and stained with Coomassie Orange dye for one hour and subsequently imaged for dye fluorescence on a Typhoon Imager. Densitometry analysis of the gel bands was conducted using ImageJ.

**Bio-magnetism in Mammalian Cells (Appendix B)**

Tagged ferritins (human codon optimized) or tagged TRPV1 (native) containing plasmids were first co-transfected into HEK293-F freestyle cells using the FreeStyle MAX System for transfection (Thermo Fischer) following its standard protocols. After 3 to 4 days of liquid culture growth (30ml in 125ml volume plastic flask shaking at 37°C), the cells were sampled and their fluorescence were measured on a plate reader or imaged on a confocal microscope. Strong magnetic field was applied by placing the plate containing the cells directly over N52 Neodymium permanent

magnetics (surface magnetic field >0.5 Tesla) for 5 to 10 minutes, with separation distance of 1 to 2 mm between the magnets and settled cells.

## References


1. Lloyd, J. R., Byrne, J. M. & Coker, V. S. Biotechnological synthesis of functional nanomaterials. *Curr. Opin. Biotechnol.* **22,** 509–515 (2011).

2. Jutz, G., Van Rijn, P., Santos Miranda, B. & Boker, A. Ferritin: A versatile building block for bionanotechnology. *Chem. Rev.* **115,** 1653–1701 (2015).

3. Xia, Y., Xiong, Y., Lim, B. & Skrabalak, S. E. Shape-controlled synthesis of metal nanocrystals: simple chemistry meets complex physics? *Angew. Chem. Int. Ed. Engl.* **48,** 60–103 (2009).

4. Stürzenbaum, S. R. *et al.* Biosynthesis of luminescent quantum dots in an earthworm. *Nat. Nanotechnol.* **8,** 57–60 (2013).

5. Komeili, A. Molecular mechanisms of compartmentalization and biomineralization in magnetotactic bacteria. *FEMS Microbiol. Rev.* **36,** 232–55 (2012).

6. Faivre, D. & Schüler, D. Magnetotactic bacteria and magnetosomes. *Chem. Rev.* **108,** 4875–98 (2008).

7. Staniland, S. *et al.* Controlled cobalt doping of magnetosomes in vivo. *Nat. Nanotechnol.* **3,** 158–162 (2008).

8. Gossuin, Y. *et al.* Looking for biogenic magnetite in brain ferritin using NMR relaxometry. *NMR Biomed.* **18,** 469–472 (2005).

9. Edelman, N. B. *et al.* No evidence for intracellular magnetite in putative vertebrate magnetoreceptors identified by magnetic screening. *Proc. Natl. Acad. Sci. U. S. A.* **112,** 262–7 (2015).

10. Eder, S. H. K. *et al.* Magnetic characterization of isolated candidate vertebrate magnetoreceptor cells. *Proc. Natl. Acad. Sci. U. S. A.* **109,** 12022–7 (2012).

11. Kirschvink, J. L., Kobayashi-Kirschvink, a & Woodford, B. J. Magnetite biomineralization in the human brain. *Proc. Natl. Acad. Sci. U. S. A.* **89,** 7683–7687 (1992).

12. Honarmand Ebrahimi, K., Hagedoorn, P. L. & Hagen, W. R. Unity in the biochemistry of the iron-storage proteins ferritin and bacterioferritin. *Chem. Rev.* **115,** 295–326 (2015).

13. Le Brun, N. E., Crow, A., Murphy, M. E. P., Mauk, A. G. & Moore, G. R. Iron core mineralisation in prokaryotic ferritins. *Biochim. Biophys. Acta - Gen. Subj.* **1800,** 732–744 (2010).

14. Ga, N. *et al.* Comparative Structural and Chemical Studies of Ferritin. *J. Am. Chem. Soc.* **130,** 8062–8068 (2008).

15. Theil, E. C., Tosha, T. & Behera, R. K. Solving Biology's Iron Chemistry Problem with Ferritin Protein Nanocages. *Acc. Chem. Res.* **49(5),** 784-791 (2016). doi:10.1021/ar500469e

16. Theil, E. C. & Behera, R. K. The Chemistry of Nature's Iron Biominerals in Ferritin Protein Nanocages in *Coordination Chemistry in Protein Cages: Principles, Design, and Applications* 3–24 John Wiley & Sons (2013).

17. Bradley, J. M., Moore, G. R. & Le Brun, N. E. Mechanisms of iron mineralization in ferritins: one size does not fit all. *J. Biol. Inorg. Chem.* **19,** 775–85 (2014).

18. Chasteen, N. D. Functionality of the Three-Site Ferroxidase Center of Escherichia coli Bacterial Ferritin (EcFtnA). *Biochemistry* **53,** 483–495 (2014).

19. Meister, M. Physical limits to magnetogenetics. *Elife* **5,** (2016). doi: 10.7554/eLife.17210

20. Brooks, R. A., Vymazal, J., Goldfarb, R. B., Bulte, J. W. M. & Aisen, P. Relaxometry and magnetometry of ferritin. *Magn. Reson. Med.* **40(2),** 227–235 (1998).

21. Nishida, K. & Silver, P. A. Induction of biogenic magnetization and redox control by a component of the target of rapamycin complex 1 signaling pathway. *PLoS Biol.* **10,** (2012).

22. Stanley, S. a, Sauer, J., Kane, R. S., Dordick, J. S. & Friedman, J. M. Remote regulation of glucose homeostasis in mice using genetically encoded nanoparticles. *Nat. Med.* **21,** 92-98 (2014). doi:10.1038/nm.3730

23. Weissleder, R., Nahrendorf, M. & Pittet, M. J. Imaging macrophages with nanoparticles. *Nat. Mater.* **13,** 125–38 (2014).

24. Ahrens, E. T. & Bulte, J. W. M. Tracking immune cells in vivo using magnetic resonance imaging. *Nat. Rev. Immunol.* **13,** 755–63 (2013).

25. Conniot, J. *et al.* Cancer immunotherapy: nanodelivery approaches for immune cell targeting and tracking. *Front. Chem.* **2,** 1–27 (2014).

26. Naumova, A. V, Modo, M., Moore, A., Murry, C. E. & Frank, J. a. Clinical imaging in regenerative medicine. *Nat. Biotechnol.* **32,** 804–818 (2014).

27. Colombo, M. *et al.* Biological applications of magnetic nanoparticles. *Chem. Soc. Rev.* **41,** 4306-4334 (2012).

28. Matsumoto, Y., Chen, R., Anikeeva, P. & Jasanoff, A. Engineering intracellular biomineralization and biosensing by a magnetic protein. *Nat. Commun.* **6,** 8721 (2015).

29. Wang, Q., Mercogliano, C. P. & Löwe, J. A ferritin-based label for cellular electron cryotomography. *Structure* **19,** 147–154 (2011).

30. Gleich, B. & Weizenecker, J. Tomographic imaging using the nonlinear response of magnetic particles. *Nature* **435,** 1214–7 (2005).

31. Glenn, D. R. *et al.* Single-cell magnetic imaging using a quantum diamond microscope. *Nat. Methods* **12,** 1–5 (2015).

32. Chen, R., Romero, G., Christiansen, M. G., Mohr, A. & Anikeeva, P. Wireless magnetothermal deep brain stimulation. *Science* **347,** 1477–80 (2015).

33. Wheeler, M. A. *et al.* Genetically targeted magnetic control of the nervous system. *Nat Neuroscience* **19,** 756-761 (2016).

34. Stanley, S. a. *et al.* Bidirectional electromagnetic control of the hypothalamus regulates feeding and metabolism. *Nature* **531,** 647–650 (2016).

35. Zhang, Z. *et al.* Functional Interactions between the Carbon and Iron Utilization Regulators , Crp and Fur , in Escherichia coli Functional Interactions between the Carbon and Iron Utilization Regulators , Crp and Fur , in Escherichia coli. *J. Bacteriol.* **187,** 980–990 (2005).

36. McHugh, J. P. *et al.* Global Iron-dependent Gene Regulation in Escherichia coli: A new mechanism for iron homeostasis. *J. Biol. Chem.* **278,** 29478–29486 (2003).

37. Carter, K. P., Young, A. M. & Palmer, A. E. Fluorescent Sensors for Measuring Metal Ions in Living Systems. *Chem. Rev.* **114,** 4564–4601 (2014).

38. Atpases, P. *et al.* Structure and mechanism of Zn2+-transporting P-type ATPases. *Nature* **514,** 518–522 (2014).

39. Waldron, K. J. & Robinson, N. J. How do bacterial cells ensure that metalloproteins get the correct metal? *Nat. Rev. Microbiol.* **7,** 25–35 (2009).

40. Wong, S. G. *et al.* The B-type channel is a major route for iron entry into the ferroxidase center and central cavity of bacterioferritin. *J. Biol. Chem.* **290,** 3732–3739 (2015).

41. Lundin, D., Poole, A. M., Sjöberg, B. M. & Högbom, M. Use of structural phylogenetic networks for classification of the ferritin-like superfamily. *J. Biol. Chem.* **287,** 20565–20575 (2012).

42. McHugh, C. A. *et al.* A virus capsid-like nanocompartment that stores iron and protects bacteria from oxidative stress. *EMBO J.* **33,** 1–16 (2014).

43. Schoonen, L. & van Hest, J. C. M. Functionalization of protein-based nanocages for drug delivery applications. *Nanoscale*. **6**, 7124-41 (2014).

44. Uchida, M. et al. Targeting of cancer cells with ferrimagnetic ferritin cage nanoparticles. *J. AM. CHEM. SOC.* **128(51)**, 16626-33 (2006).

45. Zhen, Z. et al. RGD modified apoferritin nanoparticles for efficient drug delivery to tumors. *ACS Nano.* **7(6)**, 4830-4837 (2014).

46. Vannucci, L. et al. Selective targeting of melanoma by PEG-masked protein-based multifunctional nanoparticles. *Int J. Nanomedicine.* **7**, 1489-1509 (2012).

47. Li, X. et al. Epidermal growth factor-ferritin H-chain protein nanoparticles for tumor active targeting. *Small*. **8**(16), 2505-14 (2012).

48. Kwon, C. et al. Development of protein-cage-based delivery nanoplatforms by polyvalently displaying β-cyclodextrins on the surface of ferritins through copper(I)-catalyzed azide/alkyne cycloaddition. *Macromolecular Bioscience*. **12(11)**, 1452-58 (2012).

49. Kang, H. J. et al. Developing an antibody-binding protein cage as a molecular recognition drug modular nanoplatform. *Biomaterials*. **33(21)**, 5423-30 (2012).

50. Kang, Y. J. et al. Incorporation of thrombine cleavage peptide into a protein cage for constructing a protease-responsive multifunctional delivery nanoplatform. *Biomacromolecules*. **13**, 4057-64 (2012).

51. Sage, D.L. et al. Optical magnetic imaging of living cells. *Nature* **496**, 486-489 (2013).

## CHAPTER 4

# ARTIFICIAL NEURAL NETWORKS FOR ACCURATE PROTEIN FUNCTION PREDICTION FROM SEQUENCE

## Preface

In this Chapter I introduce the application of machine learning, in particular employing artificial recurrent neural networks, combined with experimental validation to identify new proteins with particular functions. Supporting materials for this chapter are found in Appendix C.

## Abstract

Accurate prediction of the function of a protein directly from its primary amino-acid sequence has been a long standing challenge in biology. Most popular algorithms for annotating protein functions or discovering remote functional homologies rely on performing sequence alignments using heuristic scoring or probability modelling. Here machine learning using artificial neural networks (ANN) models was applied towards classification of protein function directly from primary sequence without feature engineering. The recurrent neural networks (RNN) with long-short-term-memory (LSTM) units scanning over the amino-acid sequences trained on public, annotated datasets from UniProt achieved high prediction performance for in-class prediction of four different functions, particularly compared with machine learning algorithms using sequence derived protein features. RNN models demonstrated lower sensitivity and selectivity for “out-of-class” prediction on phylogenetically distinct protein families with similar functions. Applying the trained models toward prediction of function on the UniRef100 database generated not only candidates validated by existing annotation, but also currently unannotated sequences. Several such high confidence predictions for the iron-sequestration function were experimentally validated to be iron-mineralizing even though their sequence differ significantly from known, characterized proteins and from each other. As sequencing and experimental characterization data increases rapidly, the

machine-learning approach based on RNN could be useful for discovery and prediction of homologues for a wide range of protein functions.

**Introduction**

As the cost of DNA sequencing is decreasing drastically over the last decade, the volume of biological sequences particularly for new proteins is also increasing rapidly. Discovering the functions of these new proteins not only could allow one to better understand their roles in their native contexts, but also utilize them in synthetic biology as parts that can be assembled into biological circuits and pathways for useful applications such as production of valuable compounds or treating disease. However, the experimental characterization of proteins' properties such as structure and function is currently still slow and resource-demanding using techniques such as x-ray crystallography, cryo-TEM, or recombinant protein expression and functional assays, significantly outpaced by sequencing. It is unrealistic to experimentally characterize all new protein sequences, and a predictive pipeline that can link sequencing to function to allow filtering of the vast sequence dataset to a subset of candidates of highest interest toward a particular function or application is greatly desired.

Currently there are several ways for extracting useful information from primary sequences and infer functional information based on comparison of new sequences to existing sequences of known function. One popular route is based on sequence alignment of the new and existing sequences with heuristic scoring. Prominent examples of this approach includes BLAST and FASTA. In BLAST, the score and "E-value" can indicate the degree of homology between the new sequence and existing sequences. A position-scoring matrix complemented BLAST search performed iteratively (PSI-BLAST) with appropriate "E-value" cutoff may allow discovery of

more remote homologs to the new sequence in query. Another popular algorithm inspired by works in natural language processing is the use of Hidden-Markov Models (HMM) for building “profiles” based on multiple sequence alignments to develop a probabilistic model for sequence evolution[1]. This allows the discovery of protein clusters or families that are evolutionarily related, and new query sequences may be aligned to existing profiles for identification of function. Powerful and popular as these existing approaches are for protein function annotation directly from sequence, there are limitations. Most of these methods rely on good quality sequence alignment and could fail to classify sequences coding for proteins with similar function or structure but are distant in evolutionary scale or have come to adopt similar function via convergent evolution. Notable example of convergent evolution of protein function is the protease, which has independently evolved the “catalytic triad” active site in 23 protein superfamilies. The “catalytic triad” consists an acid-base-nucleophile configuration of three amino acids arranged in spatial proximity but can be distant on the sequence. Given the difficulty in accurate prediction of three-dimensional protein folding, the “catalytic triad” is difficult to predict based on alignment of sequences.

Machine learning, a field of artificial intelligence and computer science developing diverse algorithms for capturing complex patterns from data, has become increasingly applied toward solving biological problems[2–15], and particularly has potential for complex biological problems such as protein structure and function prediction. Instead of constructing *ab initio* models of protein folding based on laws of physics as in molecular dynamics simulations to infer function, machine learning algorithms could learn complex sequence features from data associated with certain functions or labels. In supervised machine learning, the model training process

involves exposing the algorithm to data containing protein examples each pre-assigned with a particular function, determined either experimentally or via high-confidence computational predictions using tools such as HMMER. Each protein example contains a number of “features”, which may be properties of a particular protein such as its molecular weight or hydrophobicity or may just be the raw sequence of amino acids to avoid *a priori* assumptions of which set of protein properties may be necessary and sufficient for protein function. The functional labels of the protein examples are treated as dependent variables while the features are treated as independent variables whose weights or coefficients are adjusted during training to minimize discrepancy or error between predicted and true functions and produce the best fit. For the classification task of determining functional labels for proteins, several supervised machine learning methods are well-suited including logistic regression, random forest, support vector machines and artificial neural networks.

Logistic regression is a fast and reliable algorithm for classification. Similar to the common linear regression, logistic regression fits a dependent variable as a linear combination of feature variables. The key difference is that the dependent variable in logistic regression is not the target itself as a continuous variable, but rather the logit transformation of the probabilities of the labels (i.e. log odds ratio). The predicted label is assigned based on a probability threshold. Due to the non-linear transformation, the best parameters are estimated during the training process based on maximum likelihood estimation. Compared to other machine learning algorithms, the main advantages of logistic regression are speed and interpretability, as the magnitudes of the feature coefficients allows estimation of the relative feature importance.

Random forest is a robust and accurate algorithm for classification, particularly for examples where the relationship between input and output variables is hardly linear. Random forest implements an ensemble of decision trees, where each tree samples with replacement a subset of the total training samples and "votes" for a particular classification label. The label with the most votes from the ensemble becomes the final classification decision. Each decision is consisted of nodes which recursively split the dataset based on subset of the input features with thresholds in order to maximize segregation of the different classes in each step. Even though slower than logistic regression, random forest can be more powerful at capturing certain non-linear relationships and covariance in data and are also easily interpretable based on the importance values assigned to the features. Furthermore, random forest models applies bootstrap sampling for internal validation and is robust toward two common issues of datasets for machine learning: missing data and class imbalance (i.e. over-representation of one label or class versus others in the training data which can result in high-bias models that try to optimize accuracy by "ignoring" the minority class). However, decision trees models can suffer from high variance or overfitting, where the model fails to generalize to new datasets. Tuning the hyper-parameters of random forest such as the minimum samples per split or maximum tree depth to limit model complexity can mitigate overfitting.

Support vector machines represent another powerful classification algorithm that determines hyper-planes in high-dimensional feature-space that segregate the examples of different classes with maximum "margins". A non-linear kernel transformation such as using the radial basis function (rbf) can be applied to transform into a higher-dimensional space to obtain highly non-linear boundaries that best separate the example feature vectors. This "kernel trick" allows efficient

computation of SVM inner products without explicit mapping of each example feature vector into the higher dimensional space. Despite their slower performance especially using non-linear kernels, SVMs have performed well for classification tasks such as handwriting recognition.

Artificial neural network (ANN) is a machine learning algorithm inspired by the organization of the biological neural network to learn from data and perform prediction tasks. Its demonstrated high performance in recent years, combined with its flexibility and biological connections has generated significant attention and made the artificial neural network the leading candidate for artificial intelligence. Simplistically, an artificial neural network is composed of layers of neurons including an input layer, an output layer, and one or more hidden layers in between. Each layer has connections with weights to adjacent layers. An artificial neuron in this network receives inputs from neuron(s) in the preceding layer similar to dendrites in a biological neuron. The artificial neuron then performs computation to determine its output, typically represented as a weighted sum of the inputs followed by a mathematical transformation via an “activation function”. The “activation function”, analogous to a biological neuron’s decision to fire an action potential, can be the sigmoid function as in logistic regression, or others like the hyperbolic tangent (tanh) or ReLU function. The edges connecting neurons carry the weights or coefficients for the weighted sum to be activated. These weights are tuned during training of the ANN with labeled examples to minimize the prediction error, which can be represented as mean-square loss for regression or categorical cross-entropy for classification tasks. To minimize the error function, a back-propagation algorithm is applied where the gradients of the error function with respect to the outputs and the weights are propagated backwards in the ANN via chain rule, and weights are tuned

as to minimize the accumulated gradients[15]. Once the weights are optimized, new examples can be propagated in the forward direction through the network with successive matrix multiplications against the weight matrices and activations to produce prediction at the final output layer. The main advantage of the ANN is its ability to model arbitrarily complex input-output relationships, making it flexible for a variety of problems from machine vision to bioinformatics. The main disadvantages are the intensive requirements on computational resources and the potential for model over-fitting (high variance) if the training data size is small while the ANN model size and complexity (i.e. number of independent parameters) are large. For a large network, the numerous multiplications of large matrices can be slow to carry out in a serial manner using traditional central processing units (CPUs) with few threads. However, the wider adoption of graphic processing units (GPUs) with massively parallelized computation has led to significant gains in computing speed particularly for training deep ANNs with numerous hidden layers. Concurrently, the issue with model over-fitting is ameliorated by the explosive growth in the quantity of data particularly from sources as the Internet. Larger training datasets help decrease variance of complex ANN models to yield better predictions for data unseen. In machine vision for instance, optimized, deep ANNs in recent years are performing above human level at image recognition, a task traditionally perceived to be difficult for machines. In addition to employing GPU computing for speed and large sets of labelled image data from the internet and social media for training, this high-performance machine vision feat is also enabled by critical advances in deep ANN architecture. One such advance is the utilization of convolutional neural network inspired by the biological organization of the mammalian visual cortex[16]. By exploiting spatial symmetry with filters trained to recognize important patterns from

images, the complexity of the neural network could be drastically reduced. Simultaneously, the method of “Dropout”[17], a technique partially motivated by the advantages of biological sexual reproduction, has led to reduced over-fitting of ANNs by dynamically severing random subsets of connections between neurons during training. More recently, new advances in “deep-learning”, or machine learning employing deep neural networks with numerous hidden layers, has become popular for instance in using recurrent neural networks for language modeling and generation. In this application, the sequence of words or characters in the English language are learned by feeding sentences from human-written text to a recurrent neural network (RNN) which maintains memory as it recursively processes a sequence of inputs to keep track of the patterns observed. When unrolled in time, this RNN represents a deep network with as many layers as the number of input words or characters. The complexity of the network can be increased by stacking several levels of these recurrent layers and feeding the output sequence of one recurrent layer as input into the next. As the neural network gets deeper with longer sequences, the issue of vanishing or exploding gradients arises in model training via back-propagation. This is caused by repeatedly multiplying small or large error gradients which can result in zero or very large values for the changes in the front layers to prematurely terminate learning or to prevent convergence. One solution is the Long-Short-Term-Memory (LSTM) network which is capable of remembering long-term patterns[18]. The LSTM network is a special case of RNN where each neuron maintains memory in a “cell state” and with each new input and old output carefully modulates the “cell state” with a number of “gates” to selectively delete or add information (Figure 1). Each gate is mathematically represented as a sigmoid function followed by element-wise multiplication. The sigmoid function limits output to

between 0 which represents not letting any information through, to 1 which represents letting through all information. The LSTM unit first utilizes a “forget gate” between the input (including new data and value of previous hidden state) and the cell state to decide based on the new information how much of the old cell state information to keep. Then the input gate creates the new input transformed by the “input gate” that decides how much of the new information to be later added to the current cell state. Finally to decide the output value at this stage, the “output gate” sigmoid transforms the input and multiplies it by the current cell state (after modification by “forget” and “input” gates) squashed by tanh between -1 and 1 to determine the new output and hidden state values (Methods). Due to its ability to selectively add, remove or retain information in memory as needed, the LSTM has the ability to capture long-range relationships previously hindered by the issues of vanishing or exploding gradients. For example in a long sequence of amino acids that code for a protein, the LSTM network trained with protein sequences and their corresponding functions may remember key amino acids distant from each other in the sequence as important toward a function, since these residues may be in proximity to each other in the folded three-dimensional structure and contribute to an active site of the protein, even though the LSTM network does not have *a priori* knowledge of the structure.

Recurrent LSTM networks have proved powerful in both understanding and generating sequences of human text[19]. For example, with a sufficiently deep LSTM network and training with the complete works of Shakespeare, an LSTM network can compose new passages in the style of Shakespeare. DNA and amino acid sequences represent the language of biology. Our understanding of its “meaning” (i.e. fundamental biology) and ability to write it in a “meaningful” manner (i.e.

synthetic biology) are both currently limited. To accelerate this learning process, here I explore and present an approach to develop artificially intelligent systems enabled by technologies such as recurrent LSTM networks that have gained performance in understanding and writing human language to enable us to better understand and write the language of biology.

**Results**

The deep learning recurrent neural network (RNN) model for protein function prediction is first trained on a large set of protein sequences with certain known functions as labels. The training process tunes the parameters of the network by minimizing prediction errors (categorical entropy). After validating good prediction performance of the trained network using a "test" dataset of randomly chosen sequences of proteins with known functions but have never been seen during training, new sequences with unknown functions are fed to the network to make predictions of function (Figure 1a). The predicted function is eventually validated by experimental assay. Furthermore, RNN models could predict certain evolutionarily distinct "out-of-class" protein families with similar function, although the performance is worse compared to prediction of randomly chosen in-class sequences.

The deep learning model is a recurrent neural network (RNN) with one or more sets of "bi-directional" recurrent layers containing long-short term memory (LSTM) neurons processing the input sequence one residue or character at a time (Figure 1b). The forward layer scans the protein sequences from the N- towards the C-terminus and reversed for the backward layer, allowing the network to make use of context on both sides of each position rather than just what was before it when processing in one direction with a single layer. Each residue in the input protein sequence is converted into a "one-hot" vector whose elements are all 0 except at the position of the amino acid it corresponds to, where it is set to 1. Each LSTM neuron in each recurrent layer uses input "i", output "o", gate "g", and forget "f" gates to modulate the input vector and update the neuron's internal cell state "c" and hidden state "h". The gates apply matrices, whose elements are adjustable parameters to be learned, on the input and hidden state vectors at each recurrent step/layer and

subsequently normalize the results with nonlinear activation functions (Methods). Intuitively, given each new input vector (i.e. sequence residue), the gates control what and how much to add to and output from the hidden state memory, which encodes sequence patterns relevant toward particular protein function (Figure 1c). These features of the LSTM architecture allow the RNN to maintain, over many recurrent iterations, the magnitudes of both the relevant signals in feed-forward propagation as well as the error gradients in backpropagation, thereby resolving the issues of loss of contextual memory and vanishing/exploding gradients that have limited the usefulness of traditional RNNs in processing long sequences (e.g. hundreds of units/iterations). The outputs from the last LSTM neurons of the forward and reverse hidden layers are eventually fed to a fully-connected layer of artificial neurons, where each neuron represents one functional class and outputs via the "softmax" activation function the probability that the input sequence represents a particular functional class. The number of recurrent hidden layers, LSTM neurons in each layer, hidden units (i.e. hidden state vector dimension) in each LSTM neuron, and the architecture of each LSTM neuron (e.g. whether to open "peep-hole" connections") are hyper-parameters that can be optimized. For example, stacking several recurrent layers by feeding the output of each LSTM neuron in one recurrent layer as input into an adjacent recurrent layer, or increasing the number of neurons and hidden units, enable more complex or hierarchical representations at the risk of "over-fitting" or increasing prediction variance of the model. Furthermore, the number of output neurons, which represents the number of functions to be simultaneously considered (i.e. multiplex), can be varied. In this work, a single set of bi-directional recurrent layers was utilized for in-class predictions, and up to three sets were used for training toward out-of-class predictions. As protein sequences vary widely in

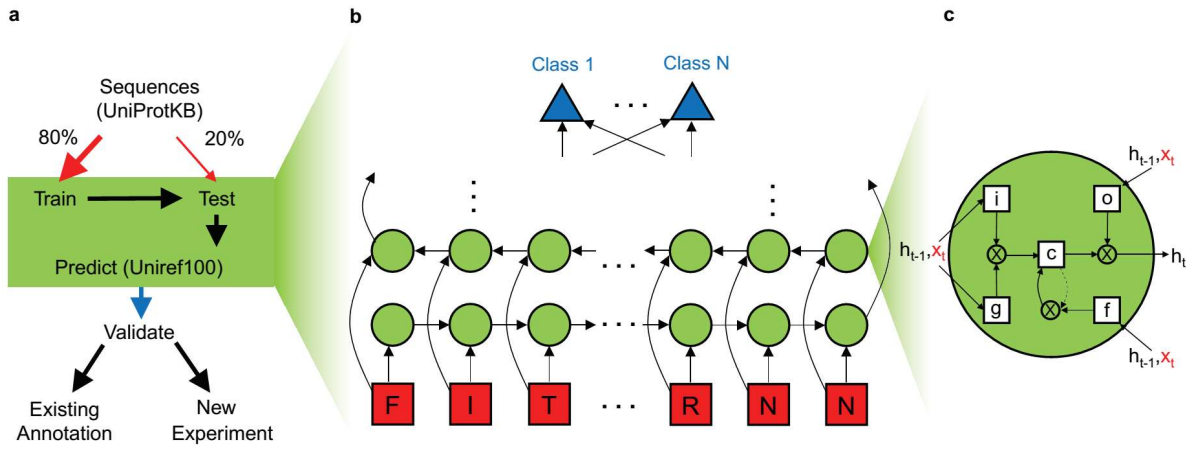



**Figure 1 Machine-learning model for protein function prediction (a)** workflow of the prediction model consists first feeding sequence dataset with known functional annotations. After training the machine-learning Recurrent Neural Network (RNN) model with 80% of the sequences chosen randomly, the last 20% of yet unseen sequences are fed to test the model prediction performance. Alternatively, the model can be tested on sequences of protein families with homologous function but distinct phylogeny from the training set (e.g. “out of class”). The tested model is used to scan and predict all proteins (including unannotated) in the UniRef100 database. The positive predictions are validated either by existing annotation (e.g. in UniProt) or experiment **(b)** The RNN model consists of arbitrary sets of forward and reverse layers of long-short term memory (LSTM) neurons taking only the amino acid letters from the sequence as input (red). The final output of the recurrent layers are combined into a fully connected layer for functional classification (blue) **(c)** Each LSTM neuron contains gates for input “i”, output “o”, gate “g”, and forget “f”, which update along with the new input the cell state “c” and hidden state “h” to encode relevant sequence patterns.

length, the number of LSTM neurons in the recurrent layer was capped, typically at 333 representing a maximum of 333 amino acids sequence or around 1 kilo-base of DNA. For proteins smaller than 333 residues the sequence was pre-padded with 0's up to a 333-digit sequence, where the digits 1 to 20 represents the 20 canonical amino acids. For functions with mostly large proteins such as the CRISPR-associated nucleases, up to 800 N-terminal amino acids were input for training, and subsequently the same RNN model was trained on up to 800 C-terminal amino acids. There are 128 hidden units in each LSTM neuron (i.e. the hidden state is represented by a 128 element vector). A high dimensional hidden state vector can encode more information to represent more complex function-related sequence features. This can be an advantage compared to some sequential models (e.g. hidden Markov model) with limited number of internal states at the expense of interpretability. Additionally, the multiple nonlinear operators of the LSTM (e.g. activation functions) allow complex updating of the hidden state memory. Adding to this flexibility, the probability of "Dropout", the random severing of connections between layers, was consistently set to 0.5. Unlike previous artificial neural network based methods, the LSTM model here does not limit itself to learning short "profiles" or motifs of pre-defined length[7,9,20] (e.g. 21 amino acid window[20]) but instead learns from the entire sequence up to a maximum length (e.g. 333, 500 or 800 from each terminus) in order to capture potentially long-range patterns.

For training, protein amino-acid sequences were obtained from the UniProt database and directly used as inputs into the neural network without any feature extraction. For prediction of a single function, the positive class contains all sequences that match the function in UniProt by keyword. The negative class contains all non-matching sequences in SwissProt (the manually reviewed database

within UniProt with currently around 550K sequences). Of the combined dataset, 80% was randomly selected and employed as “train-set” for training the neural network, and the remaining 20% was used as the “test-set” to evaluate the trained model’s performance on yet-unseen dataset. To balance the positive and negative classes, either the larger sample-size class was split into several chunks for sequential training with the smaller sample-size class, or the smaller sample-size class was over-sampled. The negative class generally greatly outnumbers the positive class and was divided into 4 or more chunks to train against the positive class sequentially. Each chunk of the negative set combined with the positive set was trained for at least 5 epochs (i.e. passes over the entire dataset) during which the categorical entropy of the predicted output compared against the expected output was minimized via the ADAM optimizer[21] with the mini-batch sampling size set to 64. Ten percent of the total data within the training dataset was used to monitor the network losses and changes in prediction accuracy during each training step. Furthermore after each chunk had been trained, the prediction performance on the “test-set” was evaluated to calculate the accuracy, precision, recall and F1 (F-measure) for the positive and negative classes. The “test-set” data was initially selected and mixed at random without applying class-balance (e.g. oversampling) in order to mimic real-life operations when the positive class is heavily under-represented.

Four functional classes were picked to test the performance of the RNN predictive model: iron sequestering proteins (class “Ferritin”), cytochrome P450 proteins (class “P450”), serine and cysteine proteases (class “Protease”) and G-protein coupled receptors (class “GPCR”). Iron sequestration is crucial for cells to maintain iron homeostasis and protect against ROS generation from iron-catalyzed

Fenton reactions[22]. Currently well-known iron sequesters across domains of life are ferritins and dps (DNA-binding protein from starved cells) proteins which form protein cages in cells that sequester and mineralize iron into inorganic nanoparticles. In addition to detoxification, the iron oxide nanoparticles synthesized could potentially be utilized towards noninvasive applications in biology such as a reporter or contrast agent for magnetic resonance imaging. I have further developed assays for measuring and validating iron sequestration and magnetic properties by proteins expressed in *E. coli* bacteria cells (Chapter 3). P450 proteins are also ubiquitous across kingdoms of life and are enzymes that act on a variety of substrates carrying out important tasks including detoxifying drugs in humans. G-protein coupled receptors are important trans-membrane proteins for cellular signal transduction and are targets for many drugs. Lastly, serine and cysteine proteases cleave peptide bonds in proteins to break them down and represent a prime example of molecular-scale convergent evolution, where different organisms independently evolved the "catalytic triad" for performing the peptide cleavage function with otherwise little homology at the overall protein sequence level.

High performance of predictions on the randomly left-out "test-set" data not seen by the model during training was obtained for all four classes of protein functions (Figure 2a). Even though accuracy is nearly 100% for all predictors, it is not the most informative measure as the negative class of proteins not possessing a particular function vastly outnumbers the positive class and a predictor could achieve high accuracy by simply only predicting negatives. But despite such challenge of "finding needle in a haystack", all functional predictors were able to achieve close to unity precision and recall in identifying the correct sequences from the "test-set", with F1 measure close to unity. The receiver operating characteristic (ROC) plots for the

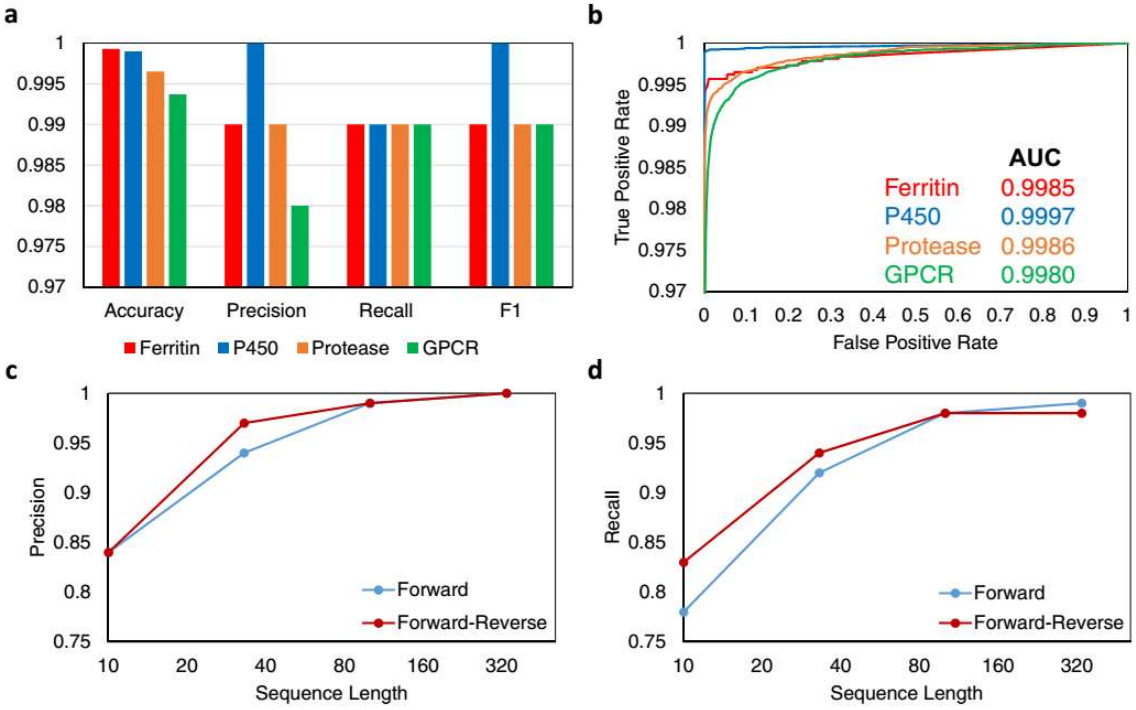



**Figure 2 RNN model achieves high prediction performance on randomly left-out testing data (a)** High prediction performance is achieved for all four tested classes: iron-sequestering (Ferritin), cytochrome P450, protease (serine and cysteine) and G-protein coupled receptor (GPCR). **(b)** Receiver-operating characteristic (ROC) of the four separate models demonstrate high Area-Under-the-Curve (AUC). For the “Ferritin” class, prediction precision **(c)** and recall **(d)** both improve to close to unity as the length of amino acid sequence shown the network increases, saturating around 333 letters.

True Positive Rate (sensitivity) versus False Positive Rate as a function of the classification threshold (between 0 and 1) and their Area Under the Curve (AUC) close to unity also demonstrates the model's ability to make strong discrimination of the positive class distribution from that of the negative class in the tested dataset (Figure 2b). However, it is important to note that these metrics do not readily apply to prediction on arbitrary datasets, particularly largest databases where class imbalance (ratio of negatives to positives) is extreme due to the negligible fraction of total proteins that have one specific function. Therefore the machine learning model performance may be negatively impacted, producing more false positive and negative predictions than expected based on the metrics reported here. Also very low false positive rate (e.g. 1E-6) would be needed to avoid large number of false positives when searching a large database (e.g. 54 million sequences in UniRef100). Lastly, as anticipated, the prediction performance in precision and recall decreases as the cutoff length of the input sequence or equivalently the depth of the bi-directional recurrent layer was decreased as demonstrated for the "Ferritin" class (Figure 2c, d), even though reducing neural network depth increases training speed. Allowing input sequence length greater than 333 amino acids significantly increases processing and memory requirements without yielding noticeable increases in prediction performance for the four protein functions of interest.

The trained and performance-validated models were used to predict whether a new sequence without assigned function could possess a potential function. Currently, state-of-art tools for remote homology search include PSI-Blast, Delta-Blast and in particular jackhmmer (part of HMMER[1]) which utilizes hidden Markov models. For comparison, HMMER and the RNN models were run on the same comprehensive UniRef100 sequence database containing numerous

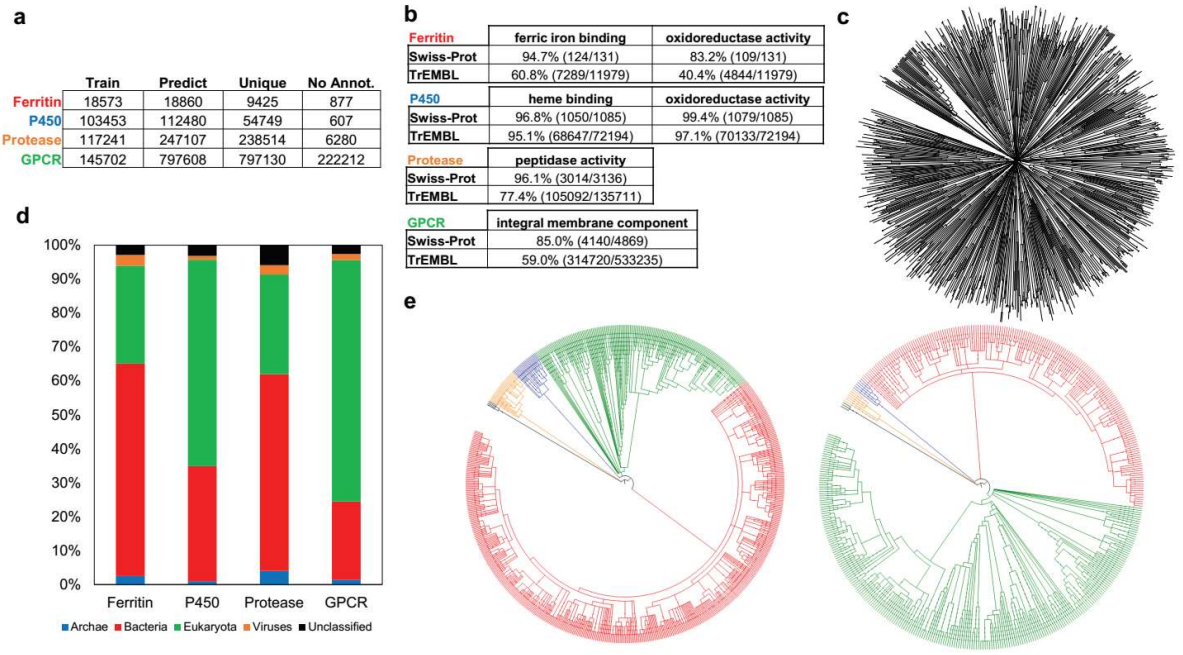



**Figure 3 Trained RNN model predicts new annotations (a)** Table listing for each function, the number of sequences used for training the RNN model, the number of *additional* sequences it predicted as positive in the UniRef100 database (“Predict”), the number of sequences not included in output of jackhmmer (iterative HMMER search) using representative starting sequence (“Unique”), and the number of sequences without any function or family annotation on UniProt and linked databases (Gene3D, InterPro, PROSITE, Pfam, SUPFAM) (“No Annot.”). **(b)** Among the “Predict” sequences, high percentages agree with manually curated Swiss-Prot annotation for expected gene ontology of each class. Agreement is worse for the automatic annotations in TrEMBL database particularly for “Ferritin” and “GPCR” functions. **(c)** Clustal Omega multiple sequence alignment of the “No Annot.” sequences for “Ferritin” function shows diverse lineages. **(d)** Taxonomy of “Predict” proteins reveals expected bias for functional class. **(d)** Taxonomy of the organism of origin for the “Predict” proteins for “Ferritin” (*left*) showing greater species diversity among bacteria (*red*) and “P450” (*right*) showing greater diversity among eukaryotic species (*green*).

uncharacterized or unannotated proteins sequences. For each of the four functional classes (Ferritin, P450, Protease, GPCR), a representative or important member was used (FTNA_ECOLI, CP21A_HUMAN, SEPR_HUMAN, FFAR2_HUMAN), respectively, as initial seed for iterative HMMER search on the UniRef100 database, and at least 5 iterations were run with a reporting cutoff threshold of e-value E=10.0 (default). Separately, the trained RNN models also predicted thousands of new hits from the UniRef100 database for each function (Figure 3a, “Predict”) *in addition to* the thousands of sequences that were used for training each model. Upon comparing the lists of outputs from HMMER to the RNN models discounting the already-annotated sequences used for training, there were still thousands of new, unique sequences predicted by the RNN model that were not shared by the HMMER output (Figure 3a, “Unique”). As a check, the majority of additional sequences predicted by the RNN model have some identification of the correct family or gene ontology in a public database obtained by other sequence or structure homology detection techniques (Figure 3b). However, there is a further subset of the predicted sequences that are unannotated and uncharacterized in UniProt (Figure 3a, “No Annot.”). For the “Ferritin” class, the “No Annot.” sequences predicted by the RNN show numerous lineages after multiple sequence alignment by Clustal Omega (using EMBL-EBI server), suggesting a set of diverse, dissimilar sequences not sharing obvious sequence patterns identifiable by alignment (Figure 3c). The statistics of the domains of origin for the predicted proteins reveal certain domain biases for function, such as bacteria for “Ferritin” class or eukaryote for P450 and GPCR, as expected (Figure 3d). Similar biases could be seen for the “Ferritin” and “P450” classes in the taxonomy of the organisms of origin for the predicted proteins (Figure 3e).

To validate the functional prediction by the RNN model of sequences without characterization or annotation in UniProt, I experimentally characterized the iron sequestration properties of ten high-scoring, “unique” predictors by the RNN model for iron sequestration proteins. The ten sequences were selected from diverse domains of life and vary widely in their amino acid lengths and composition and were predicted to be iron-sequestering with high confidence (Figure 4a). The candidates were named after their biological contexts. Homology search with these sequences as seeds using popular bioinformatics tools such as BLAST and jackhmmer using their web servers on the latest protein databases (NCBI nr, Reference Proteomes) yielded mostly proteins of unknown (only “predicted” or “hypothetical”) and uncharacterized function. However, some functional homologues were identified. For the “fungi” candidate, both web-based BLAST and jackhmmer were able to detect “ferritin-like” homologues, corroborating the RNN prediction. On the other hand, candidates “human”, “mouse”, “potato”, “cyano”, “gut” or their BLAST/jackhmmer homologues showed few entry names suggestive of other functions such as “Alternative protein NCAM1 (neural cell adhesion molecule)” for the “human” candidate and “poly-homeotic like protein” for “mouse” candidate. This could have new implications for the biological activity, particularly of iron sequestration, for these uncharacterized sequences. The remaining candidates “lancelet”, “virus”, “algae” and “archaea” yielded no hint of protein function. The DNA sequences encoding all 10 uncharacterized proteins were codon optimized, synthesized, and cloned into vectors in *E. coli* cells and expressed highly using a rhamnose-inducible, high copy number vector (the N-terminal six methionine repeat sequence of “human” was synthesized with only the last methionine due to DNA synthesis difficulty of ATG repeats and the possibility of product from translational start at the last methionine).

a

| | UniProt ID | Length (AA) | MW (kD) |
|---|---|---|---|
| fungi | H1VU60 | 262 | 28.50 |
| lancelet | C3XUE4 | 68 | 8.11 |
| virus | R4TWJ3 | 132 | 15.69 |
| algae | A0A087SKU7 | 289 | 30.98 |
| potato | M1DE37 | 85 | 9.79 |
| human | L8EBH1 | 54 | 5.54 |
| archae | M7T2B3 | 124 | 14.58 |
| mouse | E0CYV8 | 86 | 9.40 |
| cyano | B0C4S6 | 83 | 9.63 |
| gut | R9MLR9 | 110 | 12.89 |

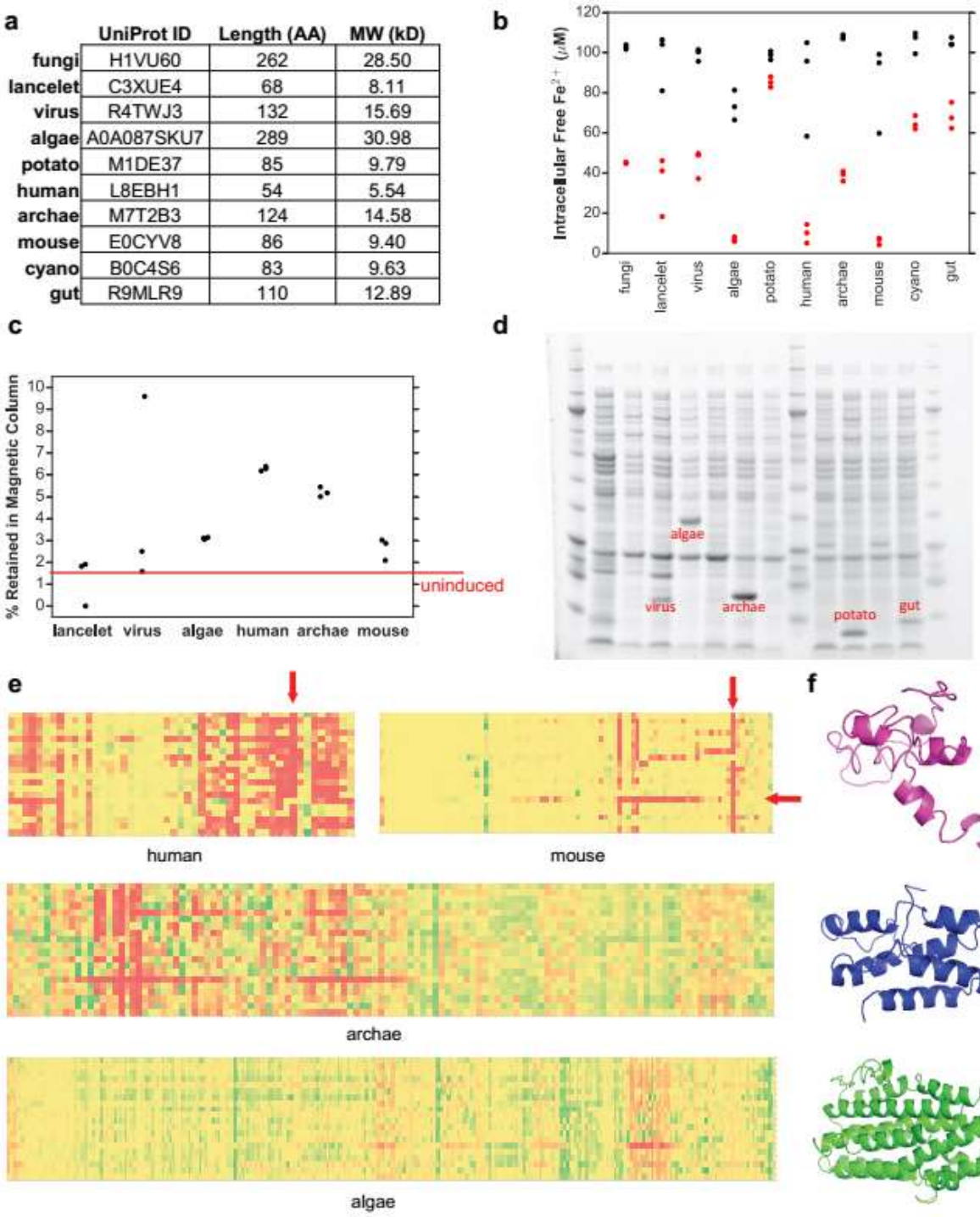


**Figure 4 Experimental validation of predicted iron sequestering proteins (a)** List of ten proteins picked from diverse biological contexts without annotation in UniProt predicted by RNN model to contain “Ferritin” like function **(b)** after cloning and expressing the proteins in *E. coli* with genetic iron sensor, the majority of the tested proteins demonstrated decreased cellular iron particularly for “algae”, “human”, “archaea” and “mouse”. **(c)** Several candidates also gave rise to increased cellular magnetism (magnetic column retention) due to possible iron biomineralization compared to uninduced cells **(d)** Bands for over-expressed proteins could be clearly observed for “virus”, “algae”, “archaea” and “potato” (did not demonstrate significant iron sequestration or magnetism) **(e)** *in silico* “saturation mutagenesis” of selected sequences using RNN model to predict effects of mutations on desired function (red=bad, yellow=neural, green=good), with residue position along horizontal axis and the 20 canonical amino acids along vertical axis. RNN model identifies key positions conserved for function (e.g. vertical arrow), and also the potentially structure-breaking mutations by mutation to proline (horizontal arrow) **(f)** structural homology models of protein candidates “mouse” (*top*), “archaea” (*middle*) and “algae” (*bottom*) using I-TASSER server (the top method in the recent CASP 2012, 2014 protein structure prediction competitions), showing diverse predicted structures.

The *E. coli* cells simultaneously contain a fluorescent, genetic iron sensor based on the *E. coli* fiu promoter that has been validated to detect intracellular iron depletion (Chapter 3). Using calibration by iron chelator bipyridine, the fluorescence values could be converted to equivalent intracellular free iron concentrations. After induction of recombinant protein expression during exponential growth phase followed by overnight growth to saturation in LB media supplemented with 100μM Fe (II) sulfate, the cells were characterized for their fluorescence by the green fluorescent protein (GFP) reporter. All ten proteins showed statistically significant increases in fluorescence, or equivalently decreases in cellular free iron concentrations upon protein expression relative to no expression/induction ($p < 0.05$ by two-tailed Student's t-test) (Figure 4b). However, the protein derived from "potato" did not dramatically change the concentrations compared to the others. To determine the proteins' ability to not only bind and sequester iron but also to bio-mineralize similar to the ferritins and dps proteins, I measured the retention level of the protein-expressing cells in high-gradient magnetic separation columns, as iron based minerals could increase magnetic moment of the cells. Some of the proteins tested, particularly "algae", "human" and "archaea", demonstrated increased magnetic retention compared to the uninduced control (Figure 4c). The expression of some of these proteins including "algae", "archaea", "virus", and the non-sequestering "potato" were clearly observed by SDS-PAGE gel (Figure 4d). The inability to observe bands for candidates "human" and "mouse" may be due to their very low molecular weight (predicted <10kD). Furthermore, the impact of mutations to the predicted sequences on the desired iron-sequestering function could be analyzed using the same trained RNN model *in silico* in the manner of "saturation mutagenesis" where residue position of a sequence is mutated to every other base.

The resulting impacts are illustrated in heat-maps with the residue positions along the sequence along the horizontal axis and the 20 canonical amino acids along the vertical axis. The negative impacts are illustrated as red and positive impacts as green. In this manner, residues “conserved” for function are easily identified by the “red columns” (Figure 4e). Furthermore, a “red row” at proline illustrates the potential helix-breaking and structure-disrupting property of proline, a chemical property that the RNN model has learned only from sequence information without a priori chemical knowledge. Further experimental testing of such mutations could enable further validation and optimization of the RNN model. Lastly, homology modeling of some of the predicted candidates using I-TASSER, the top structure prediction method in the CASP competition in 2012 and 2014, reveals diverse structures. Therefore, the model is not predictive of a particular protein fold or structure but other sequence-based features associated with function.

For machine-learning benchmark, the performance of the RNN model was compared against other popular machine learning classification models, particularly logistic regression and random forest which are known for speed, robustness and often good predictability. Furthermore, both algorithms are capable of modelling nonlinear relationships as would be expected between protein sequences and functions that would not be accurately captured by other fast machine learning methods such as linear regression. For all of these models, a set of features or independent variables are required. Using the same dataset for each of the four functional classes, 51 ProtParam features (Table S1) were extracted or calculated for each sequence and vectorized. These features include simple amino acid composition and length as well as biochemically relevant properties such as isoelectric point, molecular weight, stability index, hydrophobicity and grand average

of hydropathicity (gravy). The logistic regression and random forest models were each trained using “grid-search” over a range of values for their model hyper-parameters, such as alpha for logistic regression, and the parameter values that produced the best prediction results were selected. Comparing the “in-class” prediction performance on the four functional classes by all the machine learning methods, logistic regression was by far the fastest to train but also the least predictive (Figure 5). While random forest was slower, it achieved much better performance but still outclassed by the near perfect performance of the RNN model on the same dataset. Nonetheless, the “feature importance” of random forest models calculated for the four predictors on the 51 features reveals different biases toward different functional classes (Figure 5). The RNN model could not be simply interpreted based on these predefined features, but their best-in-class performance without “feature engineering”, like in other successful deep-learning applications, demonstrate their potential to represent and capture nontrivial and difficult-to-quantify patterns or relationship between sequence information and protein function.

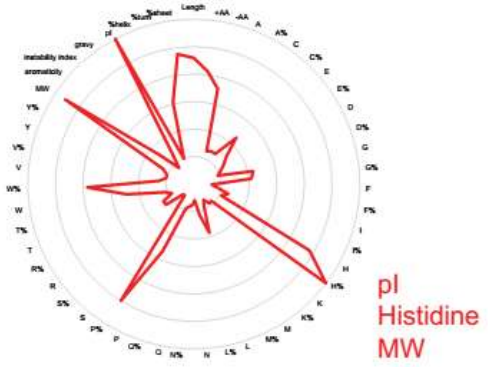
Lastly “out-of-class” prediction performance was tested, whereby the RNN models were trained on sequences from certain protein families and tested on other functionally homologous but phylogenetically distinct families. One drawback of the random splitting of UniProt dataset into train and test sets employed so far is that the two sets could contain highly similar or even identical sequences that represent homologous proteins from closely related species. Furthermore, the ability to discover proteins with homologous function that are distant in evolution from what are already known could be valuable both for studying sequence evolution as well as mining for novel proteins for particular applications like genome-editing. Here I conducted “out-of-class” prediction test on three functions, “Genome-Edit”, “Ferritin”,

and “P450”. The negative set for both training and testing was again the reviewed SwissProt database excluding members containing function of interest. For the “Genome-Edit” function, RNN was trained on the InterPro Cas9 family of proteins (IPR028629, 1201 sequences) as positive set and tested on the InterPro Cpf1 family of proteins (IPR027620, 55 sequences). Both Cas9 and Cpf1 are guided nucleases associated with the CRISPR locus. Cpf1 was discovered more recently and confer benefits such as not requiring a “tracr-DNA” for targeting and potentially higher specificity. Due to the scarcity of the positive training set (Cas9 family) relative to the set of negatives (>550,000 in SwissProt outside of Cas9 and Cpf1 family), the negative set was divided into 100 chunks and sequentially trained with the same positive set (Cas9 family). Such class-balancing or under-sampling during training was not applied during testing on the Cpf1 to more closely simulate the naturally small fraction of positives in a database. For the “Ferritin” function, RNN was trained on the InterPro non-haem ferritin family (IPR001519) along with *either* the haem-containing bacterioferritin family bfr (IPR002024) *or* the DNA-binding protein dps family (IPR002177) as positives, and tested on the remaining un-trained family. The dps differs from the ferritins or bacterioferritins prominently in assembling a cage of 12 rather than 24 monomers. The “P450” function is represented by 6 different sequence clusters/families in InterPro: B-class (IPR002397), E-class-CYP24A-mitochondrial (IPR002949), E-class-group-I (IPR002401), E-class-group-II (IPR002402), E-class-group-IV (IPR002403) and mitochondrial (IPR002399). Either the B-class (31205 sequences) or E-class-group-II (2314 sequences) was treated as the test-set, with training of RNN using the combination of the other families as positives. Taking into account the different length distributions of the protein families, the maximum recurrent depth (i.e. sequence-length) was capped at 333 for “Ferritin”,

500 for P450, and 800 for “Genome-Edit”. To remove possible false positives in the training sets, sequences shorter than 10 amino acids or longer than 1000 amino acids for “Ferritin”, “P450” functions, or 2000 amino acids for “Genome-Edit”, were filtered out before training. As the “Genome-Edit” Cas9 or Cpf1 enzyme sequences are typically over 1000 amino acids long, the RNN was trained scanning over up to the first 800 amino acids from the N-terminus and subsequently from the C-terminus. Overall, prediction performance varied more substantially among the out-of-class predictors compared to the previous random, in-class prediction performance (Table 1). Decent detection sensitivities were achieved with the left-out P450 families and for detecting bfr after training on non-haem ferritins and dps. However, sensitivity/recall was low (0.13) for detection of the 12-member caged dps from RNNs trained only on the 24-member caged non-haem ferritins and bfr. Tripling the number of recurrent layers by feeding the output sequence of one layer as input into the next, which produced a slower but deeper model with potential to encode more complex sequence patterns, increased sensitivity for detecting dps from 0.13 to 0.36 without decreasing precision. Lastly, prediction performance on Cpf1 from an RNN trained on Cas9 yielded sensitivity/recall of 0.59 after training on both N and C terminal residues (up to 800 amino acids) and averaging the prediction probabilities of processing the sequence from its two termini for final classification. Interestingly, classifying using predicted probabilities for only the N- or C-terminal residues (up to 800) significantly decreased precision (i.e. increased false positives), suggesting that multiple features along the entire sequence length (e.g. the binding and nuclease domains) may be required toward accomplishing the “Genome-Edit” function and that many other proteins may exist with only a subset of those features.

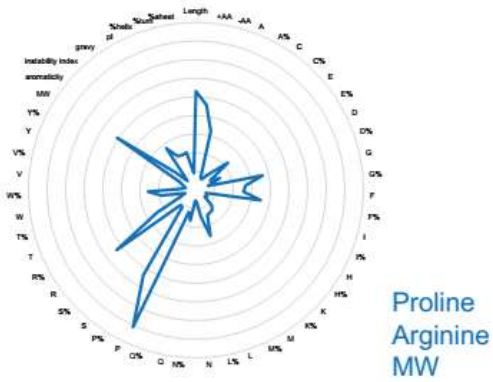
a

| Ferritin | Accuracy | Precision | Recall | F1 |
|---|---|---|---|---|
| RNN | 0.999 | 0.99 | 0.99 | 0.99 |
| LR | 0.960 | 0.69 | 0.60 | 0.64 |
| RF | 0.995 | 0.99 | 0.87 | 0.93 |



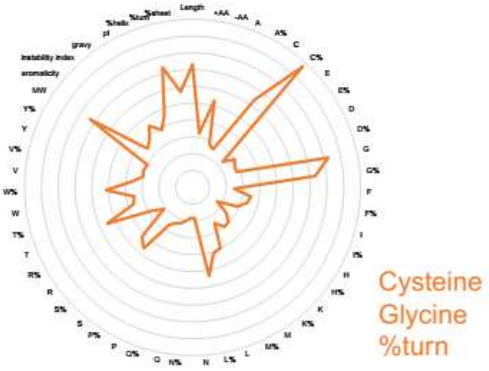

b

| P450 | Accuracy | Precision | Recall | F1 |
|---|---|---|---|---|
| RNN | 0.999 | 1.00 | 0.99 | 1.00 |
| LR | 0.909 | 0.76 | 0.65 | 0.70 |
| RF | 0.979 | 0.98 | 0.90 | 0.93 |




c

| Protease | Accuracy | Precision | Recall | F1 |
|---|---|---|---|---|
| RNN | 0.997 | 0.99 | 0.99 | 0.99 |
| LR | 0.850 | 0.68 | 0.33 | 0.45 |
| RF | 0.975 | 0.98 | 0.89 | 0.94 |




d

| GPCR | Accuracy | Precision | Recall | F1 |
|---|---|---|---|---|
| RNN | 0.994 | 0.98 | 0.99 | 0.99 |
| LR | 0.956 | 0.90 | 0.90 | 0.90 |
| RF | 0.984 | 0.98 | 0.95 | 0.96 |

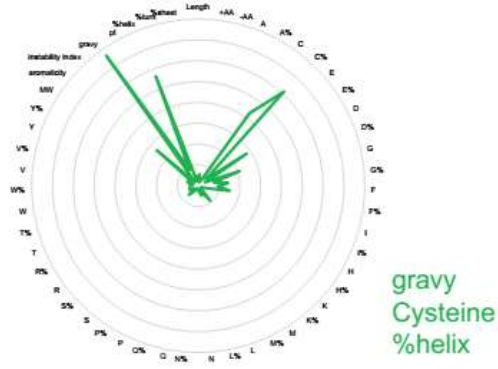



**Figure 5 Performance benchmark with other machine-learning classifiers** For each of the functions “Ferritin” **(a)**, “P450” **(b)**, “Protease” **(c)** and “GPCR” **(d)**, separate logistic regression (LR) or random forest (RF) models were trained on the same input sequence set with 51 sequence-derived ProtParam features optimized by grid-search of hyper-parameters and 5-fold cross-validation and used for prediction on 20% of randomly left-out unseen dataset. The RNN model outperforms in accuracy, precision, recall and F1. To assist understanding the learning of the models, “Feature importances” of the RF models, which achieved relatively high performance, are shown in radial-plots. The three most important features for RF predictions are listed for each function (gravy: grand average of hydropathicity).

| | precision | recall | F1 |
|---|---|---|---|
| **Ferritin-bfr** | 0.98 | 0.59 | 0.74 |
| **Ferritin-dps** | 0.96 | 0.13 | 0.22 |
| **Ferritin-dps_3X** | 0.99 | 0.36 | 0.52 |
| **P450-B** | 0.93 | 0.81 | 0.87 |
| **P450-E_II** | 0.61 | 0.91 | 0.73 |
| **CRISPR-Cpf1_Nterm** | 0.01 | 0.86 | 0.03 |
| **CRISPR-Cpf1_Cterm** | 0.09 | 0.73 | 0.16 |
| **CRISPR-Cpf1_Average** | 0.73 | 0.59 | 0.65 |

**Table 1 Out-of-class RNN classification performance** RNN models were trained toward the Ferritin, P450, Genome-Edit (CRISPR) functions using InterPro families/clusters of protein sequences. For Ferritin function, the bfr or dps family was left out as “test-set”. Tripling the number of recurrent layers for a deeper model (Ferritin-dps_3X) increased recall for predicting dps. For P450, the B class or E class group II (E_II) was left out as “test-set”. For CRISPR, the Cpf1 family was left out as “test-set”. The average of the predictions on up to 800 amino acids in the N and C termini significantly increased precision and F1, suggesting several important features throughout the entire sequence that are necessary for function.

## Discussions

In summary, this study has shown that recurrent neural network (RNN) based on LSTM can be trained to classify certain protein functions with high level of accuracy from input amino acid sequences alone. Experimental validation of the predicted iron sequestering or mineralizing proteins including some currently not easily identified by other bioinformatics methods confirm the accuracy and utility of the model for prediction.

Compared against popular sequence prediction and analysis tools such as BLAST and HMMER, the RNN model currently has several potential benefits but also limitations. One important benefit is the potential to capture obscure sequence-function relationships, allowing predictions of very remote homologies. Unlike most sequence search tools, RNN models do not explicitly rely on sequence alignments or heuristic scoring functions or similarity measures. The memory or internal state of the LSTM neuron processing entire protein sequences, unlike other machine learning methods that employ short, pre-defined motif windows[7,9,20], allows selective retention of important sequence features across long distances[18]. For instance, residues that make up an active site of an enzyme may be separated by large gaps in the protein sequence, but are in proximity of each other in 3-dimensional space. Despite much advances in recent years, the folded structure of proteins still cannot be reliably predicted from their primary amino acid sequences, which limits the prediction of protein function most often highly related to the structure. In this work, four important functional classes were selected which includes as their members proteins across domains of life that share little homology, or have converged upon the same function without common evolutionary origin as in the catalytic triad of the proteases. The ability of the RNN model to accurately make predictions for all of

these functional class from only primary sequence without structural information suggests that the RNN could represent complex patterns in the protein sequence that encode for function. However, it is important to note that the “in-class” performance measures obtained from testing on randomly selected sequences from a small and predominantly reviewed dataset (fewer than 1 million sequences) may not hold for testing on arbitrary databases (e.g. UniRef100 with 54 million sequences). In the larger databases, the proportion of members with particular function (i.e. the positive class) can be extremely small. As a result, very high performance is demanded, with false positive rates approaching zero (i.e. precision very closely to one) to avoid large number of false positive predictions. The “in-class” performance, though respectable, may not be sufficient for the large databases and will require calibration on the same test databases for comparisons against current state-of-art (e.g. BLAST). Furthermore, the “in-class” predictors’ performance may partially benefit from the high similarity or even redundancy of sequences representing homologous proteins in closely related species randomly partitioned into the training and testing sets. The “out-of-class” predictors tested on phylogenetically distinct families showed lower performance as expected. Therefore, while the RNN models can achieve some sensitivity toward new protein families with functional homology, further optimizations are necessary to improve their sensitivity and selectivity particularly for this difficult task of discovering new protein families with related functions in the large and growing sequence databases.

As a “deep-learning” model, RNN with LSTM has found success in several domains related to sequence learning, particularly language recognition and modelling, that surpassed the performance of other machine learning models particularly for learning directly from raw data[15]. However, a current limitation of

using RNN with particularly deep layers (e.g. long sequences) is the training and processing speed. This is mainly because of the large number of variables in a deep neural network model which requires training with large datasets and many operations on large matrices in the iterative optimization steps using the relatively slow gradient based, backpropagation techniques. Building Position Specific Scoring Matrices (PSSM) for PSI-BLAST or hidden Markov models for HMMER as well as searching against those models can be performed faster currently on the public servers.

Besides currently limited computing power, another limitation at the present is the data itself. While there is abundant data for accurate training for the functions of iron-mineralizing proteins, cytochrome P450s, proteases and GPCRs, there are some functions of interest that at the present do not yet have sufficient data size to produce highly predictive models. For example, in the last few years there has been exploding amount of interest and applications of oligonucleotide-targeted nucleases for genome-editing across a variety of systems. The CRISPR (Clustered Regularly Interspaced Short Palindromic Repeats) system originated from bacteria *Streptococcus pyogenes* has been particularly successful in efficient genome editing across a variety of cell types including human cell lines[23,24]. And in recent years new systems of similar function are continuously discovered via bioinformatics techniques for remote-homology prediction such as PSI-BLAST. It is of great interest to discover the whole diversity of oligonucleotide-targeted nucleases for future enhancement of genome editing applications. While there may already be some that are homologous to the known CRISPR systems by sequence or structure, there are potentially more in Nature with more remote homology not detectable by the PSI-BLAST or HMMER. The deep learning approach here employing RNN has the potential to detect those

remote candidates. However, the main challenge currently is the limited amount of public data for creating the training set, as fewer than 10,000 CRISPR/Cas9 like nucleases have been identified. Additionally, unlike the iron-mineralizing ferritins or P450s, these guided nucleases so far identified are mostly large proteins with relatively long sequences of more than 1000 amino acids. Long sequences have been particularly challenging for RNN training due to the exploding or vanishing gradient issue with back-propagation. The use of LSTM neurons allowing selective retention and forgetting of information has ameliorated the issue, but training very long sequences would require significantly more computational processing power and memory. Given significantly more computational resources and time, results here have shown that deeper RNN models could be trained on the currently available dataset to make reasonable predictions (Table 1). However, both sensitivity and selectivity could be optimized with training on the growing volume of experimental data in order to more accurately and precisely discover new functional candidates or protein families and demonstrate utility and power of the RNN predictors over the current state-of-art (e.g. PSI-BLAST).

Despite current limitations in speed and data availability for certain functional prediction applications, RNN-based deep-learning models have the potential to overcome these obstacles quickly in the coming years to become more widely applicable enabled by three trends. On the speed side, both the cost and performance of computing are improving rapidly, particularly due to the design and deployment of highly-parallelized processing architectures (e.g. graphic computing units) that are particularly well suited and have been increasingly dedicated toward training deep neural networks. On the data side, increasingly large volumes of data are collected from automated, high-throughput experimentation. In the field of

synthetic biology, first the cost of sequencing and now of synthesis of DNA has been decreasing dramatically. Large throughput sequencing, particularly of hard-to-culture environmental samples in metagenomics, has rapidly increased the database of sequences available for mining new proteins and new functions. Meanwhile, the accessibility of DNA synthesis has made it possible to quickly test new sequences of interest in relevant biological contexts and obtain valuable data such as those related to protein functions. As more validation data become available, the deep-learning model can be further trained to become more powerful at predicting desired functions. Furthermore, as RNN can be a generative model, it can be trained on proteins of a particular functional class with an auto-encoder and use the decoder to write new protein sequences that may possess that function. This is currently done for translating human languages[25,26] due to the abundance of data. It may be foreseeably applied to protein sequences in the future as the amount of data increases, but it will be significantly more challenging here due to requiring the RNN model to learn and remember not only sufficient patterns for classification of certain functions but also everything else that makes a functional protein, as often even few mutations unrelated to a particular function could cause proteins to mis-fold. At the very least, much deeper RNN models (with numerous stacked recurrent layers) and large hidden state vectors that are capable of storing more information, along with ample training dataset for not only particular function but also for other essential aspects such as proper protein folding, will be necessary to accomplish *de novo* protein “writing”. Lastly on the theoretical side, the convergence of artificial neural networks (ANN) research with the field of neuroscience where it first drew its inspiration could lead to potentially better model or computing architectures that improve both the speed and accuracy of the artificially intelligent predictors.

Given the advantages, limitations as well as the future potentials of deep-learning, the RNN-based models for protein function prediction can accomplish the most value when used in concert with other computational and experimental techniques. For a researcher interested in understanding the function of new genes or discovering new molecules toward particular application such as genome editing or drug, the researcher could first employ BLAST, HMMER or other homology search tools to quickly obtain predictions that can be validated with high-throughput experimentation. With sufficient data relating sequences to a function of interest, the researcher could then train an RNN model to high prediction capability and employ it to discover more remote candidates to test and validate as well as to understand the contributions of residues via *in silico "*saturation mutagenesis". Combining these results with automated, large format assays would enable a high-throughput "design-build-test-learn" cycle of synthetic biology that iteratively improves both the quantity and quality of output toward the engineering target. Moreover, for science, in-depth studies and interpretations of the deep neural network models and the patterns they capture could advance our human understanding of the intricate relationship between sequences, structures and the functions for a wide array of proteins in the future.

## Materials and Methods

### Computational Modelling

All computational models were written in Python and processed on the Harvard Odyssey computing cluster at Harvard University using a combination of CPU and GPU computing nodes. The recurrent neural network models were built upon the Google Tensorflow backend. The logistic regression and random forest models were built using the Python scikit-learn packages. HMMER v3.1b1 (jackhmmer tool) was deployed and executed also on the Odyssey cluster. Protein sequence and function data were obtained directly from the UniProt databases (www.uniprot.org)

For each LSTM Neuron in the RNN, its input "i", output "o", gate "g", forget "f", cell state "c" and hidden state "h" values at time "t" are determined by the following equations[12,18]:

$$i_t = \sigma(D(x_t)W_{xi} + h_{t-1}W_{hi} + b_i)$$

$$f_t = \sigma\big(D(x_t)W_{xf} + h_{t-1}W_{hf} + b_f\big)$$

$$g_t = \tanh\big(D(x_tW_{xg}) + h_{t-1}W_{hg} + b_g\big)$$

$$c_t = f_t \circ c_{t-1} + i_t \circ g_t$$

$$o_t = \sigma(D(x_t)W_{xo} + h_{t-1}W_{ho} + b_o)$$

$$h_t = o_t \circ \tanh(c_t)$$

$$\sigma(z) = \frac{1}{1 + \exp(-z)}$$

, where *W* represents weight matrix, *b* represents constant bias, *D* represents dropout (sets value to zero with probability *p*, *p*=0 in this study), ∘ represents element-wise multiplication (Hadamard product), and *tanh* represents hyperbolic tangent function.

For evaluation of machine learning performance, the metrics are computed from the number of True Positives (TP), True Negatives (TN), False Positives (FP) and False Negatives (FN) as follows:

Accuracy = (TP + TN) / (TP + TN + FP + FN)

Precision = TP / (TP + FP)

Recall = TP / (TP + FN)

F1 = 2 x Precision x Recall / (Precision + Recall)

True Positive Rate (Sensitivity) = TP / (TP + FN) = Recall

False Positive Rate = FP / (FP + TN)

The Receiver Operating Characteristic (ROC) is plotted for the True Positive Rate against the False Positive Rate as classification threshold is varied.

The performances reported in the main text and figures consider the proteins containing a particular function, the minority class, as “Positive” whereas the rest are “Negative”.

**BLAST and HMMER search of experimentally validated RNN predictions**

Default setting were used for NCBI BLAST and EMBL-EBI jackhmmer on their web-servers for searching the ten RNN predictions that were experimentally validated for possible functional homologs. Specifically for NCBI BLAST, the NCBI non-redundant protein sequences database was used for blastp. For jackhmmer run on the EMBL-EBI server, the Reference Proteomes was used, and the Cut-Off thresholds were set at default values such that Significance E-values was 0.01 for sequence and 0.03 for hit, while the Report E-values were 1 for both Sequence and Hit. Jackhmmer was iterated until convergence.

**Construction of expression vectors for predicted protein candidates**

Candidate genes for experimental validation were each cloned into a high copy-number plasmid (pUC origin of replication) with rhamnose inducible promoter (*rhaP*$_{BAD}$, with native *E. coli* transcription factors RhaS and RhaR) and kanamycin resistance cassette via Gibson Assembly. The DNA plasmid was verified by Sanger Sequencing (Genewiz) and transformed into *E. coli* BW25113 cells via electroporation. Protein expression was induced in cells by adding rhamnose to cell culture (maximum 0.2%) during log-phase growth (OD600~0.4). DNA sequences of the most relevant genes and constructs can be found in Table S2 in Appendix C.

**Iron level characterization by genetic sensor**

For the genetic iron sensor, the *E. coli* fiu promoter was cloned along with a super-folder GFP (sfGFP) reporter via Gibson Assembly into a low copy (p15A origin), chloramphenicol-resistance plasmid compatible with the ferritin-expressing plasmid. Iron levels were measured for cells containing the protein-expression and iron sensor plasmids by taking the GFP fluorescence of the culture of cells (488nm excitation by laser, 512nm emission) in 96-well plate format using the BioTek NEO plate-reader. For calibration, known concentrations of iron sequesterer bi-pyridine were added to cell cultures. The fluorescence measured were normalized to culture density by dividing by OD600 measured by the same plate-reader. The increase in normalized fluorescence of the cells was plotted against the increase in bipyridine (or consequent decrease in free iron) and modeled to determine the conversion between fluorescence reading and free iron concentration (Chapter 3).

**Magnetic Column Retention characterization**

A high-gradient magnetic column (Miltenyi LD columns) was sandwiched between two neodymium permanent magnets (K&J Magnetics Inc., BX8C4-N52) to create high magnetic field gradients inside the column. The column was first wetted by passage of 2ml of PBS 1X buffer. Then 500µl of cells re-suspended in PBS 1X buffer were added and flowed through by gravity into the elution tube, followed by addition of 3ml of PBS 1X buffer to wash through any unbound cells into the elution tube. Once dry, the column was removed from the magnets, and 3ml of PBS buffer was pushed through the column to extract the magnetically retained cells into a separate retention tube. Measuring OD600 of the elution and retention tubes allow estimation of cell counts and the percentage of total cells retained by the magnetic column.

**SDS gel analysis of protein expression**

*E. coli* cells were re-suspended in SDS Buffer (NuPAGE LDS Buffer) with reducing agent, followed by two cycles of boiling at 95°C for 5 minutes and vigorous vortex to lyse cells and denature proteins. The lysate was centrifuged to pellet cell debris, and the protein suspension was diluted and added to NuPAGE 4-12% Bis-Tris gel with MES buffer. Empty lanes in the gel were filled with equal volume of SDS buffer. After running at 200V for 35 minutes, the gel was removed and stained with Coomassie Orange dye for one hour and subsequently imaged for dye fluorescence on a Typhoon Imager.

## References


1. Finn, R. D., Clements, J. & Eddy, S. R. HMMER web server: Interactive sequence similarity searching. *Nucleic Acids Res.* **39,** 29–37 (2011).

2. J.S., B. & C.E., P. A review of protein function prediction under machine learning perspective. *Recent Pat. Biotechnol.* **7,** 122–141 (2013).

3. Rampasek, L. & Goldenberg, A. TensorFlow: Biology's Gateway to Deep Learning? *Cell Syst.* **2,** 12–14 (2016).

4. Wallach, I., Dzamba, M. & Heifets, A. AtomNet: A Deep Convolutional Neural Network for Bioactivity Prediction in Structure-based Drug Discovery. *arXiv Prepr. arXiv1510.02855* 1–11 (2015).

5. Ma, J., Sheridan, R. P., Liaw, A., Dahl, G. E. & Svetnik, V. Deep neural nets as a method for quantitative structure-activity relationships. *J. Chem. Inf. Model.* **55,** 263–274 (2015).

6. Wang, S., Peng, J., Ma, J. & Xu, J. Protein Secondary Structure Prediction Using Deep Convolutional Neural Fields. *Sci. Rep.* **6,** 18962 (2016).

7. Wu, C., Berry, M., Shivakumar, S. & McLarty, J. Neural Networks for Full-Scale Protein Sequence Classification: Sequence Encoding with Singular Value Decomposition. *Mach. Learn.* **21,** 177–193 (1995).

8. Tiwari, A. K. & Srivastava, R. A survey of computational intelligence techniques in protein function prediction. *Int. J. Proteomics* **2014,** 845479 (2014).

9. Alipanahi, B., Delong, A., Weirauch, M. T. & Frey, B. J. Predicting the sequence specificities of DNA- and RNA-binding proteins by deep learning. *Nat Biotechnol* **33,** 831–838 (2015).

10. Wu, C. H. Artificial neural networks for molecular sequence analysis. *Comput. Chem.* **21,** 237–256 (1997).

11. Hawkins, J. & Bodén, M. The applicability of recurrent neural networks for biological sequence analysis. *IEEE/ACM Trans. Comput. Biol. Bioinform.* **2,** 243–53 (2005).

12. Sønderby, S. K., Sønderby, C. K., Nielsen, H. & Winther, O. Convolutional LSTM networks for subcellular localization of proteins. *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)* **9199,** 68–80 (2015).

13. Asgari, E. & Mofrad, M. R. K. Continuous distributed representation of biological sequences for deep proteomics and genomics. *PLoS One* **10,** 1–15 (2015).

14. Jo, T., Hou, J., Eickholt, J. & Cheng, J. Improving Protein Fold Recognition by Deep Learning Networks. *Sci. Rep.* **5,** 17573 (2015).

15. LeCun, Y. *et al.* Deep learning. *Nature* **521,** 436–444 (2015).

16. Krizhevsky, A., Sutskever, I. & Hinton, G. E. ImageNet Classification with Deep Convolutional Neural Networks. *Adv. Neural Inf. Process. Syst.* 1–9 (2012). doi:http://dx.doi.org/10.1016/j.protcy.2014.09.007

17. Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I. & Salakhutdinov, R. Dropout: prevent NN from overfitting. *J. Mach. Learn. Res.* **15,** 1929–1958 (2014).

18. Greff, K., Srivastava, R. K., Koutnik, J., Steunebrink, B. R. & Schmidhuber, J. LSTM: A Search Space Odyssey. *IEEE Trans. Neural Networks Learn. Syst.* (2016). doi:10.1109/TNNLS.2016.2582924

19. Graves, A. Generating Sequences with Recurrent Neural Networks. *arXiv preprint arXiv:1308.0850* (2013).

20. Hochreiter, S., Heusel, M. & Obermayer, K. Fast model-based protein homology detection without alignment. *Bioinformatics* **23,** 1728–1736 (2007).

21. Kingma, D. & Ba, J. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* (2014).

22. Jutz, G., Van Rijn, P., Santos Miranda, B. & Boker, A. Ferritin: A versatile building block for bionanotechnology. *Chem. Rev.* **115,** 1653–1701 (2015).

23. Jinek, M. *et al.* A programmable dual-RNA-guided DNA endonuclease in adaptive bacterial immunity. *Science* **337,** 816–822 (2012).

24. Cong, L. *et al.* Multiplex Genome Engineering Using CRISPR/Cas Systems. *Science.* **339(6121)**, 819–824 (2013).

25. Dzmitry Bahdana, Bahdanau, D., Cho, K. & Bengio, Y. Neural Machine Translation By Jointly Learning To Align and Translate. *arXiv preprint arXiv:1409.0473* (2014).

26. Wu, Y. *et al.* Google's Neural Machine Translation System: Bridging the Gap between Human and Machine Translation. *arXiv preprint arXiv:1609.08144* (2016).

## CHAPTER 5

# CONCLUSION

In this thesis, I have presented three major projects exploring the boundary of synthetic biology with energy, physics and computer science around the theme of engineering and optimizing biological constructs with the potential to tackle the issues of sustainability and complexity in the longer term. The first project encompassed engineering the specificity of a central protein in microbial fatty acid metabolism to tailor production of biofuels, relevant to energy and sustainability. The second project involved engineering a ubiquitous iron storage protein for enhanced biomineralization and biomagnetism, with potential applications to magnetic-field enabled non-invasive detection and manipulation for complex biological systems. And the last project explored the potential of artificial neural networks to aid the complex challenge of identifying certain protein functions directly from their primary amino acid sequence. Beyond the immediate applications of the tools and findings of these projects, the approaches developed could be applied toward engineering or better understanding many more biological entities or systems.

In presenting the work of the three projects, I have respectively introduced three different approaches for engineering biology: rational design by human intelligence (Chapter 2), directed evolution by artificial selection (Chapter 3), and computational prediction by artificial intelligence (Chapter 4). Even though each project has biased toward one of the approaches, these three strategies are completely complementary and can be applied simultaneously to engineer and optimize any biological entity or system.

Given sufficient knowledge a biological entities or system such as the atomic structure of a protein determined by crystallography, or the detailed network of interaction among a group of molecules in a cellular context, one could make few rational designs via for example site-specific mutagenesis for proteins or knockout or

knock-in of genes in a cell and test their effects in a low throughput manner. However, for challenging problems without sufficient information on the biological construct or network, as is becoming increasingly the case for studies related to complex diseases (e.g. cancer, neuropsychiatric) or novel metabolic pathways involving many poorly characterized genes, low-throughput engineering toward a particular target such as finding a cure or optimizing the production of a new natural product could have limited success due to the need to explore without much knowledge a potentially enormous combinatorial space of parameters during the optimization procedure.

On the other hand, *in silico* approaches such as employing *ab initio* models or simulations (e.g. molecular dynamics), or in particular artificial intelligence based on machine-learning could potentially be a faster and cheaper strategy to perform efficient filtering and experimental guidance. Specifically, the computer algorithms could, in a highly parallel fashion, predict the likely outcomes from the enormous space of possibilities and suggest the most likely conditions that would achieve a desired design target. The *ab initio* models of proteins or networks are beneficial for developing or validating an understanding or intuition of the biological function with their physically relevant variables and parameters. However, they could fail to represent complex systems where certain assumptions or approximations fail, or there may be “hidden variable” that have not been characterized or taken into account. Machine learning approaches then are potentially powerful to address this type of problems where initial understanding of the entity or system is limited mainly because of their ability, particularly with deep-learning artificial neural networks, to represent and model highly complex patterns and relationships and automatically optimizing those representations or parameters by learning directly from data. Hence

machine learning is best suited for problems where data is relatively easy and cheap to collect, such as for sequencing as its cost decreases and certain types of high-throughput assays. Besides gathering as much data as possible for training, careful choices of the model type, architecture and hyper-parameters will need to be considered to find the optimal trade-off between model “bias” and “variance” and minimize prediction errors due to under or over-fitting from overly simple or complex models, respectively.

Lastly, directed evolution can be an efficient experimental optimization strategy for synthetic biology. It is essentially Nature’s engineering strategy applied on a much faster time-scale using short-generation entities such as bacteria or phages. Generally to evolve a certain entity toward a function, that function is either linked directly to survival or to other types of physical separation and distinction (e.g. magnetic selection in Chapter 3). The directed evolution has been successfully applied to optimize a variety of useful biological constructs such as antibodies or fluorescent proteins. It continues to be an essential tool to engineer many new proteins due to the intrinsic challenge in many cases of predicting effect of mutations on protein structure or function, as well as the ease and efficiency of creating large random mutagenic libraries (e.g. via mutagenic PCR or DNA-array synthesis) and assaying certain phenotypes such as cell survival (e.g. on agar plates) or fluorescence (e.g. in high-density microtiter plates).

As the costs of routine experimental procedures such as DNA sequencing and cloning decrease with increasing automation and commercial scale, and as it becomes correspondingly more feasible and efficient to perform high-throughput experiments for a greater variety of engineering targets, the growing volume of validated data could be fed back to both the human researcher for deriving better

intuition and insights, as well as to computational algorithms for developing better *in silico* models for prediction and informing more optimal designs. Hence it is imperative that researchers openly share and publish data, both successes *as well as* failures. Not only are the failed results valuable for machine learning models often as much as the successes, but they could also result in potentially large monetary and time savings for other human researchers. These benefits will inevitably accelerate scientific progress in the larger context.

Finally, by combining the three engineering approaches employing human intelligence, artificial intelligence, and artificial selection, as well as developing and applying ideas across scientific and engineering disciplines, we could accelerate our progress in engineering biology or other synthetic systems toward solving important societal problems such as sustainability and diseases, and from these efforts develop a better understanding of our universe and ourselves.

## APPENDIX A

# SUPPORTING INFORMATION FOR CHAPTER 2

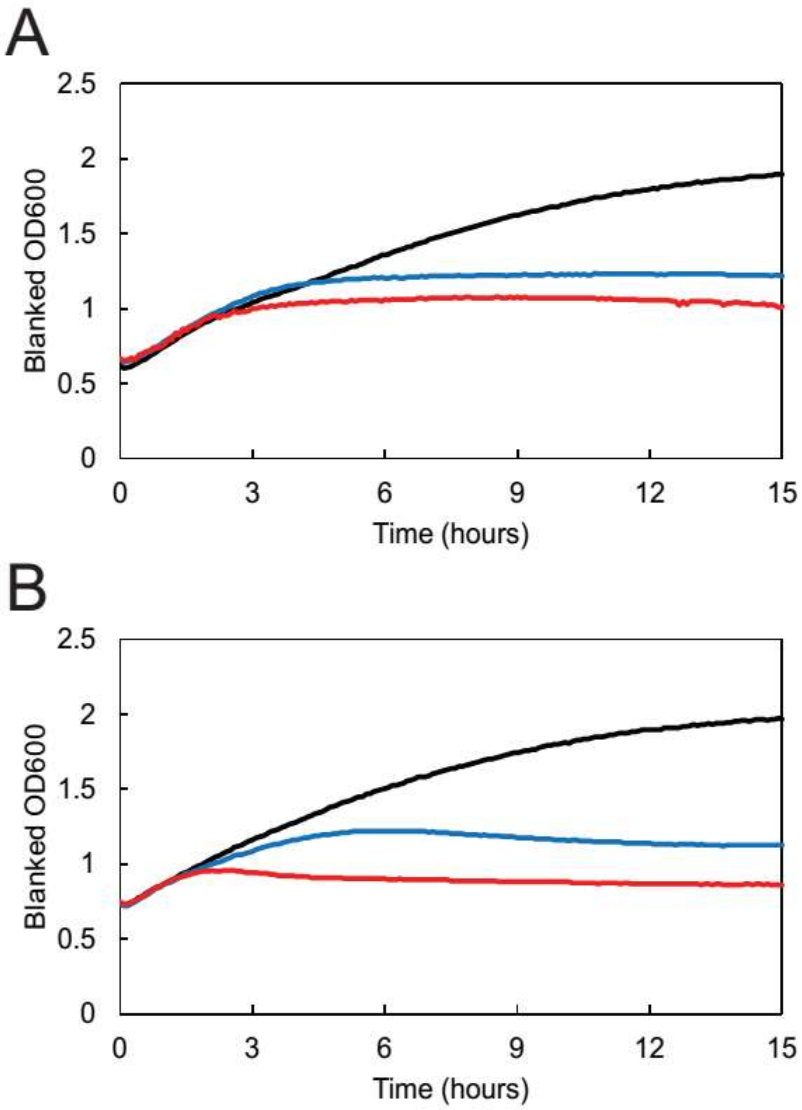



**Figure S1 Growth Suppression by ACP expression.** Growth of *E. coli* is suppressed by induction of Se-ACP expression increasing from 0mM (black), 0.1mM (blue) to saturation at 1mM (red) IPTG. The growth defect is likely due to inhibition of phospholipid metabolism by apo-ACP. Se-ACP I75W expression (**B**) shows similar growth suppression compared to WT (**A**), indicating proper folding and functionality of ACP. All mutant ACPs show similar growth curves (data not shown). Representative growth curves are shown.

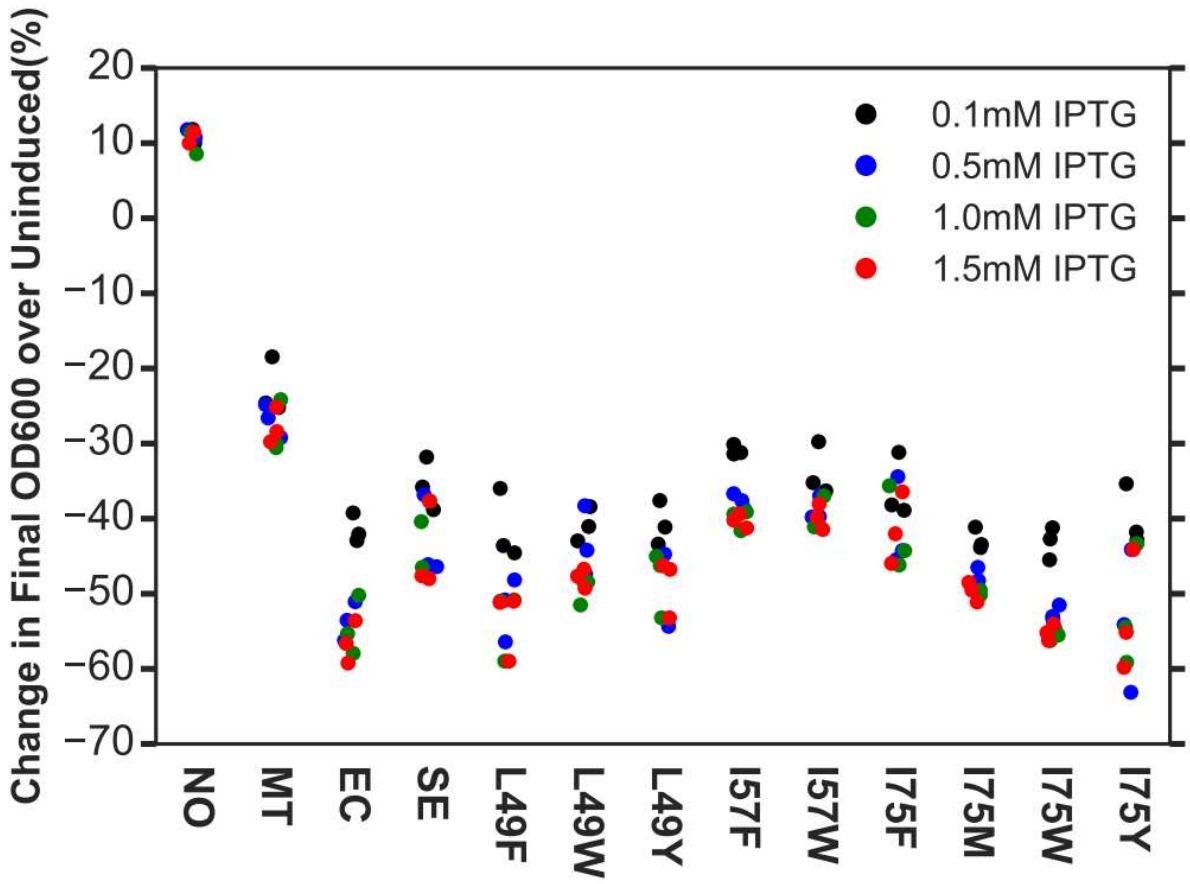



**Figure S2 Growth Suppression by ACP expression.** Shown are changes in culture OD of *E. coli* strains induced to overexpress various ACPs versus their uninduced condition (0mM IPTG). Culture densities were measured after 15 hours of growth in M9 minimal media with 0.4% glucose. When overexpressed, most mutants show equal or stronger growth suppression vs. WT Se-ACP. Data represent triplicate biological measurements.

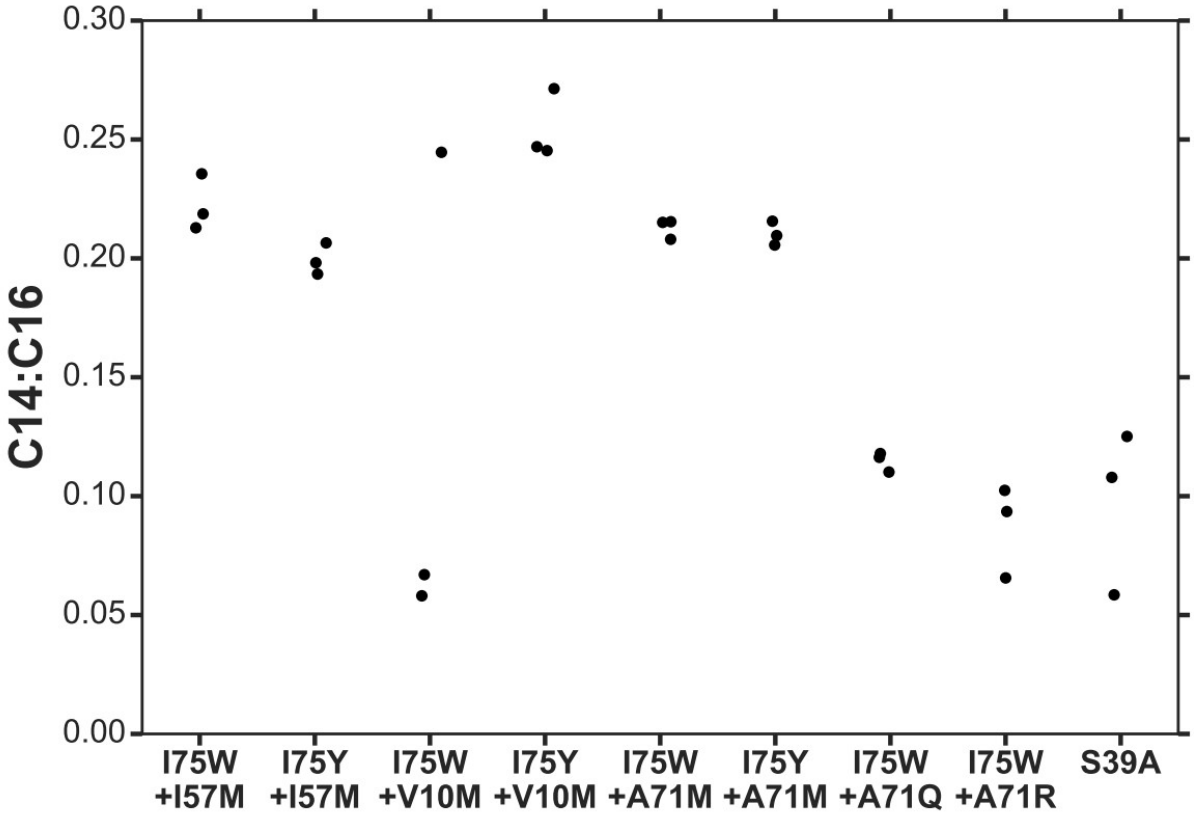



**Figure S3 GC-MS analysis of cellular lipids from double mutants.** Combining the single Se-ACP I75W or I75Y point mutants with a second set of residues mutated to methionine does not significantly change the C14:C16 ratio from that observed for the single point mutants alone. Mutating this second set of residues to arginine (A71R) or glutamine (A71Q) reduced the C14:C16 ratio to WT Se-ACP levels. Data represent triplicate biological measurements.

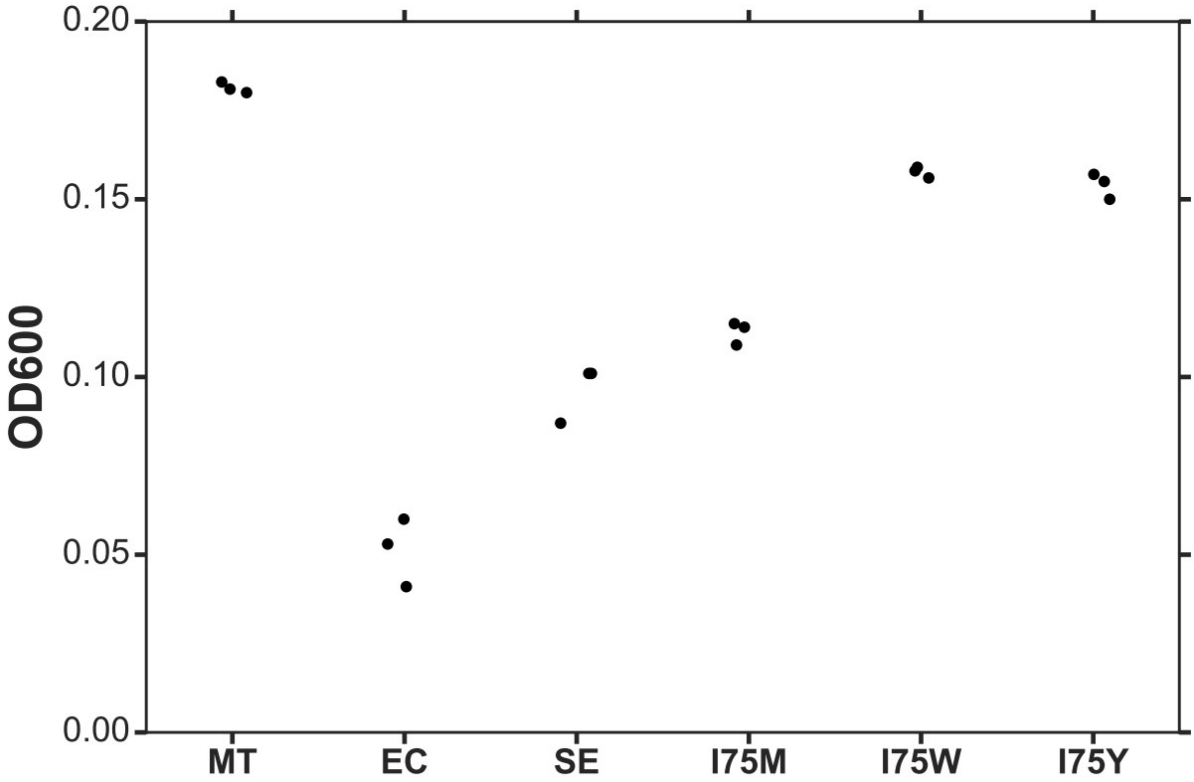



**Figure S4 Culture OD of strains co-expressing recombinant ACP and C12 thioesterase.** OD600 of the cultures were measured in stationary phase using a Biotek NEO plate reader. The mutants that showed largest increase in FFA (I75W, I75Y) measured higher OD600 compared to the wild type controls (EC, SE), indicating that increased medium chain FFA production is not a consequence of decrease growth rate.

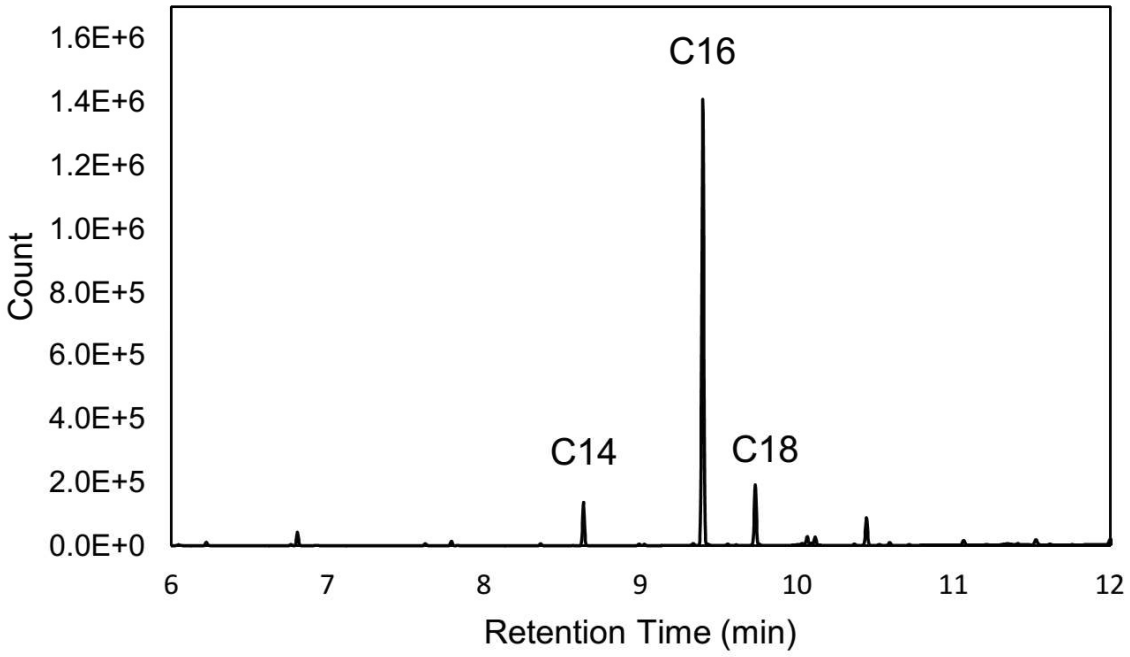



**Figure S5 GC-MS Chromatogram of FAME extracted from *E. coli* culture** FAME peaks were identified by retention time and mass spectrum. C14 and in particular C16 peaks were most dominant in all cultures. C18 peak was much less significant, and fatty acid shorter than C14 were not reliably detected.

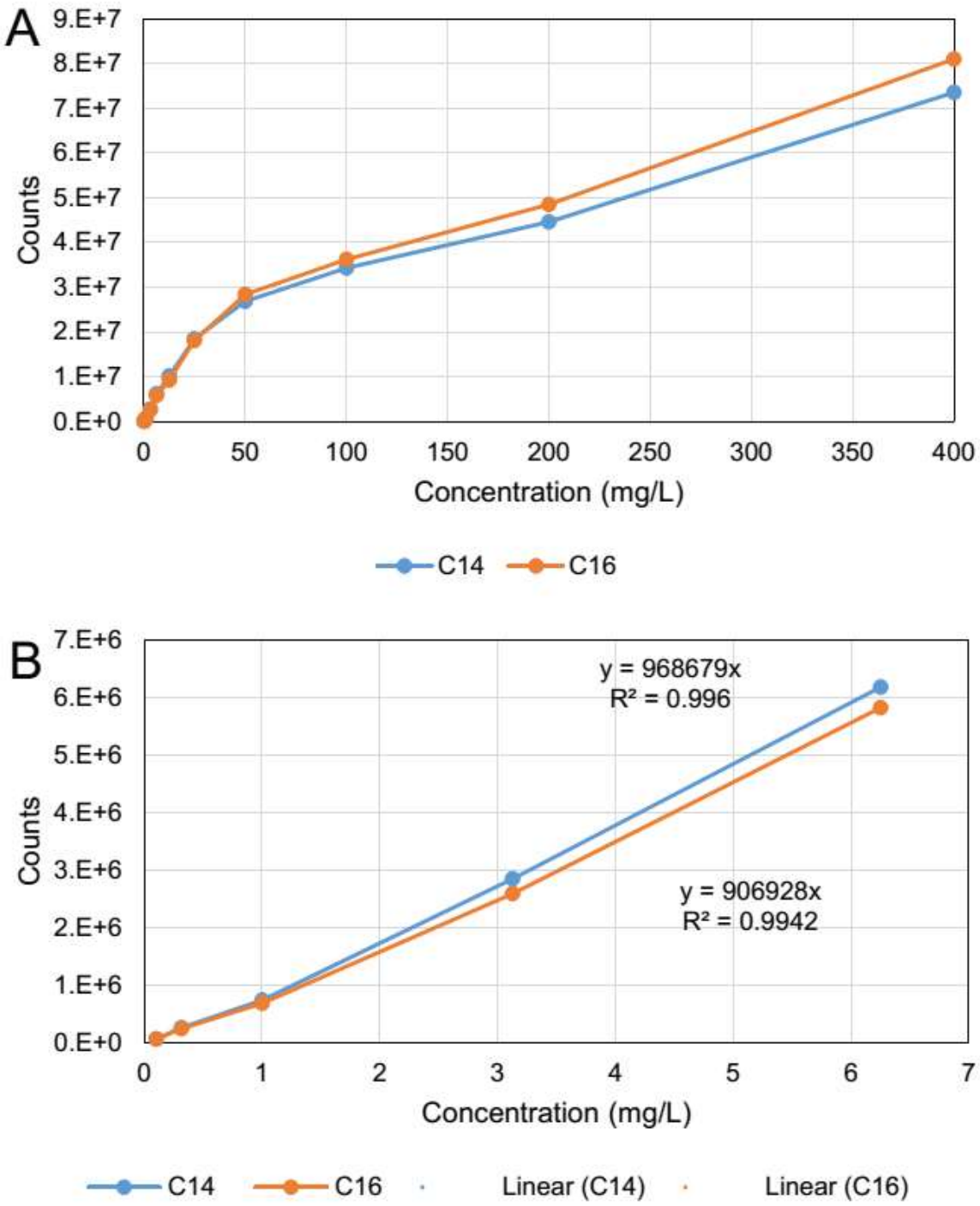



**Figure S6 GC-MS Concentration Calibration Curve for C14 and C16 FAMEs** The C14-C16 peak areas were quantified as a function of their known concentrations over the wide range from 0.1 to 400 mg/L (**A**). Linear fits were extracted from the hexane background-subtracted standard curves in the low concentration range to calibrate the concentrations of C14 and C16 from cell samples, which lie in this lower range (**B**). Eventually the calibrated mass concentrations were converted to molar concentrations by dividing by molecular mass.

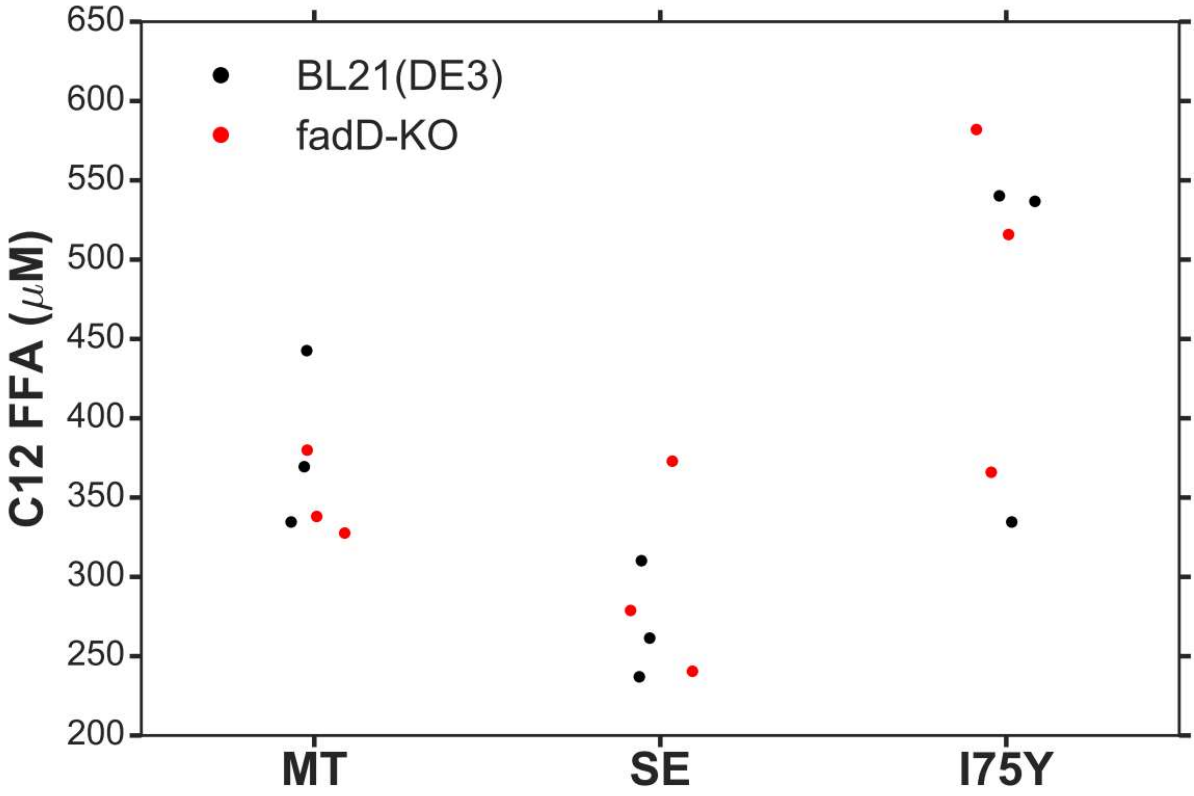



**Figure S7 Effect of Eliminating Beta Oxidation by fadD Knock-Out on FFA** C12 FFA yields were compared between wild-type BL21(DE3) and fadD knocked-Out (KO) BL21(DE3) strains after 24 hours of induced expression of C12 thioesterase (1% arabinose) and ACP (100uM IPTG), where the ACP vector contained either empty vector (MT), wild-type cyano ACP (SE), or I75Y cyano ACP mutant. The fadD-KO strains that eliminated beta oxidation of FFA did not present increased yields of C12 FFA.

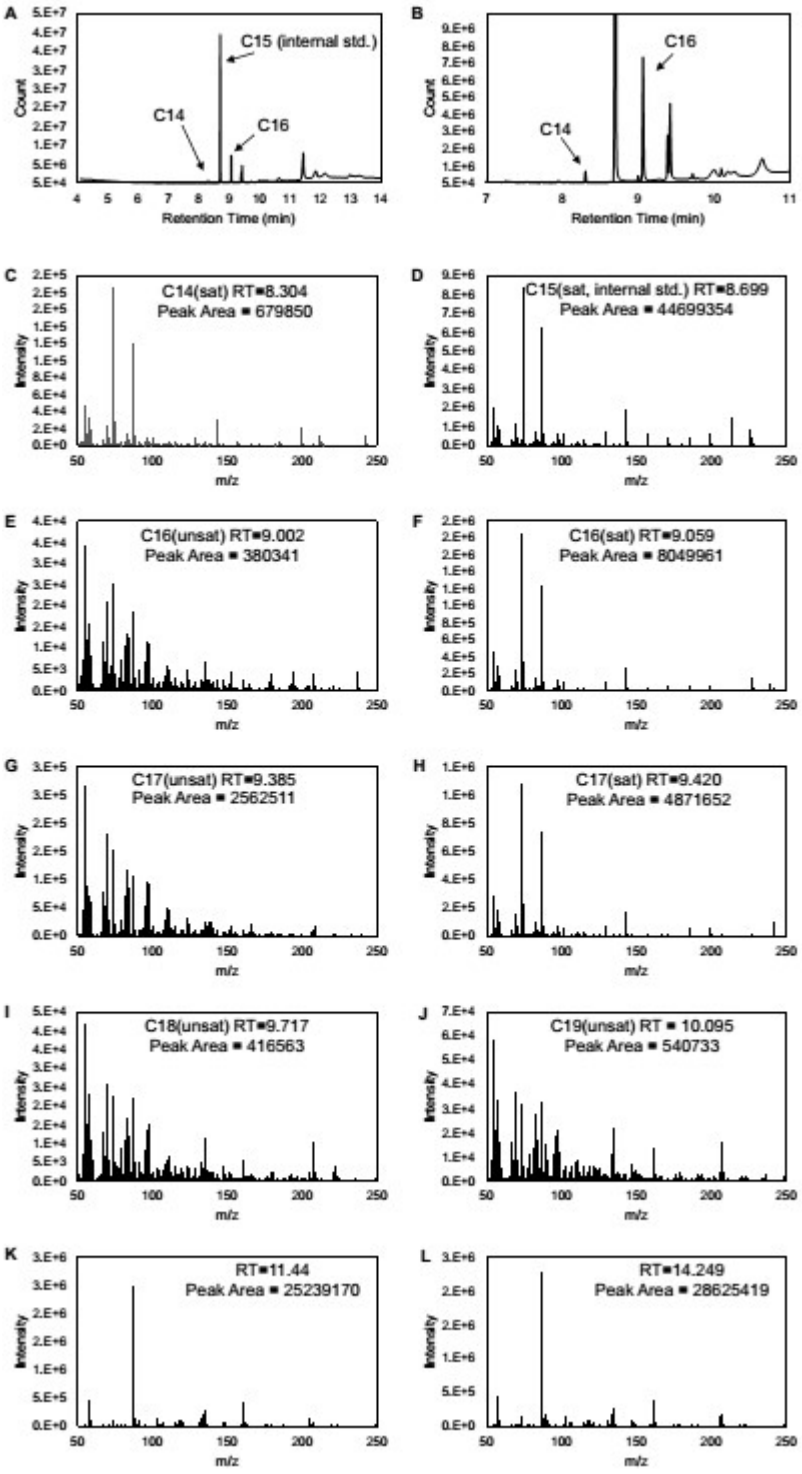



**Figure S8 GC-MS chromatogram and mass spectra of BL21(DE3) expressing WT *E. coli* ACP** Full (A) and zoomed-in (B) chromatogram in mass scan mode highlighting the saturated C14, C16 peaks from the extracted sample and the peak of saturated C15 added to during extraction as internal standard. (C-L) mass spectra of major peaks in the chromatogram, labelled by the peak identity, retention time, and integrated peak area.

| | |
|---|---|
| **EC (ACP)** | ATGAGCACTATCGAAGAACGCGTTAAGAAAATTATCGGCGAAC<br>AGCTGGGCGTTAAGCAGGAAGAAGTTACCAACAATGCTTCTTT<br>CGTTGAAGACCTGGGCGCGGATTCTCTTGACACCGTTGAGCT<br>GGTAATGGCTCTGGAAGAAGAGTTTGATACTGAGATTCCGGA<br>CGAAGAAGCTGAGAAAATCACCACCGTTCAGGCTGCCATTGA<br>TTACATCAACGGCCACCAGGCGTAA |
| **SE (ACP)** | ATGAGCCAGGAAGATATTTTTAGCAAAGTGAAAGATATTGTGG<br>CGGAACAGCTGAGCGTGGATGTGGCGGAAGTGAAACCGGAA<br>AGCAGCTTTCAGAACGATCTGGGCGCGGATAGCCTGGATACC<br>GTGGAACTGGTGATGGCGCTGGAAGAAGCGTTTGATATTGAA<br>ATTCCGGATGAAGCGGCGGAAGGCATTGCGACCGTGCAGGA<br>TGCGGTGGATTTTATTGCGAGCAAAGCGGCGTAATGATGA |
| **S39A** | ATGAGCCAGGAAGATATTTTTAGCAAAGTGAAAGATATTGTGG<br>CGGAACAGCTGAGCGTGGATGTGGCGGAAGTGAAACCGGAA<br>AGCAGCTTTCAGAACGATCTGGGCGCGGATGCGCTGGATACC<br>GTGGAACTGGTGATGGCGCTGGAAGAAGCGTTTGATATTGAA<br>ATTCCGGATGAAGCGGCGGAAGGCATTGCGACCGTGCAGGA<br>TGCGGTGGATTTTATTGCGAGCAAAGCGGCGTAATGATGA |
| **L49F** | ATGAGCCAGGAAGATATTTTTAGCAAAGTGAAAGATATTGTGG<br>CGGAACAGCTGAGCGTGGATGTGGCGGAAGTGAAACCGGAA<br>AGCAGCTTTCAGAACGATCTGGGCGCGGATAGCCTGGATACC<br>GTGGAACTGGTGATGGCGCTGGAAGAAGCGTTTGATATTGAA<br>ATGCCGGATGAAGCGGCGGAAGGCATTGCGACCGTGCAGGA<br>TGCGGTGGATTTTTGGGCGAGCAAAGCGGCGTAATGATGA |
| **L49W** | ATGAGCCAGGAAGATATTTTTAGCAAAGTGAAAGATATTGTGG<br>CGGAACAGCTGAGCGTGGATGTGGCGGAAGTGAAACCGGAA<br>AGCAGCTTTCAGAACGATCTGGGCGCGGATAGCCTGGATACC<br>GTGGAACTGGTGATGGCGCTGGAAGAAGCGTTTGATATTGAA<br>ATGCCGGATGAAGCGGCGGAAGGCATTGCGACCGTGCAGGA<br>TGCGGTGGATTTTTATGCGAGCAAAGCGGCGTAATGATGA |
| **L49Y** | ATGAGCCAGGAAGATATTTTTAGCAAAATGAAAGATATTGTGG<br>CGGAACAGCTGAGCGTGGATGTGGCGGAAGTGAAACCGGAA<br>AGCAGCTTTCAGAACGATCTGGGCGCGGATAGCCTGGATACC<br>GTGGAACTGGTGATGGCGCTGGAAGAAGCGTTTGATATTGAA<br>ATTCCGGATGAAGCGGCGGAAGGCATTGCGACCGTGCAGGA<br>TGCGGTGGATTTTTGGGCGAGCAAAGCGGCGTAATGATGA |
| **I57F** | ATGAGCCAGGAAGATATTTTTAGCAAAATGAAAGATATTGTGG<br>CGGAACAGCTGAGCGTGGATGTGGCGGAAGTGAAACCGGAA<br>AGCAGCTTTCAGAACGATCTGGGCGCGGATAGCCTGGATACC<br>GTGGAACTGGTGATGGCGCTGGAAGAAGCGTTTGATATTGAA<br>ATTCCGGATGAAGCGGCGGAAGGCATTGCGACCGTGCAGGA<br>TGCGGTGGATTTTTATGCGAGCAAAGCGGCGTAATGATGA |
| **I57W** | ATGAGCCAGGAAGATATTTTTAGCAAAGTGAAAGATATTGTGG<br>CGGAACAGCTGAGCGTGGATGTGGCGGAAGTGAAACCGGAA<br>AGCAGCTTTCAGAACGATCTGGGCGCGGATAGCCTGGATACC<br>GTGGAACTGGTGATGGCGCTGGAAGAAGCGTTTGATATTGAA<br>ATTCCGGATGAAGCGGCGGAAGGCATTGCGACCGTGCAGGA<br>TATGGTGGATTTTTGGGCGAGCAAAGCGGCGTAATGATGA |
| **I75F** | ATGAGCCAGGAAGATATTTTTAGCAAAGTGAAAGATATTGTGG<br>CGGAACAGCTGAGCGTGGATGTGGCGGAAGTGAAACCGGAA |

| | |
|---|---|
| | AGCAGCTTTCAGAACGATCTGGGCGCGGATAGCCTGGATACC<br>GTGGAACTGGTGATGGCGCTGGAAGAAGCGTTTGATATTGAA<br>ATTCCGGATGAAGCGGCGGAAGGCATTGCGACCGTGCAGGA<br>TATGGTGGATTTTTATGCGAGCAAAGCGGCGTAATGATGA |
| **I75M** | ATGAGCCAGGAAGATATTTTTAGCAAAGTGAAAGATATTGTGG<br>CGGAACAGCTGAGCGTGGATGTGGCGGAAGTGAAACCGGAA<br>AGCAGCTTTCAGAACGATCTGGGCGCGGATAGCCTGGATACC<br>GTGGAACTGGTGATGGCGCTGGAAGAAGCGTTTGATATTGAA<br>ATTCCGGATGAAGCGGCGGAAGGCATTGCGACCGTGCAGGA<br>TCAGGTGGATTTTTGGGCGAGCAAAGCGGCGTAATGATGA |
| **I75W** | ATGAGCCAGGAAGATATTTTTAGCAAAGTGAAAGATATTGTGG<br>CGGAACAGCTGAGCGTGGATGTGGCGGAAGTGAAACCGGAA<br>AGCAGCTTTCAGAACGATCTGGGCGCGGATAGCCTGGATACC<br>GTGGAACTGGTGATGGCGCTGGAAGAAGCGTTTGATATTGAA<br>ATTCCGGATGAAGCGGCGGAAGGCATTGCGACCGTGCAGGA<br>TCGTGTGGATTTTTGGGCGAGCAAAGCGGCGTAATGATGA |
| **I75Y** | ATGAGCCAGGAAGATATTTTTAGCAAAGTGAAAGATATTGTGG<br>CGGAACAGCTGAGCGTGGATGTGGCGGAAGTGAAACCGGAA<br>AGCAGCTTTCAGAACGATCTGGGCGCGGATAGCCTGGATACC<br>GTGGAACTGGTGATGGCGCTGGAAGAAGCGTTTGATATTGAA<br>ATTCCGGATGAAGCGGCGGAAGGCATTGCGACCGTGCAGGA<br>TGCGGTGGATTTTTATGCGAGCAAAGCGGCGTAATGATGA |
| **C12 thioester-ase** | ATGACGAATCTCGAATGGAAACCTAAACCCAAGCTGCCACAAT<br>TGCTCGATGACCATTTTGGCCTCCATGGCCTCGTCTTTCGCCG<br>GACTTTTGCTATTCGGTCGTACGAAGTGGGTCCCGATCGATC<br>GACCAGCATTTTGGCGGTGATGAACCACATGCAGGAAGCGAC<br>GCTCAATCATGCGAAAAGCGTGGGGATCCTGGGGGATGGTTT<br>TGGCACTACACTGGAAATGTCTAAACGCGATTTGATGTGGGTG<br>GTCCGCCGCACCCATGTTGCTGTGGAGCGGTATCCGACGTG<br>GGGCGATACCGTCGAAGTTGAGTGCTGGATCGGCGCGTCTG<br>GCAATAACGGCATGCGCCGCGATTTCTTGGTTCGTGATTGCA<br>AGACTGGTGAAATTCTCACACGCTGCACGAGCCTCTCGGTGC<br>TGATGAACACCCGCACACGTCGCCTGAGCACTATCCCAGACG<br>AAGTCCGTGGTGAAATTGGTCCTGCATTTATTGACAATGTTGC<br>GGTTAAGGATGATGAGATTAAAAAACTCCAAAAACTGAATGAC<br>TCCACGGCTGACTACATTCAAGGTGGCTTGACGCCGCGATGG<br>AATGATTTGGATGTGAATCAACATGTCAATAATCTGAAGTATGT<br>CGCATGGGTGTTTGAGACGGTTCCGGATAGTATCTTTGAAAG<br>CCATCACATTTCGTCTTTTACCCTCGAATACCGGCGGGAATGC<br>ACCCGTGATAGCGTTTTGCGGTCCCTCACGACTGTTAGCGGA<br>GGTAGTTCCGAAGCGGGCCTCGTGTGCGATCATCTGCTGCAG<br>CTCGAAGGTGGATCGGAAGTGCTCCGTGCGCGAACGGAATG<br>GCGGCCCAAACTGACCGACTCGTTTCGCGGAATCAGTGTCAT<br>TCCAGCTGAACCCCGGGTCTAG |

**Table S1 DNA sequences of relevant genes and constructs expressed and tested in Chapter 2.**

## APPENDIX B

# SUPPORTING INFORMATION FOR CHAPTER 3

Here I describe two experiments and their preliminary results showing the potential of using magnetic ferritin mutants for non-invasive *in vivo* imaging in mouse models and control of ion channels in human tissue culture. These are followed by additional supplementary figures and tables referenced in the main text.

We explored *in vivo* MRI imaging of pre-induced, magnetized *E. coli* orally gavaged to mouse (Figure S6). Both T2-weighted images using spin echo sequence (Figure S6a, c) and T2*-weighted images using gradient echo sequence were taken across a coronal slice of the body of the mouse. The control mouse fed with uninduced *E. coli* shows dark spots in the images (Figure S6a, b) possibly attributable to gas or stool as artefacts. However for the mouse fed with induced *E. coli* over-expressing magnetic ferritin m1, not only was a clear dark spot observed in T2-weighted image (Figure S6c, red arrow), but a localized distortion was observed in the same location in the T2*-weighted image (Figure S6d, red arrow) potentially indicative of local magnetic field produced from a bolus of bacteria. However, more controlled gut preparation procedures to remove gas or stool would be necessary to remove the imaging artefacts present even in the uninduced control and clearly demonstrate the effect of the magnetic bacteria. Furthermore, it would be of interest to demonstrate *in vivo* production of ferritins and hence biomagetism by a stimulus of interest (e.g. tetracycline, or inflammation) for truly noninvasive, *in vivo* biosensing using MRI.

Expression of magnetic ferritin mutants in human cells could further enable magnetic control of cellular functions. GFP-tagged and untagged ferritins were simultaneously expressed in HEK293 freestyle cells via transient transfection to form magnetic ferritin particles in cells (Figure S7a). This dual start codon design avoids steric hindrance from the larger GFP during monomer self-assembly into cages and

was observed to maintain iron sequestration in bacterial cells (data not shown). Simultaneously, human TRPV1 ion channel was fused at its N-terminus to a GFP nanobody and co-transfected. The GFP tagged ferritin cages thus bind to the TRPV1 channels. Under a strong magnetic field gradient supplied from a rare earth permanent magnet placed in close proximity, the engineered ferritins, and not the wildtype or the no-ferritin control, exhibited magnetic actuation of the TRPV1 channels indicated by transient increase in intracellular calcium level measured via X-rhod-1 red-fluorescent dye (Figure S7b). Positive control adding 10uM of capsaicin (TRPV1 agonist) shows a level increase in intracellular calcium that was also attained transiently via magnetic actuation. Positive control with ionomycin, which releases intracellular *and* extracelluar sources of calcium into the cytoplasm, drastically increases fluorescence and confirms functionality of the calcium dye (Figure S7c). Confocal microscopy images of the cells demonstrate co-localization of the GFP-expressing cells (green) and all the cells that have taken up the calcium dye (red) (Figure S7d, e, f). For greater robustness, transgene expression should be chromosomally integrated into the HEK293 cell line for stable, reproducible expression levels. Furthermore, more magnetic protein constructs may be needed for robust actuation at significantly weaker magnetic fields such as that from RF pulses used for noninvasive remote-control of transgene expressing cell *in vivo*.

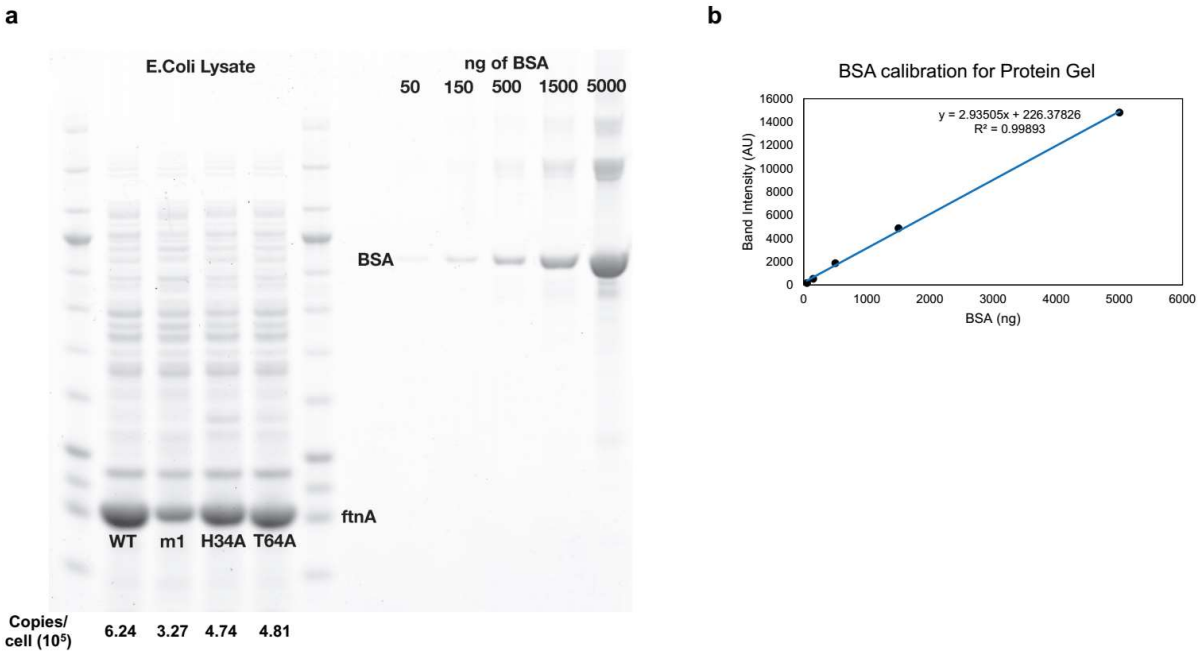



**Figure S1 BSA-calibrated SDS-PAGE gel shows lower ferritin expression level for mutants (a)** SDS-PAGE gel of cell lysate from ferritin over-expressing *E. coli* shows strongest expression for wildtype (WT) compared to the more magnetic ferritin mutants. **(b)** Band intensities versus amount of BSA loaded (in ng) produces good linear fit ($R^2$ = 0.999). The fit parameters were used along with culture OD and assumed $5x10^8$ cells per ml concentration at $OD_{600}$ = 1 to calculate the protein copies per cell in A.

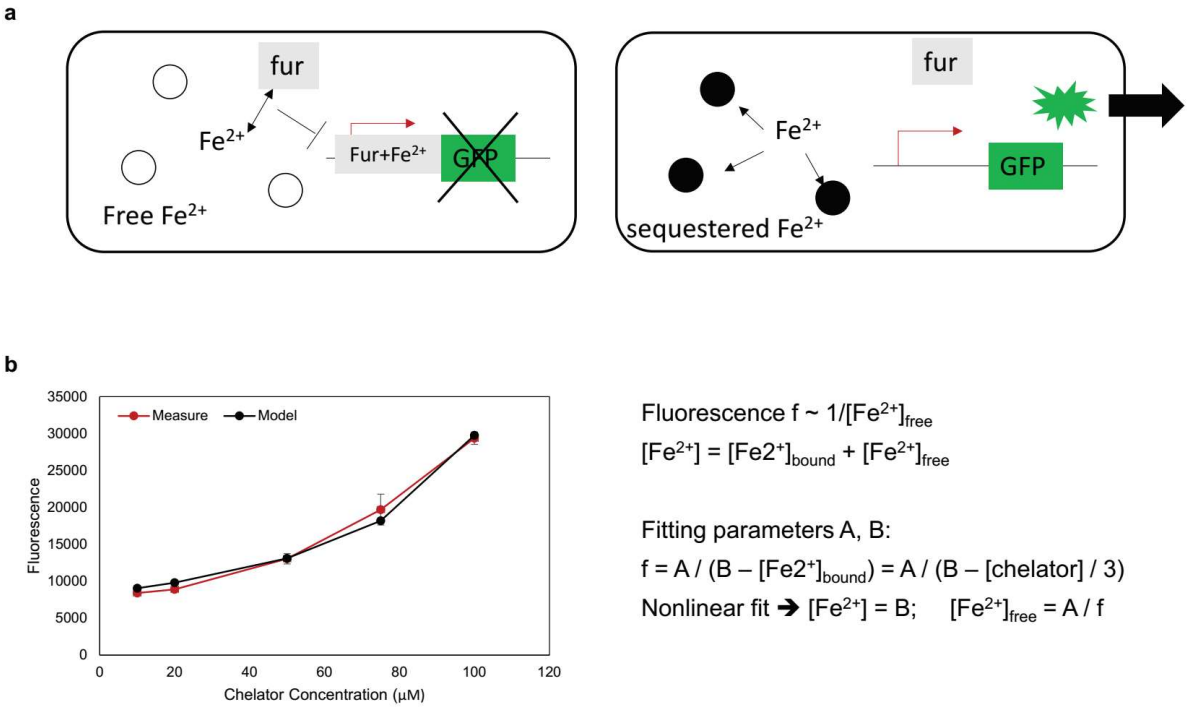



**Figure S2 Novel fluorescent genetic sensor for cytoplasmic free $Fe^{2+}$ (a)** Free $Fe^{2+}$ binds to the ferric uptake regulator (apo-fur). The Fe-bound fur binds the fiu promoter sequence to repress transcription of GFP (right). Sequestration of free $Fe^{2+}$ by ferritins increases sensor fluorescence output. **(b)** The concentration of free $Fe^{2+}$ can be deduced from the fluorescence signal by calibration of sensor output with titration of known concentrations of $Fe^{2+}$ chelator bipyridine. The simple nonlinear model produces good fit to the measured fluorescence.

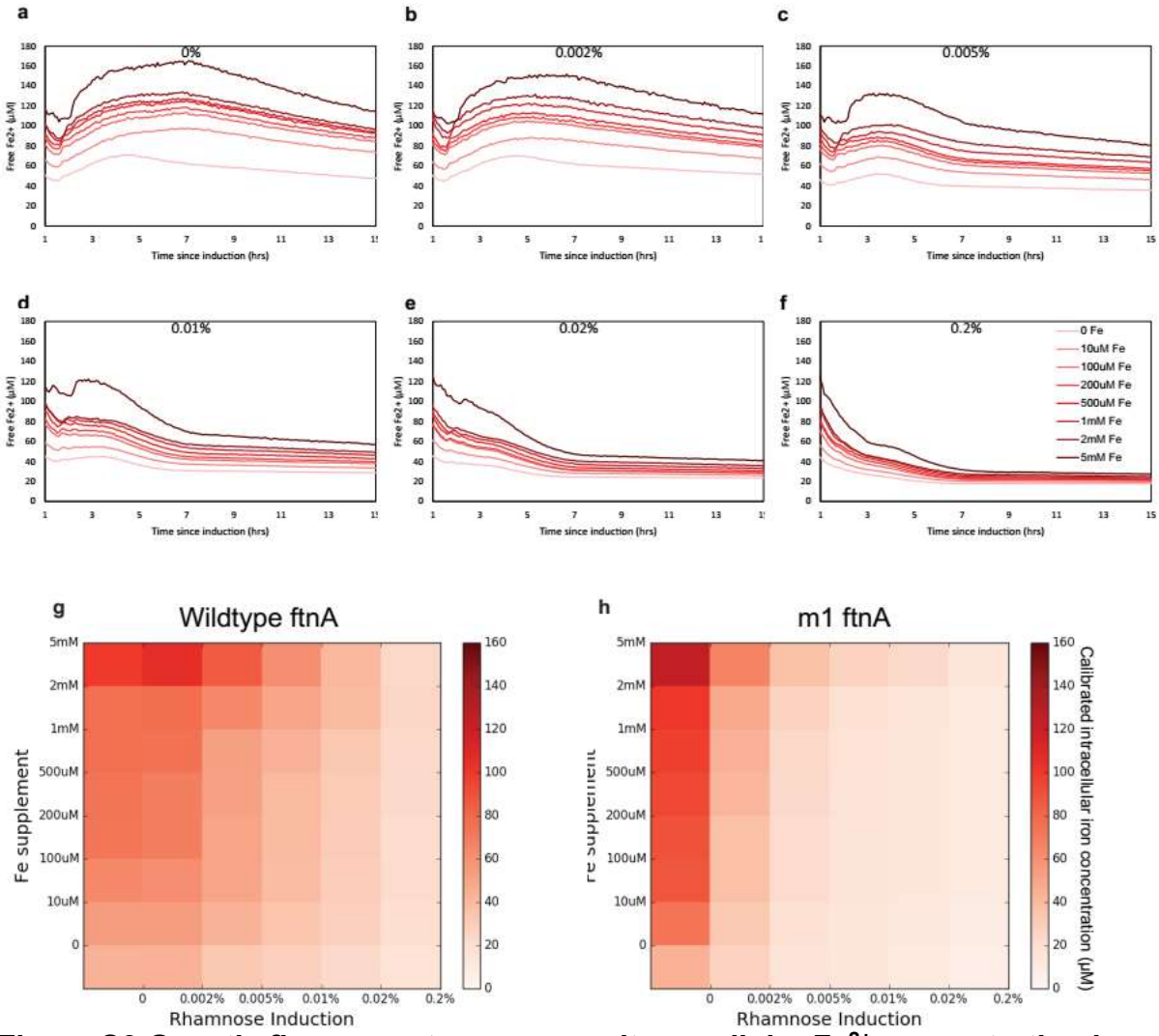



**Figure S3 Genetic fluorescent sensor monitors cellular $Fe^{2+}$ sequestration in WT and mutant ferritin expressing cells (second replicate of data in Figure 3).** (**a-f)** Calibrated free $Fe^{2+}$ concentrations in *E. coli* expressing wildtype ferritin from up to 15 hours after induction by 0% (a), 0.002% (b), 0.005% (c), 0.01% (d), 0.02% (e), and 0.2% (f) rhamnose with media $Fe^{2+}$ supplement of 0 μM, 10 μM, 100 μM, 200 μM, 500 μM, 1 mM, 2 mM, 5 mM. Without ferritins high media supplement at 5mM can dramatically alter the intracellular iron homeostatic setpoint. At high induction levels free $Fe^{2+}$ is efficiently sequestered up to the highest media supplement concentration. The heatmaps with color saturation proportional to calibrated intracellular free iron levels show that compared to the wildtype **(g)**, the best ferritin mutant **(h)** is much more effective at sequestering iron at lower protein levels (0.01% rhamnose induction) and at high environmental iron concentration (up to 2 mM). This is consistent with their greater magnetism despite lower protein expression.

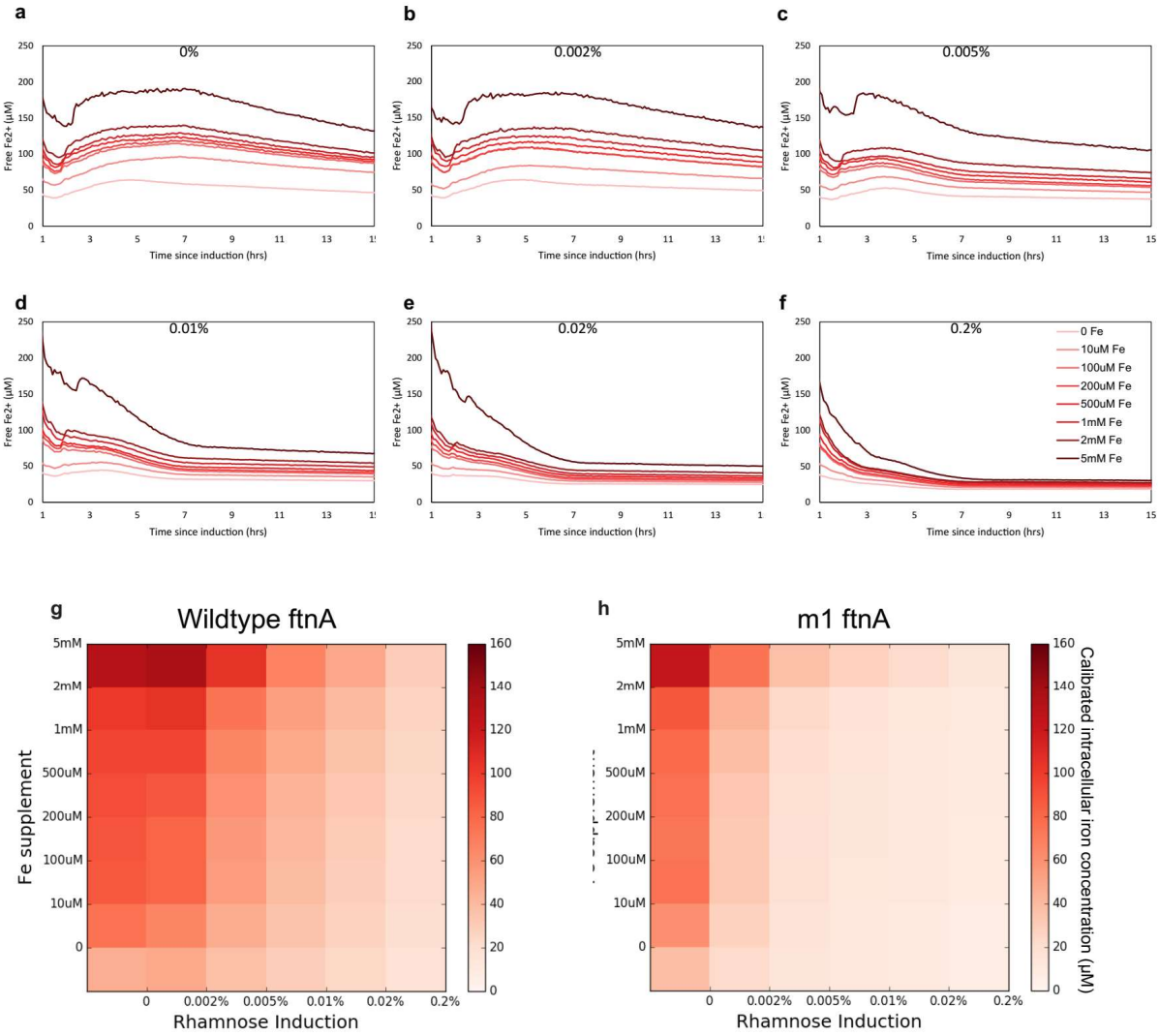



**Figure S4 Genetic fluorescent sensor monitors cellular Fe2+ sequestration in WT and mutant ferritin expressing cells (third replicate of data in Figure 3). (a-f)** Calibrated free Fe2+ concentrations in E. coli expressing wildtype ferritin from up to 15 hours after induction by 0% (a), 0.002% (b), 0.005% (c), 0.01% (d), 0.02% (e), and 0.2% (f) rhamnose with media Fe2+ supplement of 0 μM, 10 μM, 100 μM, 200 μM, 500 μM, 1 mM, 2 mM, 5 mM. Without ferritins high media supplement at 5mM can dramatically alter the intracellular iron homeostatic setpoint. At high induction levels free Fe2+ is efficiently sequestered up to the highest media supplement concentration. The heatmaps with color saturation proportional to calibrated intracellular free iron levels show that compared to the wildtype **(g)**, the best ferritin mutant **(h)** is much more effective at sequestering iron at lower protein levels (0.01% rhamnose induction) and at high environmental iron concentration (up to 2 mM). This is consistent with their greater magnetism despite lower protein expression.

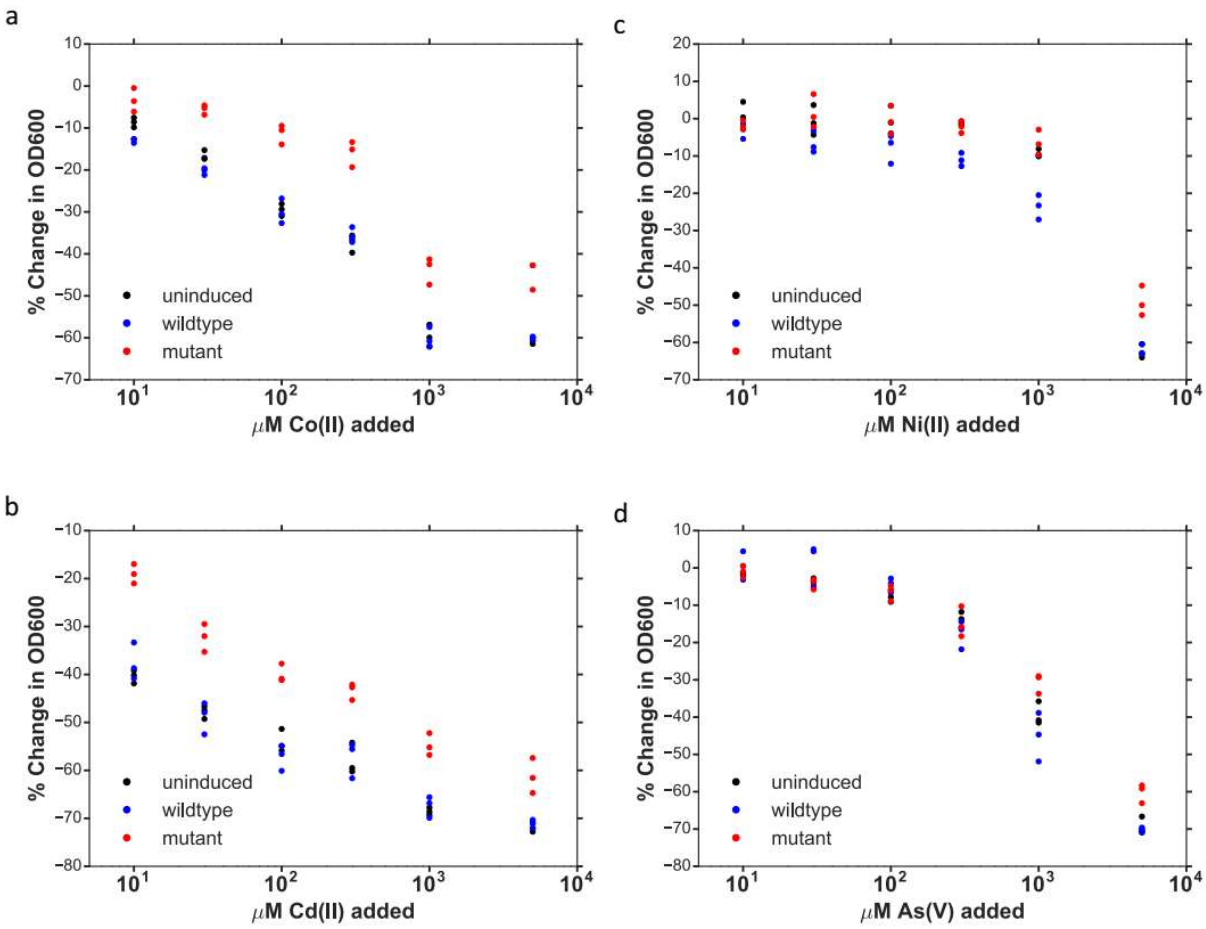



**Figure S5 Magnetic mutant ferritin over-expression reduces cellular growth defect.** Percent change in OD was measured in early saturation phase for cells expressing no (grey), wildtype (blue) or magnetic mutant (red) ferritins and incubated in LB media containing between 0 and 5mM of cobalt (a), cadmium (b), nickel (c) and arsenic (d). Expressing the magnetic mutant ferritin exhibited the least growth defect whereas expressing no or the wildtype ferritins had similarly strong growth defect. The effect is most pronounced at low concentrations for cadmium and cobalt.

| | |
|---|---|
| m1 | H34L+T64I |
| m2 | SpyTag(N-term)+R56P |
| m3 | F58L+T116R |
| m4 | SpyTag(N-term)+T64I |
| m5 | SpyTag(N-term)+A47T |
| m6 | T64I+T116R |
| m7 | R56P+H128R |
| m8 | SpyTag(N-term)+F58L |
| m9 | SpyTag(N-term)+H34A+T64A+F58A |
| m10 | SpyTag(N-term)+H128R |
| m11 | T116R |
| m12 | F58L+H128R |
| m13 | T64I |
| m14 | T116R+H128R |
| m15 | T64I+H128R |
| m16 | H128R |
| m17 | F58A |
| m18 | SpyTag(N-term) |
| m19 | T64I+F58L |
| m20 | SpyTag(N-term)+T116R |
| m21 | F58L |
| m22 | A47T |
| m23 | H34A+T64A+F58A |
| m24 | H128A |
| m25 | R56P |
| m26 | H34A |
| m27 | T64A |
| m28 | L18Q |
| m29 | K140A |
| m30 | L104Q |
| m31 | R56A |
| m32 | K156A |
| m33 | H34A+T64A+F58A+T116R+H128R |
| m34 | H34A+T64A+F58A+T116R |
| m35 | R56P+T116R |
| m36 | L10P |
| m37 | SpyTag(N-term)+H34A+T64A+F58A+T116R+H128R |
| m38 | SpyTag(N-term)+H34A+T64A+F58A+T116R |

**Table S1 List of ferritin mutants characterized with iron sensor and magnetic column**

| | |
|---|---|
| **Iron Sensor Promoter** | TACTGCAACCATCTACAAATAACCCCACCAACGTAACTGCATTCGCCT<br>CTGGATGCAGTTTTCGTCTGTTTTGTAATCAAAAAACATCAAATATGAA<br>CTCAATGTAAATAAATGTATTTCTTTTTCGCGCAATGGGTGATAGAAAA<br>TCGCTCCAAGTGATAATGCTTATCAAAATTATTATCACTTTCACGAGCA<br>CTATCACGGGATTAACAGTGGCATCGCATCCGCAGAGAGGCTTTCTC<br>GTGGCAGTGAAAATTTCAACATATAAGAAAAAGTCACCTGCAAA |
| **Iron Sensor RBS** | AGGAGGTAAAAA |
| **Iron Sensor Reporter (sfGFP)** | ATGCGTAAAGGCGAAGAGCTGTTCACTGGTGTCGTCCCTATTCTGGT<br>GGAACTGGATGGTGATGTCAACGGTCATAAGTTTTCCGTGCGTGGCG<br>AGGGTGAAGGTGACGCAACTAATGGTAAACTGACGCTGAAGTTCATC<br>TGTACTACTGGTAAACTGCCGGTACCTTGGCCGACTCTGGTAACGAC<br>GCTGACTTATGGTGTTCAGTGCTTTGCTCGTTATCCGGACCATATGAA<br>GCAGCATGACTTCTTCAAGTCCGCCATGCCGGAAGGCTATGTGCAGG<br>AACGCACGATTTCCTTTAAGGATGACGGCACGTACAAAACGCGTGCG<br>GAAGTGAAATTTGAAGGCGATACCCTGGTAAACCGCATTGAGCTGAA<br>AGGCATTGACTTTAAAGAAGACGGCAATATCCTGGGCCATAAGCTGG<br>AATACAATTTTAACAGCCACAATGTTTACATCACCGCCGATAAACAAAA<br>AAATGGCATTAAAGCGAATTTTAAAATTCGCCACAACGTGGAGGATGG<br>CAGCGTGCAGCTGGCTGATCACTACCAGCAAAACACTCCAATCGGTG<br>ATGGTCCTGTTCTGCTGCCAGACAATCACTATCTGAGCACGCAAAGC<br>GTTCTGTCTAAAGATCCGAACGAGAAACGCGATCATATGGTTCTGCTG<br>GAGTTCGTAACCGCAGCGGGCATCACGCATGGTATGGATGAACTGTA<br>CAAATAATAA |
| **Rhamnose Promoter+ RBS** | CACCACAATTCAGCAAATTGTGAACATCATCACGTTCATCTTTCCCTG<br>GTTGCCAATGGCCCATTTTCCTGTCAGTAACGAGAAGGTCGCGAATT<br>CAGGCGCTTTTTAGACTGGTCGTAGAGACCATGAAATTCTTTTAAGGA<br>GGTAAAAA |
| **WT ftnA (ferritin)** | ATGCTGAAACCAGAAATGATTGAAAAACTTAATGAGCAGATGAACCTG<br>GAACTGTACTCTTCACTGCTTTATCAGCAAATGAGCGCCTGGTGCAGC<br>TATCATACCTTCGAAGGTGCTGCCGCGTTCCTGCGCCGTCACGCCCA<br>GGAAGAGATGACGCATATGCAGCGTCTGTTTGATTACCTGACTGATAC<br>CGGCAATTTACCGCGTATTAATACCGTTGAATCTCCGTTTGCTGAATA<br>TTCCTCACTTGATGAATTATTCCAGGAAACCTATAAACACGAACAATTA<br>ATCACCCAGAAAATTAACGAACTGGCTCATGCTGCAATGACCAATCAG<br>GACTACCCAACATTTAATTTCCTGCAATGGTATGTTTCTGAGCAGCAT<br>GAAGAAGAGAAACTGTTCAAATCGATTATTGATAAATTAAGCCTGGCA<br>GGCAAAAGCGGCGAAGGTCTGTATTTTATCGACAaAGAACTCTCTACC<br>CTCGACACACAAAACTAA |
| **m1 mutant ftnA** | ATGCTGAAACCAGAAATGATTGAAAAACTTAATGAGCAGATGAACCTG<br>GAACTGTACTCTTCACTGCTTTATCAGCAAATGAGCGCCTGGTGCAGC<br>TATCtTACCTTCGAAGGTGCTGCCGCGTTCCTGCGCCGTCACGCCCA<br>GGAAGAGATGACGCATATGCAGCGTCTGTTTGATTACCTGACTGATAt<br>CGGCAATTTACCGCGTATTAATACCGTTGAATCTCCGTTTGCTGAATA<br>TTCCTCACTTGATGAATTATTCCAGGAAACCTATAAACACGAACAATTA<br>ATCACCCAGAAAATTAACGAACTGGCTCATGCTGCAATGACCAATCAG<br>GACTACCCAACATTTAATTTCCTGCAATGGTATGTTTCTGAGCAGCAT<br>GAAGAAGAGAAACTGTTCAAATCGATTATTGATAAATTAAGCCTGGCA<br>GGCAAAAGCGGCGAAGGTCTGTATTTTATCGACAaAGAACTCTCTACC<br>CTCGACACACAAAACTAA |
| **SpyTagged ftnA** | ATGGCCCACATCGTTATGGTCGACGCGTATAAGCCGACCAAACTGAA<br>ACCAGAAATGATTGAAAAACTTAATGAGCAGATGAACCTGGAACTGTA<br>CTCTTCACTGCTTTATCAGCAAATGAGCGCCTGGTGCAGCTATCATAC<br>CTTCGAAGGTGCTGCCGCGTTCCTGCGCCGTCACGCCCAGGAAGAG |

| | |
|---|---|
| | ATGACGCATATGCAGCGTCTGTTTGATTACCTGACTGATACCGGCAAT<br>TTACCGCGTATTAATACCGTTGAATCTCCGTTTGCTGAATATTCCTCAC<br>TTGATGAATTATTCCAGGAAACCTATAAACACGAACAATTAATCACCCA<br>GAAAATTAACGAACTGGCTCATGCTGCAATGACCAATCAGGACTACCC<br>AACATTTAATTTCCTGCAATGGTATGTTTCTGAGCAGCATGAAGAAGA<br>GAAACTGTTCAAATCGATTATTGATAAATTAAGCCTGGCAGGCAAAAG<br>CGGCGAAGGTCTGTATTTTATCGACAaAGAACTCTCTACCCTCGACAC<br>ACAAAACTAA |
| **A0A072C8A3** | ATGCATAAGGCCTCCGAACACCTTCAGGCCTGGTTGCGCGACGCTCA<br>CGCGATGGAAGAGCAGGCAATCACAATGCTGACTTCCCAGTCTCGTC<br>GCCTTGAAAACTATCCTGAGCTTAAAGCGCGTATTGATCGCCATTTGC<br>AGGAAACCCGTGACCAAGCCGCCATGTTGAAGCGTTGTTTGGAGCGT<br>TTGCACGGAGGGACATCAGCCGTTAAAGATATTAGTGGTAAAATTGTA<br>GCGATCGGTCAAGGCCTTTCTGGGTTATTCGTATCCGACGAAGTGGT<br>TAAAGGCTCCTTGGCAAGTTACACTTTTGAGCACATGGAAATCGCTTC<br>ATATAAGATCCTGATTGCCGCCGCAGACTACGCAGGCGACCAGGAGA<br>CCAAACGTGTTTGCGAAACTATTTTGCAACAAGAAGTCGAGATGGCC<br>GAATGGTTGGACCAACATAGCGCCGAGATCACGCGTACATTCTTAGA<br>ACGTGACCAACGCGGGGTTACTGCGAAGCAC |
| **F7XJ33** | ATGGCAACAGCAAACGAGCACTTGGTCGCGTGGTTACGCGATGCCCA<br>TGCCATGGAAGAACAGGCAATCACAATGCTTACGAGTCAGAGCTCCC<br>GTCTGGAGAACTATCCCGAGCTGAAGACACGCATCGACCGTCATCTT<br>CAGGAAACAAAGGACCAAGCAGCAATGCTTGAACGTTGCTTAGAACG<br>CTTGGATGGGGGTACTTCGAGCATTAAGGATATTAGCGGCAAGATCG<br>TGGCCTTCGGGCAGGGTTTAAGCGGGTTATTTGTTTCGGACGAGGTA<br>GTCAAGGGGACCTTGGCAAGCTATACCTTTGAGCACATGGAAATCGC<br>CAGCTATCGTATCCTGATCGCGGCGGCGGAACAGGCTGGCGACCAA<br>GAGACTAAACGTGTTTGCGAGAGCATTTTACAACAAGAAATCGCAATG<br>GCAGAGTGGTTAGCGCAGAACGCAGGGGAGATCACACGCAAGTTTTT<br>AGAACGCGATCAGCGTGACGTGACGGCCAAACAC |
| **Q1QGY5** | ATGAGTCGTATCGAGGAAAATCTGATGGCGTGGTTACGCAATGCCCA<br>TGCTATGGAGGAACAGGCAGTCACTATGTTGACCAGTTTGGCCTCTC<br>GTACTGGCGATTATCCAAATGTCAAAGCACGCATTGAATCACATCTGG<br>CAGAAACGAAGCGTCAGGCCGAAGCTCTGGAAGAATGCATTAAGCGC<br>CGTGGTGGAGAGACAAGTACGTTGAAGGACTTAACTGCGAAAATGCT<br>GGCTTTCGGTCAAGGTTTGTCGGGAATGTTTGTCGATGATGAGATTGT<br>AAAGGGGGCAATGGCTTCTTATACATTCGAGCACATGGAGGCTGCTG<br>CATACCGCGTGTTGATCGCCGCGGCGGAGGCGGTCGGGGATACGCA<br>GACACAGGCAGTATGCGAGCGTATTCTGCAAGAGGAGTTGTCTATGG<br>CGTCGGATCTTGAAGATCACCTTCCCGAATTGACGCGCAAGTATCTG<br>AATCGTTTAAAGCCCATGCGCATGAGCTCC |

**Table S2 DNA sequences of relevant genes and constructs expressed and tested in Chapter 3.**

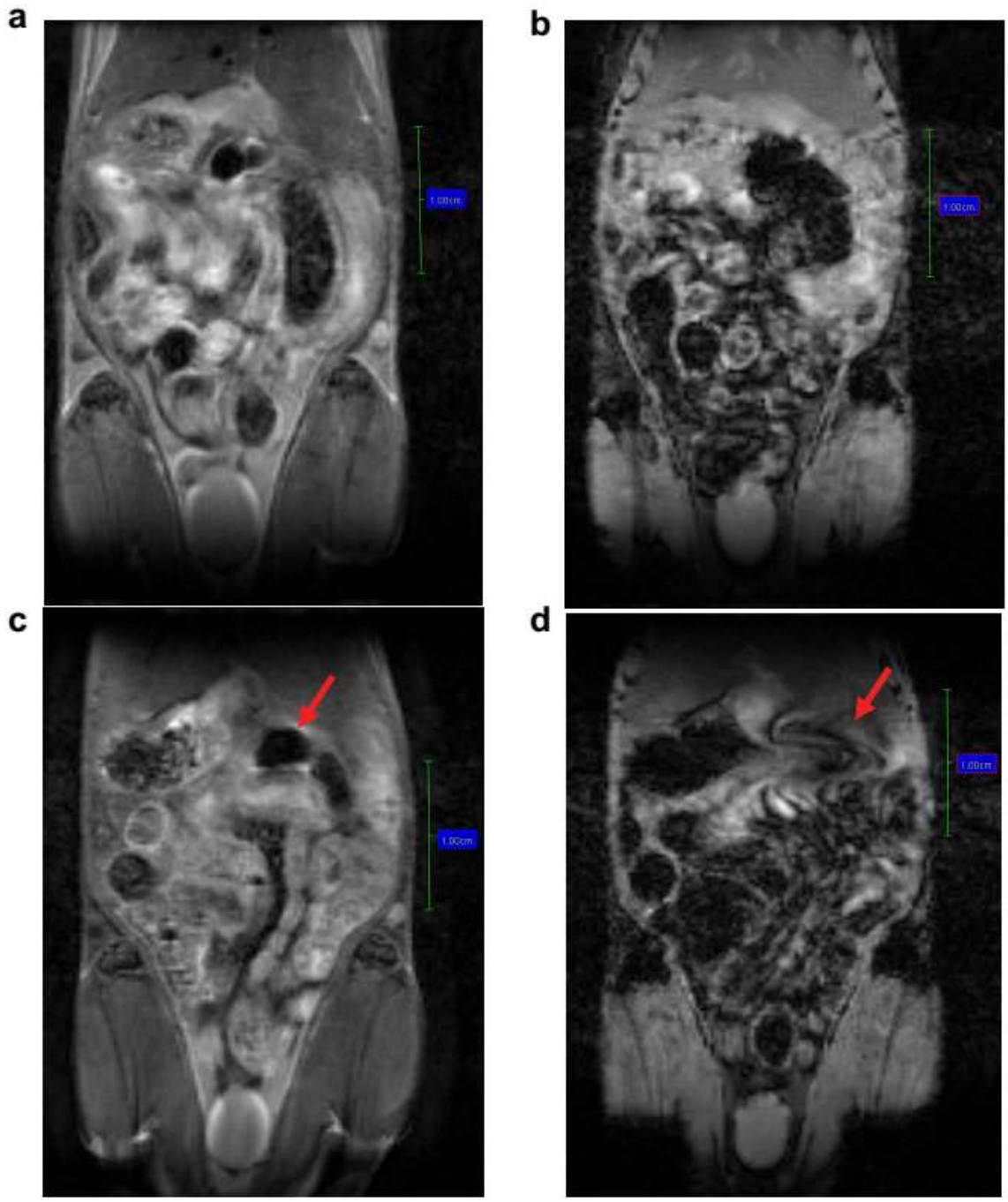



**Figure S6 *in vivo* MRI imaging of *E. coli* gavaged to mouse (a)** T2-weighted image for the mouse gavaged with uninduced *E. coli* along with **(b)** its T2*-weighted image show some dark artefacts due possibly to gas or stool. **(c)** T2-weighted image and **(d)** T2*-weighted image for a separate mouse gavaged with *E. coli* pre-induced to over-express ferritin mutant m1 show a clear dark spot in the T2-weighted image corresponding to a localized distortion in the T2*-weighted image (red arrows) suggestive of bolus of magnetized bacteria passaging through the gut. The green scale bar represents 1cm.

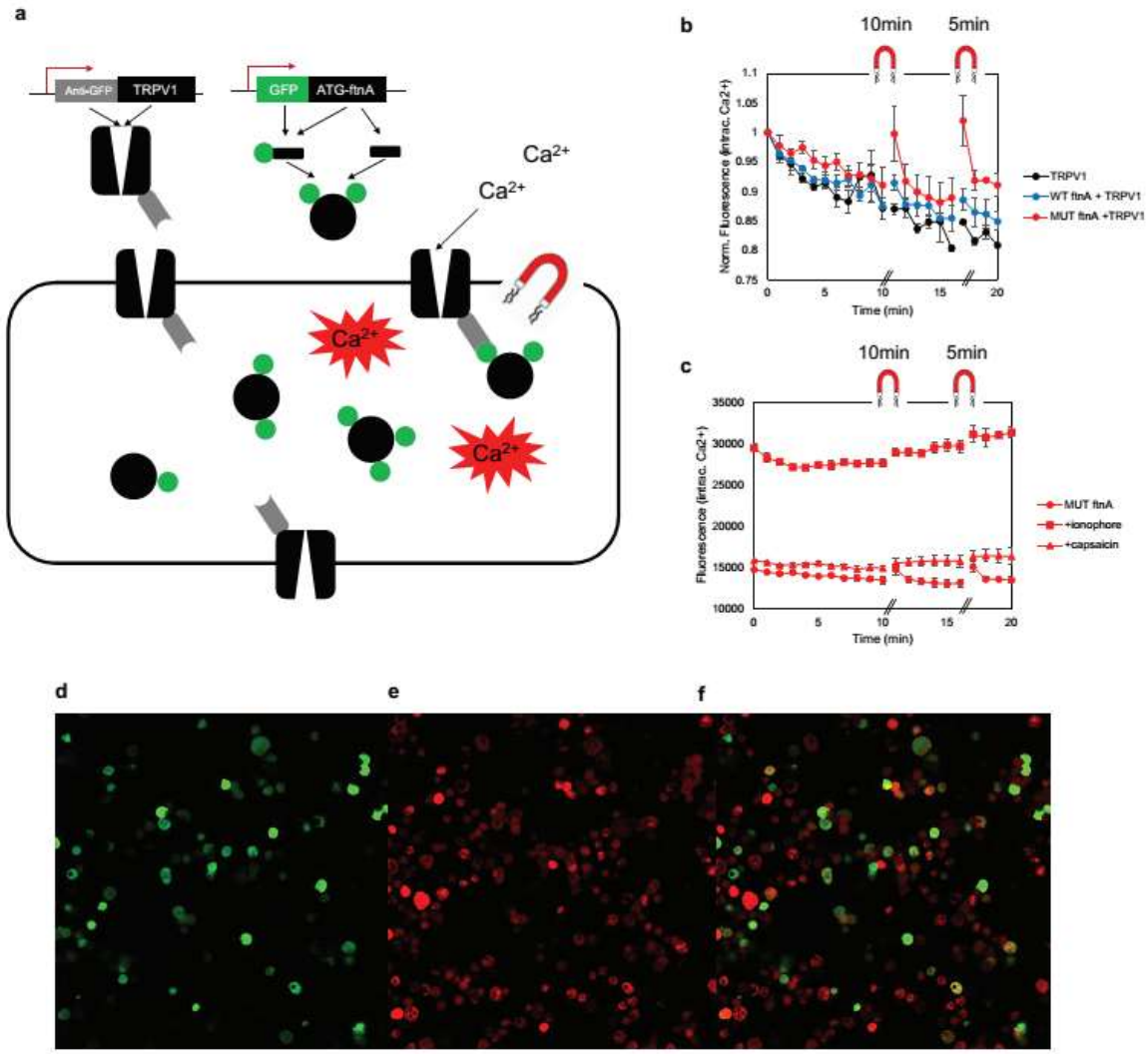



**Figure S7 Genetically-encoded biomagnetism controls signaling in mammalian cells (a)** Schematic of HEK293 cells co-transfected to express anti-GFP nanobody tagged TRPV1 calcium channel and GFP tagged ferritin (wildtype or mutated). The magnetic field induces pressure on the channel to open it for calcium influx. **(b)** Calcium measurement by fluorescent dye (X-Rhod-1) in HEK293 cells expressing either TRPV1 alone (black), or TRPV1 along with wildtype ferritin (blue) or best mutant ferritin m1 (red). Static magnet was applied for 10 minutes after 10 minutes of measurement, and 5 minutes at 16 minutes of measurement. Only the mutated ferritin plus TRPV1 shows increases in fluorescence corresponding to calcium intake. **(c)** Positive controls for the mutant ferritin and TRPV1 co-transfected cells show that the addition of 10uM ionophore drastically increases overall calcium signal (square), and addition of 100uM TRPV1 agonist capsaicin also increases intracellular calcium at a constant level (triangle), whereas the application of magnetic field to non-treated cells increases fluorescent to the same level only transiently (circle), consistent with the lack of magnetic stimulus required for channel opening. Confocal microscopy images of cells co-transfected with gfp-tagged ferritins and TRPV1 channel fused to anti-GFP nanobody showing the gfp channel alone **(d)**, red channel for calcium dye imaging **(e)** and overlay **(f)**.

## APPENDIX C

# SUPPORTING INFORMATION FOR CHAPTER 4

| | | |
|---|---|---|
| **length** | **A** | **A%** |
| **# of Positive +AA** | **C** | **C%** |
| **# of Negative -AA** | **E** | **E%** |
| **MW** | **D** | **D%** |
| **aromaticity** | **G** | **G%** |
| **instability index** | **F** | **F%** |
| **gravy** | **I** | **I%** |
| **Isoelectric Point (pI)** | **H** | **H%** |
| **%helix** | **K** | **K%** |
| **%turn** | **M** | **M%** |
| **%sheet** | **L** | **L%** |
| | **N** | **N%** |
| | **Q** | **Q%** |
| | **P** | **P%** |
| | **S** | **S%** |
| | **R** | **R%** |
| | **T** | **T%** |
| | **W** | **W%** |
| | **V** | **V%** |
| | **Y** | **Y%** |

**Table S1 ProtParam Features for training of logistic regression and random forest models**. “Length” is the total number of amino acids in the sequence. The capital letter represents the number of times the corresponding amino acid (e.g. A = Alanine) appears in the sequence. The capital letter followed by “%” sign represents the percentage of that amino acid in the entire sequence (e.g. A% = A / length). The positive amino acids (+AA) feature sums the total number of Arginine (R) and Lysine (K) in the sequence, which are charged positively at pH 7. Similarly, the negative amino acids (-AA) feature sums the total number of Aspartic Acid (D) and Glutamic Acid (E) in the sequence. “MW” is the theoretical molecular weight calculated from the sum of weights of each amino acid in the sequence. The other parameters are calculated using the ProtParam module of the Biopython package. Specifically, aromaticity is the total percentage of Phenylalanine (F), Tyrosine (Y) and Tryptophan

(W) in the sequence. “Instability index” is calculated according to Guruprasad et al. 1990[1]. The grand average of hydropathicity (gravy) is calculated as the sum of hydropathy values of all the amino acids divided by the number of residues in the sequence according to Kyte and Doolittle 1982[2]. The percentages of helix (%helix), turn (%turn) and sheet (%sheet) are calculated as sum of percentages of amino acids in the sequence that tend to be in helix (V, I, Y, F, W, L), turn (N, P, G, S), and sheet (E, M, A, L). The documentation and source code for the ProtParam package can be accessed at http://biopython.org/DIST/docs/api/Bio.SeqUtils.ProtParam.ProteinAnalysis-class.html, with additional documentation at http://web.expasy.org/protparam/protparam-doc.html

**References:**

1. Guruprasad, K., Reddy, BVB., Pandit, MW. Correlation between stability of a protein and its dipeptide composition: a novel approach for predicting in vivo stability of a protein from its primary sequence. *Protein Engineering*. **4,** 155-161 (1990).

2. Kyte, K. and Doolittle, RF. A simple method for displaying the hydropathic character of a protein. *Journal of molecular biology*. **157,** 105-132 (1982).

| | |
|---|---|
| **fungi** | ATGGCTCCCAAGAGCTTGCTTCAACACGCAGGGCTGCTGGCCCTGGCTA<br>GCGGTGCTTTTGCAGTACCGTTTGTCTCCAAAGCTCAGACCACGGTTACC<br>TCAGAACCTACGATTACAGCGTCCCAAGTACCACACACTAACGTAACGTC<br>TCATGGTCCTTACACTGGTCCAAGTCCTACCACCACCGGAGCAATTTCAA<br>CAAGTGTTTTGGCATCTGCCGTCCCTATTCTGCCTCCCCCCGACGACGCC<br>TATGACTACCCCGCCGACGGAAAGTTGCACGGGGATCAACCCGCGCCTT<br>ATACTCCAAGTGGAGGCATCGGGACTAATGGATCGGCCCCAGTTTACCG<br>CGTCCAGTCCGATTTTGACTACCAATCCCTTGCTTTGGCATTGTATCAAGA<br>ATACATTGAGCTGGACCTGTTTCATTGGGGGTTAGATACGTATAGTGAAG<br>CAGATTTTGCCGAATTGGGCCTGAATGCTGAAGACCGCTTTCTTTTACAG<br>CATATGGCGGACCAGGAGATTGGGCATGCCACAGTTATTACTAATCTGTT<br>AGGGCCTCAAGCACCCCGCCAGTGTACGTACAATTATCCAGTGTCCAATC<br>TGCGCGAATACATTGATTTTAACCAAAAGTTAACCCGTTGGGGCGAGGCA<br>GGGGTGTACGGATTCTTACCCCACCTTAATTCCGGACCTGCAGCGCAATT<br>ACTGTTACAAAGTATCACTGTAGAAGCGCGTCAACAGATGATTTTCCGCC<br>AATTCGAGGGTTTATTTCCAATGCCCGAATGGCATACT |
| **lancelet** | ATGGCTCATGCTCTTAGCATTGCTAAATGGCACTCTCTGGAGACTGATTG<br>GCAGTGTTCAAAGAACATGTCTTACACCCTTACTGAAAAGAGCATGGCCC<br>ATGCCTTGCCGATTGCTAAGTGGCATAGTTTAGAAACAGATTGGCAATGC<br>AATAAGAACATGTCATATACCTTAACAGAGAACCAGTGGCACATGCATTAT<br>CTT |
| **virus** | ATGGATAACATCGACTACGACAACGGGTTTATCTGCACATACAATTTAATT<br>GAGGACGACCAGGAAAGCATTATCTGTTACCAGGCGCAATTACTTCAGGC<br>ACTGAAGCAGCCTGTATATGATGACGACAAGATTAGTATCATCACGGGTA<br>AAATCTATGATTTACTGAAAGATAACGAGGAAATTCAGGAAATTCTTAATAT<br>TTTGTCTGAGAAGCTGGTCCTGTTTCAATTTTTCAAATCGAATAATAAGGA<br>ATTGGACAACACGTTCGTCTTTCCTATGCTGTTTTCTTTCGAGTATTTCCA<br>CTTATTCCATAAACTTCTTTCCAATTATATCTCAAATAAGTCGCTGCAAGAG<br>AACCATTTCTCTCAAATTAAGCAACTGATTTGCGACAAC |
| **algae** | ATGGACCACGGTACGGTTGCAGGGGCAGTCGCAACCCAAGCTCCCCCAT<br>GCCCCGCCACGGTAGTGCCCGCATCCAAACCCGGACTTATCTCCCATGA<br>CGAGGTCCGCGCATTGGAATCGTCGTTGGACCACGGTCACTCCGTCCGC<br>GCCCTGTACCGTTTTGACGCAGCTTTACGCGCGTTATCCCTGGATGAGTT<br>CGCCGCTAAATTACGCCCAGTCATGCTGGAATTAGCGTTTAACCTGGAAA<br>GTTTTAACCCGGCGTTCACGGCTGTCTTACTGAAAGTACAGAGTACTATC<br>GGGTTATCGGGTTACCATGCACGCGACGAGGCCTTAAAGCGCCTGATCC<br>TGCCTTTGGCAGGAGAATACGGTTTACACGCTGATCAGCCGCAAGGGGC<br>AACCCATCGCGCTCTTTTCGCGGAGTTCTATGAGGATTTGCTTGGCGAAG<br>CGCTTGAGCCCGCGATGCGCGCTGCAGCTACATCATTTCCGCCAGCTGC<br>CGCAATCTTGTACTCCCAGATGATGCGTGACGTGCTTGGTGGGGGAGGA<br>CGCACCGGCGATGCTCAAGAGCAAGCCTCGTATGCTTTGGGGTACAACC<br>TTGCAATTGAATACCTGGCGGACGTTGAGAAACGCCTGGAATTAGAAGCG<br>TTCCGTCACTTGGACGAGCGTGTCCTTGCACCCGCTGGACGTCGTTTACG<br>TGCTTGGGAGTTTCTGGAAGTACATGCTGAAGGCGAGGCAGAACACGCG<br>GCAATCGGCCACGCTGCTGCTTGCGGATTGGTTCCTGCGGACCACGCAG<br>GGTTACTGCATGCCGCTGCTGCCGACCATGACCGCGACTTTGCGGCCTT<br>CTACGACGCGTTGGCAGCAATGCTGGAG |
| **potato** | ATGGAAGCTAGTGGTAATCGCCAGGCGATCACTGAACTTCTGGCACTTAC<br>TTACCGCAAGAAAGATGTACCTAAAATGGATTTGCAAGATATTGCGGAAAT<br>GTTCCATCTGTCTGCGTTAATCGCCATGAAATTTGCTTTTCACAATTTGAA<br>GGATTCGGATCATCAAATGCGTAACCGTACAATCAAGGCTTTGATCGCCT<br>TAGAAGTTTTGAATTACCAATATTACTTTGCAACGATGTCGTCATCCGCGA<br>AT |

| | |
|---|---|
| **human** | ATGAAGATGATGTTGAGCAAGCAGAAGAGTGTTCTTCGTATGCGCTCGTC<br>ACCAAAGGTTCAGAGCTCGAAGCGTCTGGGTTGGCAAGGATGCCTGTGG<br>GTTCATAGCGTCGGCCATTCACCACAACGCTCGGATACGAGCGCGGTT |
| **archaea** | ATGTCTAAAATGGACCCTTTTCTTCGTGGTTTTTCAGAACAATTCATGGAT<br>ATCCATTTGGAGGAAAAAGAGATCCAGAATATTAAGGGTATCATTAACATT<br>TTCAAAGGAGGCATTAAGAGCAATTTGGACGCGGAAATCGGCTTTTTACT<br>TGGTTATGCTTACGCAGAATTGTTGATGCAATTCTTAATCTTGAAAAACCG<br>TCTTCCAAACAAAGATGAGAGTAGCGAGTACTACACTATTATGAAACGCC<br>GCTTTCCCGAAATCATTCAACAAATTAAAAAGCAGAAAAAGAGCGAGCTG<br>CTGGATCGTGATGACGTAGTCATTAATGTATCCGAAGTGAATGTCGAGCA<br>GATGAACACTATCCAGGAG |
| **mouse** | ATGGAGACTGAATCGGAGCAGAACAGCTCCTCGACTAACGGATCAAGTTC<br>TTCAGGTGCCTCTAGTCGTCCCCAAATTGCCCAAATGAGTCTGTATGAAC<br>GTCAAGCCGTGCAGGCACTGCAGGCACTGCAACGTCAGCCCAATGCCGC<br>CCAGTATTTCCATCAGTTCATGTTGCAACAGCAATTGTCGAATGCACAACT<br>GCACAGCCTGGCCGCAGTACAGCAAGTGCGTGGGCATGATTGCTGTCAG<br>AGTACCGGC |
| **cyano** | ATGGCTCAATGGGAATCAATCGACCACGATGGTCTGTTTAAGGAGTTAAT<br>TGAAACATTCTTCTGGGAGTTTTTGAATTTGTTTCTGCTTCAGGTCTTCGA<br>CTATGTCGAGCGTGGACCAGTCACATTCTTGTTCCAAGAAGTGTACTCCT<br>CAATCGGGGCCGAGGAGCGTCGCATTATCGACTTACTTGGCGCCGGGTC<br>GTTAGCCCTGACTATTTCCCCTCTTCTGAATGTTAAGTCCTGTTGGATT |
| **gut** | ATGAAAAACGATGCAAGCCTTCCGATTGTGTTGATGTTTCACGATAGCGA<br>CTATGCCTTAGTGGTGGCCGATAATCTTGAAGACTTCATCTTCTATAAAAT<br>GTTGATGGCTGCCTTTAAGGTAGATGATGAATGTGACGATGAAACTTTCAA<br>GGCCGAGCTTCAACGTGTGCTGTCCACTCACAAGAAGTATTTGAAGAAGG<br>AGTATGTGAAAGTCCTTGAGCAGCTTTATGACGGTCCCATCGCGAAACTG<br>TCGGATTGCGATTTGGAGAAGTTGTATATTAAGGAGACTGCTTTCACCCG<br>TTTTAACGAAGAACTGTATTGCGGTATT |

**Table S2 DNA sequences of relevant genes and constructs expressed and tested in Chapter 4.**