Statistical Models of the Protein Fitness Landscape: Applications to Protein Evolution and Engineering

Thesis by

Philip A. Romero

In Partial Fulfillment of the Requirements

for the Degree of

Doctor of Philosophy

California Institute of Technology

Pasadena, California

2012

(Defended December 16, 2011)

Acknowledgements

I wish I could stay a grad student at Caltech indefinitely. I appreciate the many aspects that make Caltech a unique and stimulating environment. Caltech's close community has felt like a family to me (and literally is a family). Over the past six years, I have truly experienced the time of my life. I've made many great friends and learned so much about science and research. Every day I have been thankful for the opportunity to pursue graduate studies at this wonderful institution.

I would like to start off by thanking my advisor, Frances Arnold. Frances has provided important and interesting problems, while allowing the freedom to approach them in unique and creative ways. She has instilled in me the ability to identify interesting scientific questions and to always maintain a practical perspective. The lessons learned in her lab will be invaluable for the rest of my scientific career. I am grateful for the stimulating research environment that Frances has created, which includes a diverse group of engineers, biologists, chemists, and computational scientists, all collaborating on important research problems.

While at Caltech, I have had the opportunity to work with many exceptional scientists. I appreciate the assistance and guidance that every Arnold lab member has provided to me over the years. I would like to specifically thank several group members. Jesse Bloom served as a mentor when I first started in the lab and introduced me to many concepts in molecular evolution. Chris Snow taught me nearly everything I know about computational methods and inspired many of my current interests. Matt Smith, the last remaining computational group member, has been great to bounce ideas off of and always provided valuable feedback. And of course I would like to thank the lab managers, Geethani Bandara and Sabine Bastian for keeping the lab running, providing experimental guidance, and not getting too frustrated with my slow turnaround times for multichannel pipette

repairs. In addition, I am grateful to all the chimeraologists over the years that have contributed to the massive data set that I analyze on a daily basis.

I have also had the opportunity to work with several great collaborators outside the Arnold lab. The MRI project was a collaboration with Mikhail Shapiro, a graduate student from MIT, whose energy and enthusiasm are contagious. It was also a real pleasure to work with Everett Stone, a postdoc from UT Austin, on the arginase SCHEMA library. Professor Andreas Krause's class on active learning captured my imagination and since we've worked on several of the algorithms presented in this thesis.

I am grateful to my committee members Harry Gray, Steve Mayo, Doug Rees, and Zheng-Gang Wang. They have provided helpful guidance and were always available if I had questions. I also enjoyed interacting with Steve Mayo's group and participating in their group meetings.

I would like to acknowledge the Biochemistry and Molecular Biophysics (BMB) option. Specifically, I would like to thank Alison Ross and Doug Rees for managing a great program and for always supporting the students' best interest. I would also like to thank the first people I met when I arrived at Caltech, my BMB classmates Nick Ballor, Jason Cute1, Kelly Dusinberre, Russ Ernst, Peera Jaru-Ampornpan, Kyle Lancaster, Cam Liu, John Ngo, Fred Tan. They have been great friends over the past six years.

Last, I would like to thank my family. My parents for granting me the freedom to make mistakes and grow from these experiences, and for always providing unconditional support. My two younger brothers, which for some unusual reason have always served as role models for me. After attending colleges in different states, we have been so fortunate to spend the last five years at Caltech together. I would also like to thank Peter and Shalini Venturelli for their support and advice during my graduate studies. Finally I would like to thank my wife, Ophelia, for her love and encouragement. It's been great to learn the ins and outs of science and research together.

Abstract

Understanding the protein fitness landscape is important for describing how natural proteins evolve and for engineering new proteins with useful properties. This mapping from protein sequence to protein function involves an extraordinarily complex balance of numerous physical interactions, many of which are still not well understood. Directed evolution circumvents our ignorance of how a protein's sequence encodes its function by using iterative rounds of random mutation and artificial selection. The selection criteria is based on experimental measurements, which permits the optimization of protein sequence properties that are not understood. While directed evolution has been useful for exploring protein fitness landscapes, these searches have been relatively local in comparison to the vast space of possible protein sequences. Here, we present several classes of statistical models that map protein sequence space on a larger scale. We use these simple models to interpret data from SCHEMA recombination libraries, understand the evolutionary benefit of intragenic recombination, and design optimized protein sequences. By training on directly on experimental data, these models implicitly capture the numerous and possibly unknown factors that shape the protein fitness landscape. This provides an unrivaled quantitative accuracy across a massive number of protein sequences.

Contents

Acknowledgements **iii**

Abstract **v**

1 Introduction: Exploring protein fitness landscapes by directed evolution **1**

1.1 Abstract 1

1.2 Introduction 1

1.3 Protein Fitness Landscapes 4

1.4 Strategies for Directed Evolution 6

1.5 Lessons from Directed Evolution 12

1.6 Conclusions 17

1.7 Figures 19

2 Directed evolution of a magnetic resonance imaging contrast agent for noninvasive imaging of dopamine **24**

2.1 Abstract 24

2.2 Introduction 25

2.3 Results 26

2.4 Discussion 33

2.5 Methods 35

2.6 Supplementary Material 42

2.7 Figures 46

3 SCHEMA-designed chimeras of Human Arginase I and II reveal sequence elements important to stability and catalysis 56

3.1 Abstract 56

3.2 Introduction 57

3.3 Materials and Methods 59

3.4 Results 62

3.5 Discussion 66

3.6 Figures and Tables 69

4 Random field model of the protein recombinational landscape 73

4.1 Abstract 73

4.2 Introduction 73

4.3 Results 75

4.4 Methods 85

4.5 Discussion 88

4.6 Supplementary Material 91

4.7 Figures and Tables 96

4.8 Supplementary Figures and Tables 99

5 Gaussian process models of the protein fitness landscape 101

5.1 Abstract 101

5.2 Introduction 101

5.3 Results 104

5.4 Materials and Methods 113

5.5 Discussion 119

5.6 Figures 123

5.7 Supplementary Figures 126

5.8 Supplementary Tables 131

Bibliography **136**

List of Figures

1.1 Directed protein evolution traverses a fitness landscape in sequence space 19

1.2 Schematic of a directed evolution experiment 20

1.3 Recombination generates highly mutated sequence libraries 21

1.4 Directed evolution of cytochrome P450 BM3 for propane hydroxylation 22

1.5 Stability-mediated mutational epistasis 23

2.1 Ligand binding to BM3 heme domain changes MRI contrast and optical absorption . 46

2.2 Screen-based isolation of BM3h mutants with enhanced dopamine affinity 48

2.3 Evolved sensors produce strong and specific MRI signal changes in response to dopamine 49

2.4 BM3h-based sensors measure dopamine release in cell culture 50

2.5 BM3h-8C8 reports dopamine in injected rat brain 52

2.6 Stability of BM3h complexes 53

2.7 Dopamine binding to BM3h-B7 and BM3h-8C8 is reversible 53

2.8 BM3h-8C8 reports DA release from PC12 cells 54

2.9 Histological data from a rat injected with 500 μM BM3h-8C8 and 500 μM DA 54

2.10 Statistical analysis of K^+ stimulation data 55

2.11 Time courses of MRI signal in the six individual animals 55

3.1 Arginase SCHEMA library 69

3.2 Arginase long-term stability 70

3.3 Correlation between relative metal affinity and catalytic efficiency 70

3.4 Schematic of potential arginase inactivation mechanisms 71

4.1 Effect of homologous substitutions on the fraction of functional sequences 98

4.2 Comparison between library properties and their expected values 99

4.3 Comparison of contact parameters estimated on four different recombination libraries 100

5.1 Predictive ability of Gaussian process models 123

5.2 Gaussian process models for P450 enzymatic activity and binding affinity 124

5.3 Upper confidence bound sequence optimization 125

5.4 Gaussian process models with alternate kernel functions 126

5.5 Schematic representation of single- and double-crossover chimeric P450s 126

5.6 Uncertainty in Gaussian process predictions 127

5.7 Schematic representation of the experimentally designed chimeric P450s 128

5.8 Absorbance spectra of the 26 experimentally designed chimeric P450s 129

5.9 Additional Gaussian process models for P450 enzymatic activity 130

5.10 Gaussian process model predictions on an independent enzymatic activity data set . . 132

5.11 Schematic representation of the sequences from the UCB sequence optimization . . . 133

List of Tables

3.1 Chimeric arginase data 71

3.2 Parameters for the arginase long-term stability regression model 72

4.1 Summary of nine protein recombination libraries 97

4.2 Contact parameters estimated from four recombination libraries 99

5.1 Thermostabilities of unpublished single- and double-crossover chimeric P450s 131

5.2 Experimental data collected on chimeric P450s 134

5.3 Pairwise correlations between the measured enzymatic activities and binding affinities 135

Chapter 1

Introduction: Exploring protein fitness landscapes by directed evolution

A version of this chapter has been published as [1].

1.1 Abstract

Directed evolution circumvents our profound ignorance of how a protein's sequence encodes its function by using iterative rounds of random mutation and artificial selection to discover new and useful proteins. Proteins can be tuned to adapt to new functions or environments by simple adaptive walks involving small numbers of mutations. Directed evolution studies have shown how rapidly some proteins can evolve under strong selection pressures and, because the entire 'fossil record' of evolutionary intermediates is available for detailed study, they have provided new insight into the relationship between sequence and function. Directed evolution has also shown how mutations that are functionally neutral can set the stage for further adaptation.

1.2 Introduction

Millions of years of life's struggle for survival in different environments have resulted in proteins providing diverse, creative and efficient solutions to a wide range of problems, from extracting energy from the environment to repairing and replicating their own code. Good solutions to biological problems can also be good solutions to human problems — proteins are widely used in the food, chemicals, consumer products, and medical fields. Not content with nature's protein repertoire, however, protein engineers are working to extend known protein function to new environments or tasks [2, 3, 4, 5] and to create new functions altogether [6, 7, 8].

Despite major advances, a molecular-level understanding of why one protein performs a certain task better than another remains elusive. This is perhaps not surprising when we remember that a protein often undergoes conformational changes during function and exists as a dynamic ensemble of conformers that are only slightly more stable than their unfolded and non-functional states and that might themselves be functionally diverse [9]. Mutations far away from active sites can influence protein function [10, 11]. Engineering enzymatic activity is particularly difficult because very small changes in structure or chemical properties can have big effects on catalysis. Thus, predicting the amino acid sequence, or changes to an amino acid sequence, that would generate a specific behavior remains a challenge, particularly for applications requiring high performance (such as an industrial enzyme or a therapeutic protein). Unfortunately, where function is concerned, details matter, and we just don't understand the details.

Evolution, however, had no difficulty generating these impressive molecules. Despite their complexity and finely tuned nature, proteins are remarkably evolvable: they can adapt under the pressure of selection by changing their behavior, function, and even fold. Protein engineers have learned to exploit this evolvability using directed evolution — the application of iterative rounds of mutation and artificial selection or screening — to generate new proteins. Hundreds of directed evolution experiments have revealed the ease with which proteins adapt to new challenges [12]. Notable recent examples include a recombinase evolved to remove proviral HIV from the host genome (providing a new strategy for treating retroviral infections) [13], a cytochrome P450 fatty acid hydroxylase that was converted into a highly efficient propane hydroxylase (thereby proving that a cytochrome P450 is fully capable of hydroxylating small alkanes, even though most propane-using organisms use structurally and mechanistically unrelated enzymes) [14], a more than 40 °C increase in the thermostability of lipase A (extending its application in biocatalysis to a whole new set of environments) [15], and a variant of green fluorescent protein that tolerates having all its leucine residues replaced with a non-natural amino acid, trifluoroleucine [16]. Roger Tsien won the Nobel Prize last year for his work on the fluorescent proteins that have transformed biological imaging [17]. Directed evolution had a key role by improving many features of fluorescent proteins, including emission and excitation properties, quantum yield, multimerization state, and maturation rate [5, 18].

Directed evolution has become a common laboratory tool for altering and optimizing protein function (as well as the function of other biological molecules and systems, including RNA, DNA regulatory elements, biosynthetic pathways, and genetic regulatory circuits [19, 20, 21]). To understand the power, and the limitations, of directed evolution, it is helpful to view it as a biological optimization process. We therefore introduce the concept of evolution on a fitness landscape in protein sequence space and use this framework to explain directed evolution strategies. Data from laboratory evolution experiments have revealed important features of this fitness landscape and the types of trajectories that can traverse it efficiently. This landscape picture can help explain why decomposing a large functional hurdle into a series of smaller ones and exploiting protein modularity and structural information are useful strategies for dealing with the combinatorial explosion of possible paths in an evolutionary search. This also helps us to appreciate the power of recombination to generate functional sequences with numerous (mostly neutral) mutations, novel combinations of which can give rise to new protein behaviors and therefore new starting points for optimization of protein function.

There is little doubt that directed evolution is one of the most effective and reliable approaches to engineering useful new proteins. Perhaps less well appreciated, however, is how much our understanding of protein function and evolution has been enriched by data from these experiments. Directed evolution allows us to disconnect a protein from its natural context and observe how adaptation to different functional challenges can occur. These experiments can explore the boundaries between biological relevance (the ability of a protein to contribute to the reproductive fitness of an organism) and what is physically possible (the ability of a protein to carry out a specific function *in vitro* or *in vivo*) in ways that studies on natural proteins alone cannot. Directed evolution can test alternative adaptive scenarios, explore the range of possible solutions to a given functional challenge, examine relationships between different protein properties (for example, trade-offs, in which improvements in one property are accompanied by losses of another) and provide biophysical explanations for evolutionary phenomena. Much has been discovered since these topics were first reviewed in the context of temperature adaptation [22, 23]. In this review, we revisit some of these early lessons and discuss new ones that have emerged.

1.3 Protein Fitness Landscapes

In his influential 1970 paper, John Maynard Smith eloquently described protein evolution as a walk from one functional protein to another in the space of all possible protein sequences [24]. He arranged all proteins of length L such that sequences differing by one amino acid mutation were neighbors. Although the distance between any two sequences is small (that is, it equals the number of mutations required to interconvert the sequences and is therefore $\leq L$), this high-dimensional space contains an incomprehensibly large number of possible proteins. For even a small protein of 100 amino acids there are 20^{100} ($\approx 10^{130}$) possible sequences — more than the number of atoms in the universe. Searching in this space for billions of years for solutions to survival, nature has explored only an infinitesimal fraction of the possible proteins [25]. Furthermore, natural evolution keeps only sequences that are biologically relevant; others are discarded, even if they represent solutions to other interesting problems. There are so many proteins waiting to be discovered and we can only dream about the extent of their capabilities. Directed evolution is one way to extend protein function to new, non-natural tasks and convert dreams into actual proteins.

Each sequence in Maynard Smith's protein space can be assigned a 'fitness', which in natural evolution is a measure of the host organism's ability to reproduce in a given environment: fitter organisms reproduce faster and their genes spread throughout the population [26]. When artificial selection is imposed, fitness is defined by the experimenter. High-fitness sequences satisfy all of the criteria for a protein to function as desired, or at least to perform well in the assay used for screening, and might include the ability to recognize one substrate but not another, to be expressed at high levels in a particular host organism, to not aggregate, and to have a long lifetime. Protein evolution can then be envisioned as a walk on this high-dimensional fitness landscape, in which regions of higher elevation represent desirable proteins, and iterations of mutation and artificial selection continuously discover new sequences further uphill, with higher fitnesses (Figure 1.1*A*).

As with any optimization problem, the structure of the objective function (the fitness landscape) influences the effectiveness of a search strategy [27]. Possibilities range from smooth, single-peaked 'Fujiyama' landscapes to rugged, multi-peaked 'Badlands' landscapes [28] (Figure 1.1*B*). The rougher the landscape, the harder it is for evolution to climb. Local optima create traps that evolution cannot escape from unless a side-step or even a temporary decrease in fitness is permitted, or if multiple simultaneous mutations enable a jump to a new peak. The easiest landscape to climb is one that offers many smooth, uphill paths to the desired fitness (the Fujiyama landscape).

This terrestrial landscape analogy should be interpreted cautiously, however, because it cannot accurately represent the numerous possible paths that evolution can take to higher fitness (or the even larger number of possible downhill paths). Although it is easy to visualize being caught on a local optimum in a three-dimensional landscape, a local optimum in protein sequence space (in which all possible mutations are deleterious) might be rare, unless stability has been compromised and few new mutations can be accepted. For example, the introduction of stabilizing mutations can increase a protein's mutational robustness, opening new routes for further adaptation [29, 30].

The vast size of sequence space makes it impossible to characterize (or even model) more than a minute fraction of this fitness surface. Despite this, several important features have emerged from accumulated experimental studies. The first is the low overall density of functional sequences: the vast majority do not code for any functional protein, much less the desired protein [31, 32, 33]. Another important feature is the uneven distribution of functional sequences. Although representing a very small fraction of all possible sequences, functional sequences are often next to other functional sequences [34, 35, 33]. Maynard Smith recognized that this feature was a requirement for evolution by point mutation to be successful. Evolution can step one mutation at a time only if there is a continuous network of functional proteins, otherwise mutation would always lead to lower fitness and evolution would stop [24]. Proteins are in fact robust to mutation — a significant fraction of possible mutants retain their fold and function [36, 37].

Whereas natural evolution can discover new protein functions along circuitous paths that involve many neutral or even slightly deleterious mutations, directed evolution does not have that luxury. Because the possible evolutionary paths grow exponentially as mutations accumulate and there are too many ways to take neutral or deleterious steps that do not ultimately lead uphill, directed evolution is largely constrained to moving continuously uphill in an adaptive walk [38]. This is often not a severe limitation because many interesting proteins are accessible by short and simple adaptive walks. Although the resulting proteins, or even the mutations, might not be the same as those discovered by more convoluted paths to the same fitness level, they nonetheless provide valuable insights into protein function and routes of adaptation.

1.4 Strategies for Directed Evolution

Before we describe some of the key lessons that directed evolution studies have taught us about protein function and evolution, we briefly discuss the experimental strategy. How the experiment is performed obviously influences the outcome and, therefore, the information that is extracted from it. Finding a sequence that performs a desired function in a vast space of possible sequences that is only sparsely populated with functional ones might seem like a daunting task. Inefficient searches of this space could take essentially forever and the task of the protein engineer is to choose a strategy that will reach the objective and do so quickly and easily. Starting with a functional protein, directed evolution uses repeated generations of mutation to create functional variation and selection of the fittest variants to direct the search to higher elevations on the fitness landscape. It involves four key steps (Figure 1.2). First, identifying a good starting sequence; second, mutating this ‘parent’ to create a library of variants; third, identifying variants with improved function; and last, repeating the process until the desired function is achieved. There are many options for the implementation of each step, the choice of which can greatly affect both the efficiency and the endpoint of an evolutionary search.

Directed evolution (and, indeed, natural evolution) relies on the ability of proteins to function over a wider range of environments or carry out a wider range of functions than might be biologically relevant at a given time and therefore selected for. This ability to tolerate a non-natural environment or to exhibit ‘promiscuous’ functions at some minimal level provides the jumping-off point for optimization towards that new goal. A good parent protein for directed evolution, therefore, exhibits enough of the desired function that small improvements (expected from a single mutation) can be reliably discerned in a high-throughput screen [38]. It is also easy to work with and sufficiently stable to accommodate multiple, potentially destabilizing, mutations if the target function is some other property. Some proteins are much more evolvable than others [12, 30, 39, 40]. Possible molecular mechanisms that contribute to evolvability have been discussed, including the key role of the chemical mechanism in enzyme functional evolution [41, 42] and the idea that evolvable proteins exist in multiple closely related but functionally diverse conformations, the distribution of which is easily altered by mutation [9]. These ideas, however, are still largely speculative, and little other than the ability to accept mutations [30, 43] has been conclusively shown in laboratory evolution experiments to contribute directly to allowing one protein to adapt to a new challenge more readily than another protein. A good heuristic indicator of a protein family’s evolvability is its natural functional diversity [40, 44]. Proteins that have adapted to exhibit a range of functions across their family, for example members of an enzyme family that accepts a wide range of substrates (although individual enzymes in the family might be specific) are likely to be adaptable in the laboratory.

The next step is to create a library of variants. As screening is often the most difficult experimental step, the library is usually created to generate the highest probability of finding improved proteins given the screening capability. Because most mutations are deleterious and multiple mutations frequently inactivate proteins (see below), this usually involves a low mutation rate (one or two amino acid substitutions per gene). If screening is not difficult (for example, there is a good genetic selection), then the library can be constructed to generate the largest potential improvement. This might mean a slightly higher mutation rate [45]. In either case, mutations can be introduced randomly [2] or, if structural or mechanistic information is available, they can be made in a more directed manner [46, 47, 48] in an effort to increase the frequency of improved proteins and reduce the load in the next step.

Screening (with high-throughput functional assays) or selection (for example, a genetic selection in which hosts with improved proteins out-compete the others) is used to identify the library members improved in the target property. A good screen or selection accurately assesses the target properties. The rule 'you get what you screen for' is always useful to remember — screening (or selecting) for something else is risky [49]. It is also important not to demand too much improvement in a single generation. The hurdle must be tuned to the screening capacity and should usually be no greater than the improvement that can be provided by a single mutation. If the desired function is beyond what a single mutation can accomplish, the problem can be broken down into a series of smaller ones that can be solved by the accumulation of single mutations, for example by gradually increasing the selection pressure or evolving against a series of intermediate challenges [14]. The process of mutation and selection is repeated until the fitness objective is met; the number of iterations required obviously depends on the starting fitness and the improvement that can be achieved in each round, but is often only five to ten generations.

1.4.1 Mutational steps

An evolutionary search relies on the presence of functional diversity in a population, which is the result of underlying genetic variation. At the molecular level, this genetic variation can take many forms; for example, point mutations, insertions, deletions, recombination, and circular permutation [50, 51, 52]. To search efficiently and minimize the screening load, the underlying genetic variation should be set to generate the highest probability of improvement. Statistically, random mutations tend to be quite harsh, usually decreasing activity and sometimes destroying it altogether. Typically, 30–50% of single amino acid mutations are strongly deleterious, 50–70% are neutral or slightly deleterious and 0.01–1% are beneficial [12, 30, 37, 53, 54, 55, 56]. If the fitness landscape is Fujiyama-like with many smooth uphill paths, only beneficial mutations need to accumulate (either in multiple rounds of mutagenesis and screening or by recombining beneficial mutations found in each round

[57, 58]) until the desired fitness is reached. In a single-peaked landscape, all beneficial mutations make a cumulative contribution to the desired function and all paths uphill eventually converge to the same optimal solution.

Of course, no real protein landscape consists of a single peak. Most mutations are deleterious and therefore most paths end downhill, with inactive proteins, rather than uphill at fitter sequences. Furthermore, epistatic interactions occur when the presence of one mutation affects the contribution of another mutation. Such epistatic interactions lead to curves in the fitness landscape and constrain evolutionary searches. Extreme forms of epistasis, in which mutations that are negative in one context become beneficial in another (so-called sign epistasis [59]), create local optima on the landscape that can frustrate evolutionary optimization. Epistatic interactions are a ubiquitous feature of protein fitness landscapes [60, 61]. We argue, however, that they are not important for most optimizations by directed evolution, which instead follow one of many smooth paths that bypass the more rugged, epistatic routes on this high-dimensional surface [62, 63, 64]. Among the numerous mutational trajectories between a starting point and a solution, smooth uphill paths can often be found (Figure 1.1*C*).

1.4.2 Dealing with the combinatorial explosion

Knowing of epistatic interactions and local fitness optima, some protein engineers worry about the need to make and find multiple mutations at one time. If multiple mutations are needed to climb the peak, the combinatorial explosion of mutational possibilities makes them especially challenging to find. For even a small protein of 100 amino acids, there are 1,900 single amino acid mutants and more than 1.5 million double mutants. The number of possible sequences increases exponentially with the number of mutations and a complete sampling of even just the double mutants is beyond the capacity of most screens. Higher-throughput screening approaches have been developed to enable sampling of more mutants and more combinations of mutations [4, 65, 66]. These screens can allow multiple paths to be explored simultaneously, increasing the probability of discovering good adaptive routes to higher fitness. However, higher-throughput screens or selections usually come at the cost of decreased accuracy, especially when a surrogate function that is more amenable to high-throughput measurement is substituted for the desired function. Furthermore, increasing the mutation rate to capture rare synergistic mutations can make it more difficult to identify improved single-mutation variants because common deleterious mutations will tend to mask the rare beneficial ones. It is often better, therefore, to focus on sampling single mutants with a higher quality, lower-throughput screen rather than on increasing the throughput to capture multiple simultaneous mutations. Although a search through single adaptive steps cannot find mutations exhibiting negative epistasis, there are usually other, step-wise adaptive routes to the objective.

The high dimensionality of sequence space that makes finding simultaneous beneficial mutations so difficult can be reduced by taking advantage of structural, functional or phylogenetic information to focus mutations to those residues that are most likely to lead to the desired properties. For example, the modularity of protein structures permits the separate optimization of protein domains [14, 67]. Phylogenetic analyses suggest that nature might separately optimize other, structurally non-obvious subunits, or 'sectors' [68], which could prove to be appropriate targets for directed evolution. The search space can also be reduced by focusing mutations to specific residues in a domain; for example, in an active site or binding pocket in which functional changes might be more likely to occur [12, 46, 69, 70, 71]. This strategy only works, however, when the experimenter is able to select the right residue combinations for random mutagenesis, leaving out the possibility of finding surprising and informative solutions elsewhere. Numerous studies have shown, for example, that many activating mutations lie outside enzyme catalytic sites and exert their influence through mechanisms that might not be obvious from structural analysis [10, 11, 72].

1.4.3 Alternative search strategies

Evolution by the accumulation of single mutations has proven to be very effective at optimizing a function or property that already exists or can be reached through a series of intermediate steps. Some functions, however, simply can not be reached through a series of small uphill steps and instead require longer jumps that include mutations that would be neutral or even deleterious when made individually. Examples of functions that might require multiple simultaneous mutations include the appearance of a new catalytic activity or an activity on a substrate for which the parent and its single mutants show no measurable activity. Because most mutations are deleterious, the probability that a variant retains its fold and function declines exponentially with the number of random substitutions [36, 37], and random jumps in sequence space uncover mostly inactive proteins. Thus, new functions are extremely difficult to obtain without altering some aspect of the search.

One approach is to create a new starting point – a parent protein with at least some minimal function and improve that by directed evolution [8]. Where natural examples of a desired function are not practical or might not even exist, emerging protein design tools have identified functional sequences [6]. Expanding the sequence space by the incorporation of non-natural amino acids can also introduce a whole array of new functions and directed evolution can do the fine-tuning that might be needed to optimize these novel designs [16]. Another approach is to find more conservative ways to make multiple mutations; for example, using computational protein design tools to identify sets of mutations that are likely to be compatible with structure retention [47].

An approach to making multiple mutations that is used extensively in nature is recombination. Naturally-occurring homologous proteins can be recombined to create genetic diversity within protein sequence libraries [73, 74, 75] (Figure 1.3*A*). It has been shown that mutations made by recombination are much less disruptive and generate functional proteins with much higher frequency than random mutations [56] (Figure 1.3*B*). Recombination methods based on DNA sequence hybridization direct crossovers to regions of high sequence identity and are generally limited to sequences that are very similar (with more than 70% identity) [75], whereas various sequence-independent methods can recombine at random [76, 77] or user-specified sites [78, 79]. Recombining homologous proteins by choosing crossovers based on structural information allows the construction of libraries of chimeric proteins that simultaneously exhibit high levels of functionality and genetic diversity [80]. In all cases, the chimeric proteins inherit the best (and worst) residues the parents have to offer, in new combinations that are not observed in nature.

Chimeric proteins can differ by tens or even hundreds of mutations from their parent sequences and still function. The conservative nature of recombination can be exploited to make whole families of novel enzymes. For example, in one set of more than 6,000 chimeric cytochrome P450 proteins with an average of 70 mutations from the closest parent, approximately half folded properly, and at least 75% of these folded P450 proteins displayed enzymatic activity [80].

The new combinations of residues can give rise to novel properties [81]. Because many of the mutations made by recombination are neutral or nearly neutral, recombination is an efficient way to generate the neutral drifts (the accumulation of neutral mutations) that have been shown to lead to increases in promiscuous functions [82, 83] and mutational robustness [84, 82]. For example, members of the chimeric cytochrome P450 library exhibited higher enzymatic activity than any of the three parents across a panel of 11 non-native substrates that included substrates on which the parent enzymes showed no measurable activity [85]. Several P450 chimeras were also more thermostable than the most thermostable parent enzyme, and dozens of thermostable chimeras could be readily identified based on a small sampling of the library [86] (Figure 1.3*C*). This approach was subsequently used to generate dozens of highly stable, highly active fungal cellobiohydrolase II enzymes that degrade cellulose into fermentable sugars (for biofuels applications, for example) [79].

1.5 Lessons from Directed Evolution

In addition to generating a plethora of novel proteins, directed evolution studies have elucidated available pathways and molecular mechanisms of adaptation, shown a key role for stability in epistasis and evolvability, identified important evolutionary trade-offs in protein properties, and revealed the simultaneously conservative and exploratory nature of recombination, all of which have shed light on long-standing questions in protein chemistry and evolutionary biology. First and foremost, directed evolution experiments have shown time and again how rapidly proteins can adapt to exhibit new functions and properties. Protein behavior can change dramatically on mutating a very small fraction of the protein sequence. Directed evolution also provides a detailed view of the adaptive process.

A directed evolution approach to studying sequence-function relationships circumvents several challenges associated with inferring mechanisms of adaptation using comparisons of evolutionarily related natural amino acid sequences [22, 23]. Such studies are confounded by the numerous, mostly neutral mutations that accumulated during divergence of the sequences and the complex and largely unknown selection pressures under which the natural sequences evolved. By contrast, the sequences generated by directed evolution contain a small number of adaptive mutations that accumulated under well-defined selective pressures. Furthermore, performing the evolution in the laboratory permits access to the full 'fossil record' of evolutionary intermediates, the sequences, structures, and functions of which can be analyzed in an attempt to explain how new properties were acquired [11, 44, 72, 87]. Fasan and co-workers analyzed selected intermediates that arose during the directed evolution of a cytochrome P450 fatty acid hydroxylase into a highly efficient and highly specific propane monooxygenase [14, 72] (Figure 1.4). The gradual increase in activity on propane (as measured by total turnovers of propane to propanol — the property targeted during directed evolution) was accompanied by other interesting changes in the enzyme's behavior, the most notable of which was the decrease in thermostability (as measured by T_{50}; the temperature at which 50% of the proteins are inactivated in 10 minutes). Activating mutations came at the cost of thermostability, to the point that it became necessary to incorporate stabilizing mutations (generation nine in Figure 1.4) before further increases in activity could appear. This apparent trade-off between functionally beneficial mutations and thermostability reflects the fact that most mutations are destabilizing and therefore most activating mutations are also destabilizing. Because evolution favors the most likely solutions over rarer ones, it favors marginal stability in the absence of selection for higher stability. It also favors properties that are compatible with marginal stability [33]. Such trade-offs have also been shown to constrain the evolution of antibiotic resistance enzymes [88] and will be discussed further below.

The mutations that accumulated in the heme domain of the cytochrome P450 are depicted in Figure 1.4*B* and are color-coded according to the generation in which they appeared. Many of the mutations that conferred the increased activity on propane lie outside the substrate-binding pocket, where they influence substrate recognition through mechanisms that are difficult to discern from crystal structures or modeling. That the effects of the adaptive mutations are difficult to rationalize, much less predict, underscores how little we understand of how sequence determines protein structure and function. Directed evolution deals with the details of molecular interactions, and it is hoped that these details will eventually help protein design efforts [8].

Directed evolution can explore alternative evolutionary scenarios; for example, to identify other possible solutions to the same functional challenge or to address whether multiple paths can lead to the same solution, as was done with a laboratory-evolved β-lactamase variant that contains 5 mutations responsible for a 100,000-fold increase in cefotaxime resistance [63]. In this study, the authors constructed variants with all 32 (2^5) combinations of the adaptive mutations, representing all intermediate sequences along all 120 (5!) possible mutational pathways. They were able to estimate the probability of each pathway based on the relative change in antibiotic resistance conferred to the bacteria by each mutation along each path. Whereas most of the possible paths were constrained by epistasis and were therefore highly unlikely, there were 18 different, simple uphill walks to the final solution.

1.5.1 Empirical landscapes

Even the earliest directed evolution experiments noted how rapidly proteins could adapt to new selective pressures [2, 58], indicating the ready availability of smooth uphill paths in the fitness landscapes. Stability, the ability to tolerate new environments and low-level side reactions or promiscuous functions usually respond well to directed evolution. One study used a well-controlled set of experiments to select for six different promiscuous activities starting from three different enzymes [12]. After two rounds of directed evolution, yielding just 14 mutations, the promiscuous enzyme activities (k_{cat}/K_M) had increased by up to 150-fold over the activities of the parent enzymes. Interestingly, these newly evolved activities came at little cost to the native enzymatic activities, suggesting a particular robustness of the native functions to mutation and supporting a scenario for evolution of new activities that allows both the native and novel activities to be displayed in the same gene for some period of time [9].

As well as demonstrating the availability of smooth uphill paths, directed evolution has provided insight into the molecular epistasis that curves the landscapes. Several studies have revealed a key link between stability and epistasis, where the effect of a mutation can be conditional on the stability of the parent sequence [36, 43, 9] (Figure 1.5). This was demonstrated, for example, in a study of cephalosporin antibiotic resistance mutations in β-lactamase, in which the fitness effects of several active site mutations were found to depend on the presence of a stabilizing M182T mutation [88] (Figure 1.5*A*). These epistatic interactions are the result of catalytically beneficial but destabilizing mutations in the active site that cannot be tolerated unless the stabilizing M182T mutation is present. Without M182T, the active site mutations destabilize the enzyme to the point that total activity is compromised.

Many examples of stability-mediated epistasis are best explained in terms of a protein stability threshold, whereby stability is under selection only insofar as it allows a protein to fold and function [36, 43, 82] (Figure 1.5). The consequences for evolution are profound: a protein with low stability cannot accept more than a small fraction of the possible mutations because most mutations are destabilizing. Thus, it can become trapped on a local optimum, unable to go further. As illustrated in Figure 1.5*B*, proteins enjoying a larger margin above the minimal stability threshold can explore many more mutations and can therefore continue to adapt to other tasks, such as acquiring activity towards a new substrate or partner [30]. Stability-mediated epistasis is a mechanism whereby neutral mutations can shape the available adaptive pathways during natural evolution as well as in the laboratory. Experience has shown that when an evolutionary search in the laboratory seems to have exhausted all options for further uphill steps, the incorporation of stabilizing mutations is able to open up new adaptive routes [14].

Despite being performed on different protein folds with selection for different protein functions, the repeated evaluation of thousands of random mutations has revealed the general features of protein fitness landscapes. In addition to the uphill paths that lie alongside numerous less favorable, epistatic routes there are an even larger number of side-steps in the protein fitness landscape. The high frequency of neutral mutations observed during evaluation of random mutant libraries suggests a myriad of sequences with essentially equivalent fitness. This is consistent with the existence of natural protein homologs that differ at several positions, the majority of which are functionally neutral. Even sequences that are highly optimized are probably just one of many potential solutions to a given functional challenge. Indeed, it is probably more accurate to imagine protein evolution occurring on neutral networks, rather than on fitness landscapes in which each neighbor has a different fitness [29, 62]. This pervasive neutrality is exploited when families of functional proteins are constructed by recombination of homologous proteins [79, 80].

As discussed above in the context of stability-mediated epistasis, mutations that are neutral in one context might not be neutral in all and therefore can provide new opportunities for evolution. Directed evolution has shown an important role for stabilizing mutations (which can be functionally neutral or only slightly deleterious) in adaptation. Laboratory evolution experiments have also shown that purposefully accumulated neutral mutations alter promiscuous activities and create new starting points for subsequent adaptive evolution [82, 83, 89]. Genetic drift and pre-existing diversity might have a similarly important role in natural adaptive evolution [62].

1.5.2 Directed evolution to understand natural evolution?

An overall picture of the protein function landscape is therefore emerging from accumulating directed evolution data. This picture offers a description of the physical features that all proteins (synthetic or natural) must exhibit and the effects of mutations on these features. Extending the lessons learned from directed evolution to natural evolution, however, requires caution because these search processes operate under, for example, different time scales, population sizes, mutation rates, and strength of selection. Furthermore, natural evolution works on a different fitness landscape and it is unclear how the protein fitness assayed during directed evolution is related to the organismal fitness that natural evolution optimizes. Differences reflect the consequences of interactions between the protein and the cellular environment and might include constraints related to metabolic burden, regulation, non-specific interactions, and other factors.

The ability to disconnect a protein from its *in vivo* function is a valuable asset of directed evolution because it allows the exploration of physically possible proteins without the often-severe constraint of their being biologically relevant and contributing to organismal fitness. Thus, directed evolution can be used to identify which features of proteins are dictated by their physical properties versus those that are due to biological constraints or evolutionary origins and history. The laboratory evolution of the cytochrome P450 propane monooxygenase (Figure 1.4), for example, showed the physical possibility, and indeed the ready availability, of such an enzyme, even though known organisms that live on small alkanes use mechanistically and evolutionarily unrelated enzymes for this transformation [72]. Another example is the generation of proteins with combinations of properties that are usually not found in natural proteins, such as high catalytic activity at low temperature and high stability at elevated temperature [22, 23]. When properties seem to trade off like this, it might be tempting to infer that such trade-offs are dictated by physical requirements, such as the incompatibility between molecular rigidity that is needed for high stability and the flexibility that is required for catalytic activity [90, 91]. If stability and enzymatic activity placed mutually exclusive demands on protein flexibility, then highly active, highly stable enzymes could not exist (a statement that protein engineers did not want to hear). Directed evolution, however, has little trouble finding enzymes that are both highly active and highly stable when the experiments select for both properties [92]. Clearly, such proteins are far rarer than highly active, marginally stable proteins and, without a good reason, natural sequences would not exhibit both features [22, 23, 33, 93].

1.6 Conclusions

Despite the vast size of sequence space and the complex nature of protein function, the Darwinian algorithm of mutation and selection provides a powerful method to generate proteins with altered functions. This simple uphill walk on a fitness landscape in sequence space works because proteins are wonderfully evolvable and can adapt to new conditions or even take on new functions with only a few mutations.

In addition to providing useful proteins, directed evolution experiments have also taught us how proteins adapt and shed light on processes at work during natural evolution [22, 62, 94]. These experimental results allow us to look at sequence data in a functional context, providing a bridge between long-separated fields of evolutionary and molecular biology [95]. Directed evolution experiments have been used to address important evolutionary questions about the average effects of mutations, mechanisms of functional divergence, evolvability, and evolutionary constraints [12, 82, 93, 96].

With the growing number of applications for engineered proteins, directed evolution will continue to be an important strategy for making proteins that are well adapted to new environments and new functions. More advanced high-throughput screens and higher quality sequence libraries will make the searches easier and will enable evolution to solve increasingly complex problems. Advances in our understanding of proteins can be incorporated into library design, and the rapidly decreasing cost of DNA synthesis will relieve many sequence construction constraints. Directed evolution will help teach us how biological systems adapt to changing demands; it might also help us to address some of today's most challenging problems of providing effective treatments for disease or producing fuels and chemicals from renewable resources.

1.7 Figures

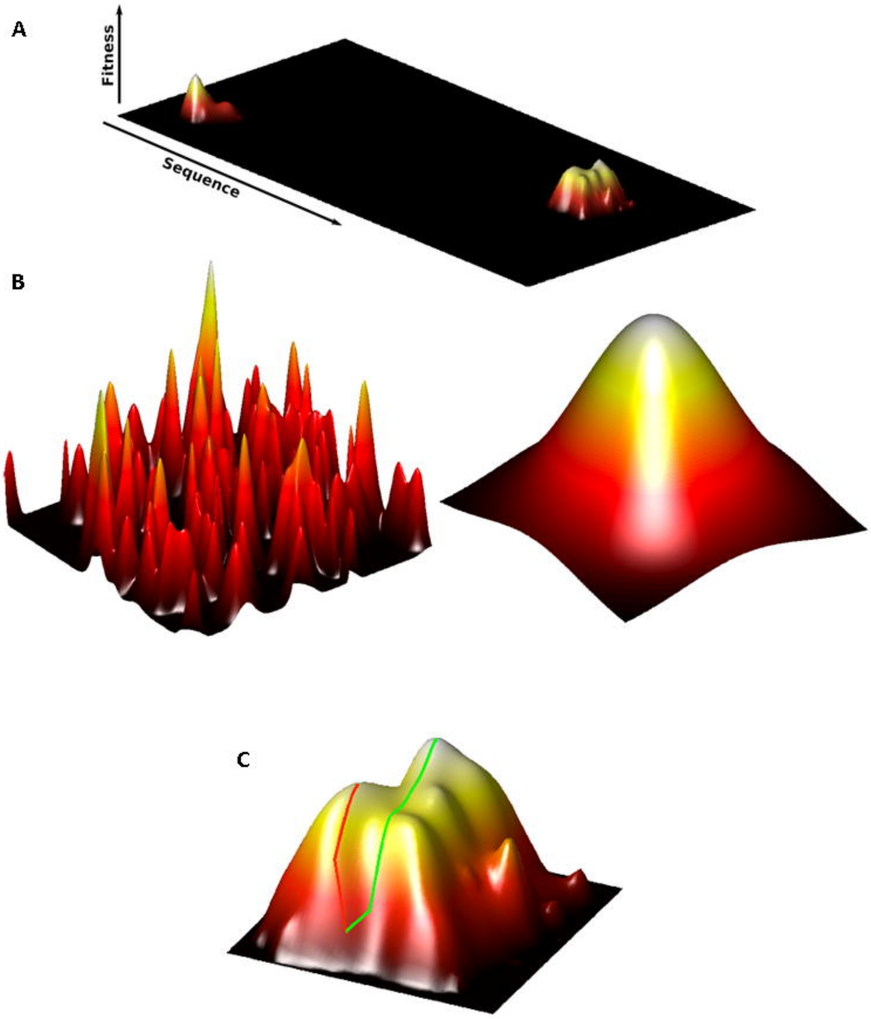


Figure 1.1: Directed protein evolution traverses a fitness landscape in sequence space. This fitness is the measure of how well a given protein performs a target function. (*A*) The plot of fitness against sequence creates the landscape for evolution. The transition through black-red-orange-yellow represents increasing fitness. Although the details of this landscape are unknown, it is believed that most sequences do not function (black) and that the rare functional sequences encoding natural proteins are clustered near other functional sequences. However, this popular three-dimensional representation does a poor job of illustrating the numerous paths available to evolution and the numerous sequences in functional regions that do not encode functional proteins [97]. (*B*) Similar to natural protein evolution, directed evolution moves along networks of functional proteins that differ by a single amino acid, because selection requires a continuous uphill walk and does not permit the fixation of non-functional sequences. Epistasis occurs when the effect of one mutation depends on the presence of another, which can create landscape ruggedness and local optima. Landscapes could range from the rugged 'Badlands' landscape (left panel), which is nearly impossible to climb by mutational steps, to the 'Fujiyama' landscape (right panel), in which any beneficial mutation brings the search closer to the optimum [28]. (*C*) The presence of local optima might restrict some of the mutational paths uphill (red line). However, the large number of alternative routes leaves plenty of adaptive paths to a fitness optimum (green line).

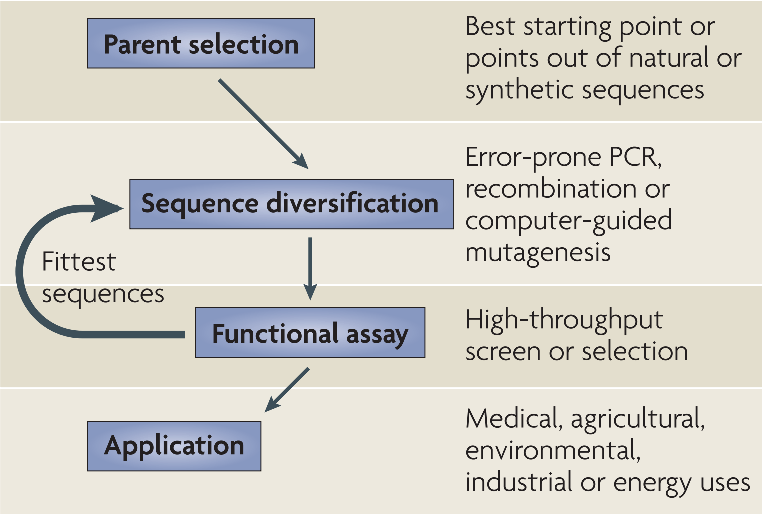

Parent selection
Best starting point or points out of natural or synthetic sequences
Sequence diversification
Error-prone PCR, recombination or computer-guided mutagenesis
Fittest sequences
Functional assay
High-throughput screen or selection
Application
Medical, agricultural, environmental, industrial or energy uses

Figure 1.2: The objective of directed evolution is to create a specific protein function through successive rounds of mutation and selection, starting from a parent protein with a related function. There are numerous options for implementing each step in the process, the choice of which can greatly affect the efficiency and success of the protein sequence optimization. A parent sequence (or sequences) is chosen based on its perceived proximity to the desired function and its evolvability. This parent sequence is then mutated to form a library of new sequences (error-prone PCR or other methods can be used to incorporate mutations randomly, recombination can be used to introduce mutations from other functional sequences or mutation sites can be chosen based on functional and/or structural information). These mutated sequences are evaluated for their ability to perform the desired function using a high-throughput screen or artificial selection. The fittest sequence (or sequences) is used as the parent for the next round of directed evolution, and this process is repeated until the engineering objective is met (usually after five to ten generations).

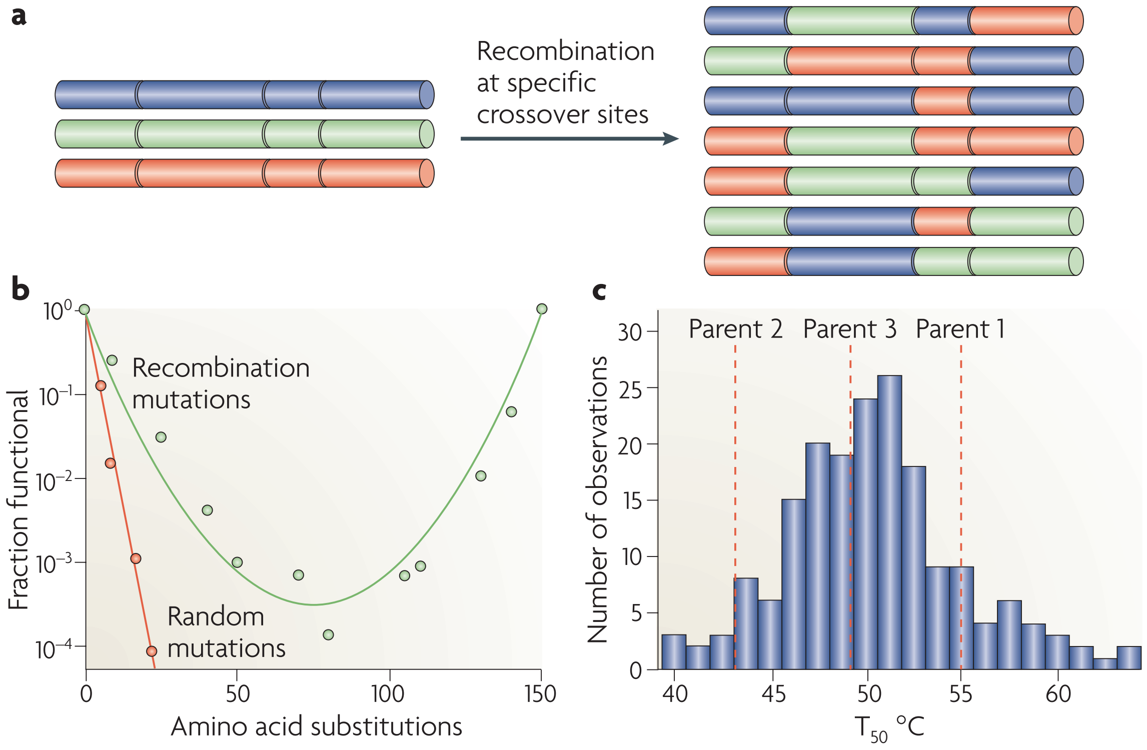
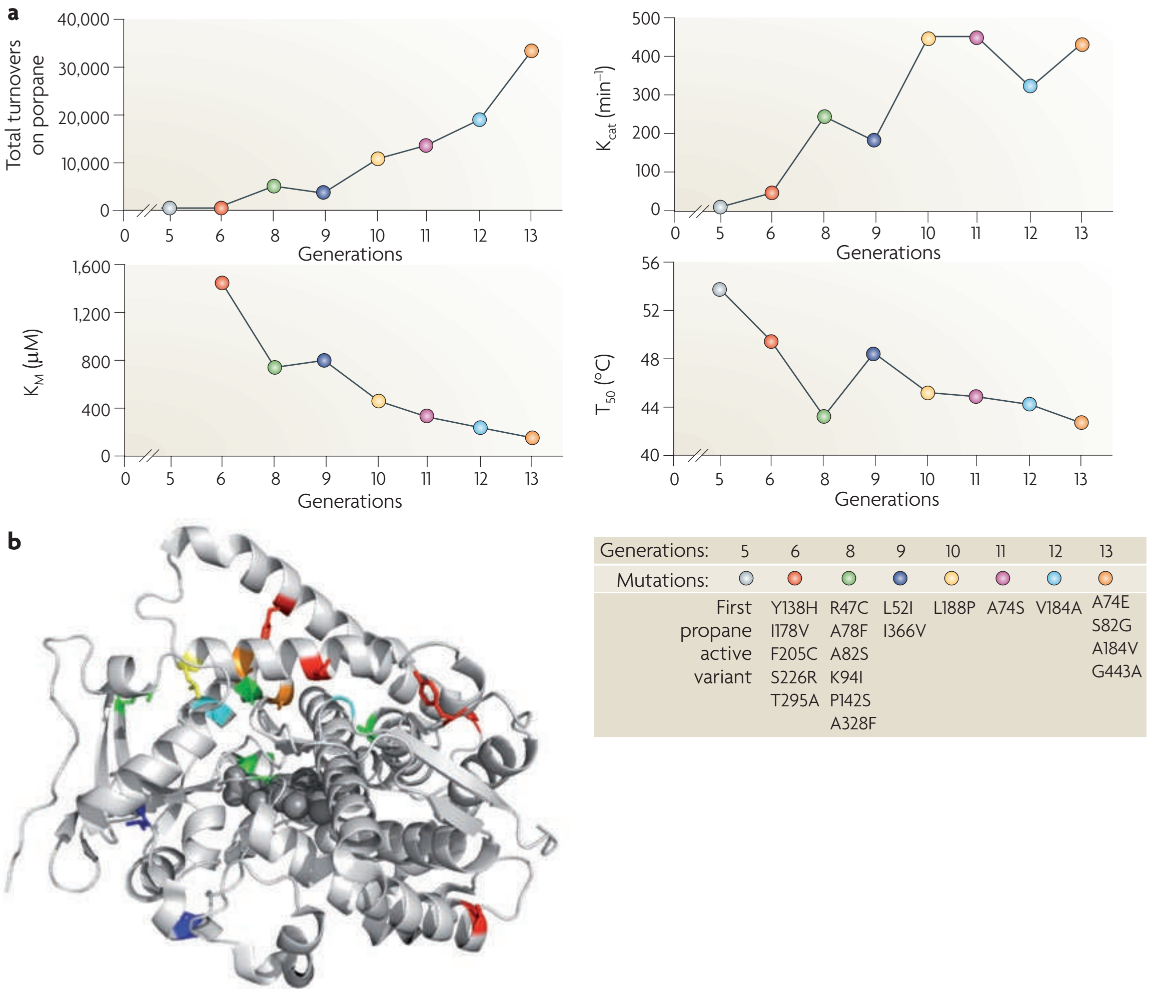


Figure 1.3: (A) Recombination generates highly mutated sequence libraries. Multiple homologous parent sequences are divided into fragments, which can be chosen to minimize structural disruption [73], and these fragments are recombined to form a combinatorial library of chimeric proteins. (B) The mutations from homologous recombination (green) are much more conservative than random mutations (red). In β-lactamase, chimeras with high levels of amino acid mutations (around 75) are 10^{16} times more likely to fold than sequences with 75 random mutations [56]. (C) Chimeric proteins contain new combinations of beneficial mutations. The histogram shows the distribution of thermostabilities (T_{50}; the temperature at which 50% of the proteins are inactivated in 10 minutes) of 184 randomly selected chimeric cytochrome P450 enzymes made by structure-guided recombination. The thermostabilities of the three parents are marked by dashed red lines [86]. A significant fraction of chimeras are more thermostable than any parent from which they are derived.



Generations:	5	6	8	9	10	11	12	13
Mutations:	●	●	●	●	●	●	●	●
	First propane active variant	Y138H I178V F205C S226R T295A	R47C A78F A82S K94I P142S A328F	L52I I366V	L188P	A74S	V184A	A74E S82G A184V G443A

Figure 1.4: Cytochrome P450 BM3 from *Bacillus megaterium* catalyses the hydroxylation of long-chain fatty acids and has no measurable activity on propane. This enzyme was converted into a highly efficient and specific propane monooxygenase over 13 rounds of directed evolution [14, 98, 99]. The large change in substrate specificity was achieved using an incremental approach that first involved screening on an intermediate substrate. Because the native substrate contains a long alkyl chain and the target function was activity on a short alkane, an intermediate-length alkane (octane), towards which the parent enzyme had low but measurable activity, was chosen as the initial directed evolution target. Once high octane activity was achieved, the selective pressure was switched towards activity on propane. (*A*) Selected kinetic and biophysical properties of evolutionary intermediates from later generations (with generation five being the first propane active variant) [72]. Total catalytic turnovers (moles of propanol produced per mole of P450), K_M and k_{cat} are reported for propane hydroxylation. Thermostability is shown as T_{50} (the temperature at which half of the enzyme inactivates after a 10 minute incubation). Variants were selected for total propane activity in all generations, except for generation nine, which was selected for T_{50}. The mutations acquired during each generation are listed. Even small numbers of mutations can be responsible for large functional changes. (*B*) The crystal structure of the fifth generation P450 heme domain (Protein Data Bank identifier: 3CBD), with the locations of the mutations from subsequent generations color-coded as in part A. Beneficial mutations are distributed over the heme domain and many are tens of Å from the catalytic iron.

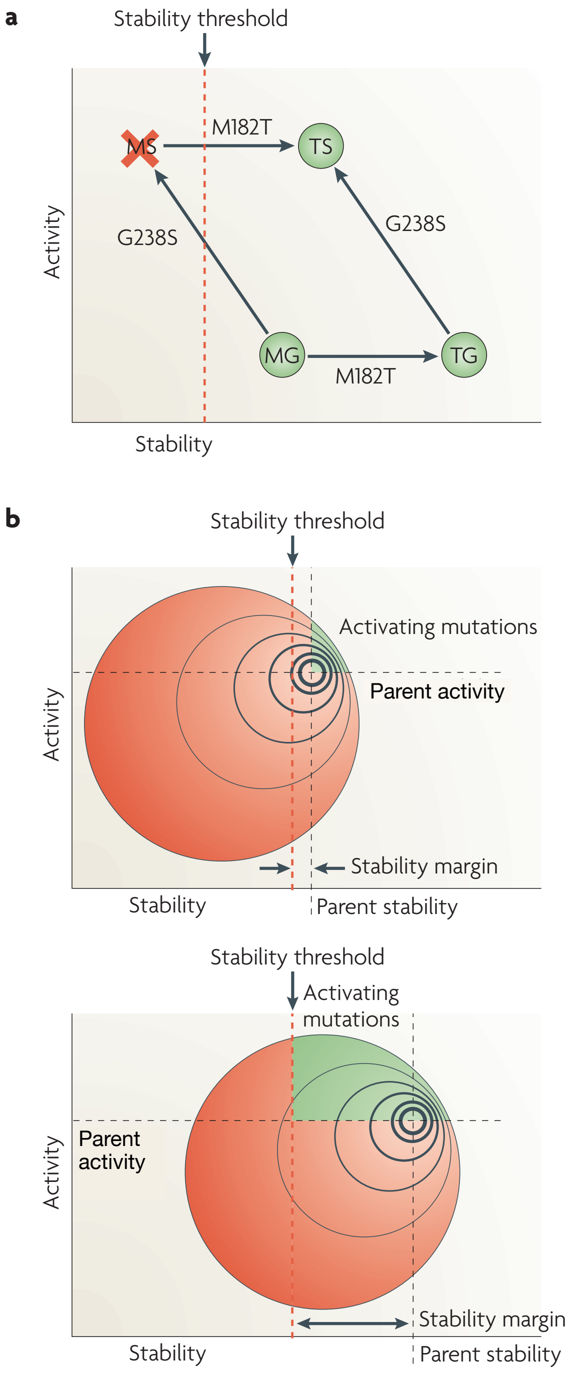


Figure 1.5: Laboratory evolution studies have found many examples of mutational epistasis that are related to protein stability. The relationship between protein stability and epistasis is best explained in terms of a protein stability threshold, whereby stability is under selection only insofar as it allows a protein to fold and function [36, 43, 82]. (*A*) Epistasis can arise as the result of the protein stability threshold. The G238S active-site mutation in this β-lactamase increases enzyme activity on cephalosporin antibiotics [88]. However, this mutation cannot be accepted into the wild-type sequence (MG) because the resulting protein (MS) is not sufficiently stable. Sequences with the beneficial G238S mutation can instead be reached by first finding the functionally neutral, but stabilizing, M182T mutation (sequence TG) and then incorporating the G238S mutation (sequence TS). (*B*) Because most mutations are destabilizing, many of the single mutants of a protein close to the stability threshold (top panel) will be unstable and therefore inactive (red). This leaves few active mutants having beneficial mutations (green). A more stable protein (bottom panel) will be more tolerant to mutation, making more beneficial mutations available

Chapter 2

Directed evolution of a magnetic resonance imaging contrast agent for noninvasive imaging of dopamine

A version of this chapter has been published as [100].

2.1 Abstract

The development of molecular probes that allow *in vivo* imaging of neural signaling processes with high temporal and spatial resolution remains challenging. Here we applied directed evolution techniques to create magnetic resonance imaging (MRI) contrast agents sensitive to the neurotransmitter dopamine. The sensors were derived from the heme domain of the bacterial cytochrome P450-BM3 (BM3h). Ligand binding to a site near BM3h's paramagnetic heme iron led to a drop in MRI signal enhancement and a shift in optical absorbance. Using an absorbance-based screen, we evolved the specificity of BM3h away from its natural ligand and toward dopamine, producing sensors with dissociation constants for dopamine of 3.3–8.9 μM. These molecules were used to image depolarization-triggered neurotransmitter release from PC12 cells and in the brains of live animals. Our results demonstrate the feasibility of molecular-level functional MRI using neural activity-dependent sensors, and our protein engineering approach can be generalized to create probes for other targets.

2.2 Introduction

MRI is a uniquely valuable tool for studying the brain because MRI scans are noninvasive and can provide information at relatively high spatial resolution ($< 100\ \mu$m) and temporal resolution ($\sim$1 s) from living specimens. Functional imaging (fMRI) of brain activity is possible with MRI methods sensitive to cerebral hemodynamics [101]. The most common fMRI technique, blood-oxygen-level-dependent (BOLD) fMRI, is based on oxygenation of hemoglobin, an endogenous oxygen-sensitive MRI contrast agent present in the blood [102]. Although BOLD fMRI has had a tremendous impact in neuroscience, the method provides only a slow and indirect readout of neural activity, owing to the complexity of neurovascular coupling [103]. Considerably more precise measurements of brain function would be possible with MRI sensors that were directly and rapidly responsive to neurochemicals involved in the brain's information processing [104].

The challenging process of developing sensors for next-generation neuroimaging could be greatly accelerated using advanced molecular engineering techniques. Directed evolution is a molecular engineering method that employs successive rounds of mutagenesis and selection to generate proteins with novel functionality, starting from a molecule with some of the desired properties of the end product [36]. This technique could be applied to evolve MRI sensors from proteins that are magnetically active (for example, paramagnetic) and have tunable ligand-binding or catalytic properties.

The flavocytochrome P450-BM3 (BM3), a fatty acid hydroxylase from *Bacillus megaterium*, contains a paramagnetic iron atom embedded in a solvent-accessible substrate-binding pocket, suggesting that it could produce ligand-dependent MRI signal changes. BM3's binding specificity is also highly tunable, as demonstrated by previous efforts to identify novel enzymatic activities through directed evolution of this protein [105, 98, 106, 80]. If BM3 variants could be engineered to act as MRI sensors, they would be genetically encodable, an added advantage over synthetic molecular imaging agents.

We sought to apply directed evolution of BM3 to develop MRI sensors for a key signaling molecule in the brain, the neurotransmitter dopamine. To our knowledge, no MRI contrast agent for sensing dopamine (or any other neurotransmitter) currently exists, but there is considerable interest in measuring dopamine-related activity by MRI [107]. Dopamine is of particular significance because of its roles in learning, reward, and motor coordination [108], and because the dysfunction of dopaminergic systems underlies addiction [109] and several neurodegenerative diseases [110]. Existing techniques for measuring dopamine *in vivo* are either invasive point-measurement methods [111, 112, 113] or positron emission tomography procedures [114] with low spatial and temporal resolution. MRI could be used successfully for dopamine measurement if combined with an imaging agent capable of responding quickly, reversibly and specifically to extracellular dopamine fluctuations from < 1 μM to tens of micromolar [115, 116]. To be comparable with established functional brain imaging techniques, interaction of dopamine with the probe should also produce image signal changes on the order of 1% or more *in vivo* [117]. Here we show that directed evolution of BM3 is capable of producing dopamine sensors that largely meet these specifications.

2.3 Results

2.3.1 P450 BM3 reports ligand binding in MRI

To evolve dopamine probes for MRI, we focused on the heme domain of BM3 (BM3h), a 53 kDa moiety that is catalytically inactive in the absence of the full protein's reductase domain [118]. BM3h contains a single iron(III) atom (mixed spin 1/2 and 5/2) [119] bound to a hemin prosthetic group and axially coordinated by residue Cys400 on the protein. In the absence of substrates, the remaining coordination site is filled by a water molecule [120]. Interaction of the heme iron with exchanging water molecules at this axial site promotes T_1 relaxation in aqueous solutions [121] and is therefore predicted to modulate MRI contrast. To determine the extent of this effect, we used a spin echo pulse sequence in a 4.7 T MRI scanner to measure the proton relaxation rate as a function of protein concentration in PBS; the slope of this relationship (T_1 relaxivity, or r_1) provides a standard measure of the strength of a contrast agent. For BM3h in the absence of ligands, an r_1 value of 1.23 ± 0.07 mM^{-1}s^{-1} was obtained. Addition of a saturating quantity of the natural BM3 substrate, arachidonic acid (400 μM concentration), resulted in an r_1 of 0.42 ± 0.05 mM^{-1}s^{-1}

(Figure 2.1*A*). This ligand-induced decrease in relaxivity, probably arising from the displacement of water molecules at the BM3h heme, enabled quantitative sensing of arachidonic acid using MRI (Figure 2.1*B*) and suggested that BM3h could serve as a platform for molecular sensor engineering.

We next tested whether dopamine or related compounds could serve as unnatural ligands to BM3h when applied at high enough concentrations. As measured by MRI, addition of 1 mM dopamine to BM3h in fact induced a drop in r_1 to 0.76 ± 0.03 mM^{-1}s^{-1} (Figure 2.1*A*). Binding of arachidonic acid is known to induce a change (blue shift) in BM3h's optical absorbance spectrum because of perturbation of the electronic environment of the heme chromophore [122]. To determine whether the relaxation change induced by dopamine also reflects interaction with the BM3h heme, we measured optical spectra of the protein in the presence and absence of 1 mM dopamine. The interaction produced a small but clearly discernable red shift of λ_{max}, from 419 to 422 nm (Figure 2.1*C*), indicative of ligand coordination to the heme iron [122]. This suggests that dopamine (at 1 mM) directly replaces water as an axial metal ligand in the BM3h substrate-binding pocket and that directed evolution of BM3h binding specificity could therefore improve the protein's relative affinity for dopamine. In addition to providing mechanistic insight, the correspondence between optical and MRI measurements of ligand binding to BM3h implied that either modality could be used to obtain quantitative binding parameters. We monitored the difference between absorption at two wavelengths as a function of ligand concentration to determine binding isotherms for arachidonic acid and dopamine (Figure 2.1*D,E*). For BM3h, the apparent K_d for arachidonic acid was 6.8 ± 0.5 μM; the K_d for dopamine was 990 ± 110 μM. Goals for the production of BM3h-based MRI sensors thus included decreasing the affinity for arachidonic acid, increasing the dopamine affinity by at least two orders of magnitude and maintaining or enhancing the relaxivity changes observed upon ligand binding.

2.3.2 Directed evolution of dopamine-responsive BM3h variants

To create an MRI sensor for dopamine using directed evolution, we developed a customized screening methodology (Figure 2.2*A*). Results shown in Figure 2.1 suggested that either MRI-based or optical assays could be used to distinguish BM3h mutants with differing ligand affinities. We chose an absorbance assay for our screen because lower protein concentrations (∼1 μM) could be used in this format. Input to each round of screening consisted of a library of BM3h mutants, each with an average of one to two amino acid substitutions, generated by error-prone PCR from the wild-type (WT) gene or a previously selected mutant. We transformed DNA libraries into *Escherichia coli*. We grew and induced approximately 900 randomly selected clones in microtiter plate format, then prepared cleared lysates for optical titration with dopamine and arachidonic acid in a plate reader. Titration data were analyzed to determine K_d values for both ligands. An average of 79% of assayed mutants had sufficient protein levels (absorbance signal $>$ 30% of parent) and clean enough titration curves ($R^2 > 0.8$) for K_d estimation. Mutant affinities appeared to be distributed randomly about the dissociation constant measured for the corresponding parent protein, but we were able to identify individual clones with desired affinity changes in each round (Figure 2.2*B*). From each screen, we chose eight to ten mutants on the basis of their estimated K_ds, purified them in bulk, re-titrated them to obtain more accurate estimates of their dopamine and arachidonic acid affinities, and examined them with MRI to ensure that robust ligand-induced changes in r_1 could be detected. On the basis of these assays, we chose as a parent for the next round of evolution the mutant showing the best combination of relaxivity changes, improved dopamine affinity and decreased affinity for arachidonic acid.

After carrying out the screening strategy over multiple rounds, we found a steady trend in the distribution of K_d values toward greater affinity for dopamine and less affinity for arachidonic acid (Figure 2.2*B*–*D*). Little change in binding cooperativity was observed, and changes in partial saturation generally occurred over 100-fold ranges of dopamine concentrations. Five rounds of evolution yielded a BM3h variant with eight mutations (Figure 2.2*E*), four near the ligand-binding pocket and four at distal surfaces of the protein. One residue (Ile263) was first mutated to threonine (third round), then to alanine (fourth round). The clones selected from rounds 1, 3, and 5 had two new mutations each. We did not determine the individual contributions of these mutations to the observed changes in affinity. We introduced the mutation I366V by site-directed mutagenesis before the fifth round to enhance thermostability and tolerance of BM3h to further mutation [14, 30]; it did not noticeably affect dopamine binding affinity.

The mutant proteins selected after the fourth and fifth rounds of evolution, denoted BM3h-8C8 and BM3h-B7, had optically determined dissociation constants of 8.9 ± 0.7 μM and 3.3 ± 0.1 μM, respectively, for dopamine, and 750 ± 140 μM and 660 ± 80 μM, respectively, for arachidonic acid. The T_1 relaxivity of BM3h-8C8 was 1.1 ± 0.1 mM^{-1}s^{-1} in the absence of ligand and 0.17 ± 0.03 mM^{-1}s^{-1} in the presence of 400 μM dopamine (Figure 2.3*A*). For BM3h-B7, the corresponding r_1 values were 0.96 ± 0.13 mM^{-1}s^{-1} and 0.14 ± 0.04 mM^{-1}s^{-1}. Both sensor variants showed a dopamine concentration-dependent decrease in T_1-weighted MRI signal (up to 13% with 28.5 μM protein) that could be fitted by binding isotherms with estimated K_d values of 4.9 ± 2.7 μM for BM3h-8C8 and 2.7 ± 2.9 μM for BM3h-B7 (Figure 2.3*B,C*). For both BM3h variants, the stability, reversibility and rate of dopamine binding were established using spectroscopic assays (Supplementary Figures 2.6 and 2.7).

We investigated the reporting specificities of BM3h-8C8 and BM3h-B7 for dopamine by measuring MRI signal changes that resulted from incubation of 28.5 μM of each protein with 30 μM of either dopamine or one of eight other neuroactive molecules: norepinephrine (a neurotransmitter formed by catalytic hydroxylation of dopamine), 3,4-dihydroxy-L-phenylalanine (DOPA, the biosynthetic precursor to dopamine), serotonin, glutamate, glycine, -aminobutyric acid (GABA), acetylcholine, and arachidonic acid (Figure 2.3*D*). Of these potential ligands, only dopamine, norepinephrine, and serotonin elicited substantial changes in the T_1 relaxation rate ($1/T_1$). For BM3h-8C8, the $1/T_1$ reductions produced by norepinephrine and serotonin were 0.0076 ± 0.0023 s^{-1} and 0.0041 ± 0.0020 s^{-1}, respectively, compared to 0.0182 ± 0.0006 s^{-1} for dopamine; for BM3h-B7, norepinephrine and serotonin induced $1/T_1$ decreases of 0.0112 ± 0.0024 s^{-1} and 0.0171 ± 0.0005 s^{-1}, respectively, compared to 0.0208 ± 0.0002 s^{-1} for dopamine. We measured the affinities of BM3h-based dopamine sensors for these competitors spectroscopically (Figure 2.3*D*, inset). For BM3h-8C8, measured K_ds were 44 ± 3 μM and 80 ± 8 μM for norepinephrine and serotonin, respectively, and for BM3h-B7 the K_d values were 18.6 ± 0.4 μM and 11.8 ± 0.1 μM, respectively. Although both BM3h-8C8 and

BM3h-B7 show substantially higher affinity for dopamine than for norepinephrine (fivefold and sixfold, respectively) or for serotonin (ninefold and fourfold, respectively), the BM3h-8C8 variant is more specific for sensing dopamine at concentrations above 10 μM. In settings where dopamine is known to be the dominant neurotransmitter, BM3h-B7 may provide greater overall sensitivity.

The specificity data also provided a possible indication of the geometry of dopamine binding to the evolved BM3h proteins. Only monoamines showed affinity for BM3h-8C8 and BM3h-B7, whereas two catechols that lack primary amines, epinephrine and 3,4-dihydrophenylacetic acid, showed no measurable affinity (data not shown). Combined with the spectral evidence that dopamine directly coordinates the BM3h heme (Figure 2.1*C*), the titration results therefore suggest that the dopamine amine may serve as an axial ligand to the BM3h heme in the sensor-analyte complexes we examined.

2.3.3 BM3h-based sensors detect dopamine released from PC12 cells

We asked whether BM3h mutants produced by directed evolution could sense dopamine release in a standard cellular model of dopaminergic function. We applied an established protocol [123] to test the ability of our sensors to measure dopamine discharge from PC12 cells stimulated with extracellular K^+ (Figure 2.4*A*). Cells were cultured in serum-free medium supplemented with dopamine to promote packaging of the neurotransmitter into vesicles. After pelleting and washing, we resuspended cells in a physiological buffer containing 32 μM BM3h-B7 and either 5.6 or 59.6 mM K^+ (cells in the low-K^+ condition were osmotically balanced with Na^+). T_1-weighted MRI images (spin echo TE/TR = 10/477 ms) obtained with BM3h-B7 showed a $4.0 \pm 0.5\%$ reduction in signal intensity in the supernatant of K^+-stimulated cells, compared with cells for which isotonic Na^+ was used as control (Figure 2.4*B*). This corresponded to a $54 \pm 4\%$ decrease in sensor r_1 (Figure 2.4*C*). Given the dopamine dissociation constant of BM3h-B7 and its relaxivities under ligand-free and dopamine-saturated conditions, and assuming negligible dilution of the sensor after mixing with cells, we estimated supernatant dopamine concentrations of 60.3 ± 7.9 μM for stimulated cells and 22.2 ± 1.1 μM for controls. These estimates were in reasonable agreement with an independent quantification of dopamine release measured using an enzyme-linked immunosorbent assay (ELISA), which yielded concentrations of 54 ± 9 μM and 13 ± 2 μM for stimulated and control cells, respectively (Figure 2.4*D*). We were also able to use BM3h-8C8 to image dopamine release from PC12 cells. Under experimental conditions similar to above, BM3h-8C8 had a $37 \pm 2\%$ reduction in r_1 in the supernatant of K^+-stimulated cells relative to Na^+ controls (Supplementary Figure 2.8).

2.3.4 Dopamine detection in the brain of living rats

As an initial test of the ability of BM3h-based sensors to measure dopamine concentrations in intact animals, we injected BM3h-8C8 in the presence or absence of exogenous dopamine into the brains of anesthetized rats. We chose this simple experimental protocol for validation of the sensor because it guaranteed the presence of reproducible and unambiguous micromolar-level dopamine concentrations, suitable for evoking robust responses from our sensors *in vivo*. We obtained T_1-weighted MRI scans (fast spin echo TE/TR 14/277 ms, 8.9 s per image) continuously during 0.5-μl-min^{-1} paired infusions of 500 μM BM3h-8C8 with and without 500 μM dopamine, via cannulae implanted stereotaxically into the left and right striatum. Dopamine-dependent contrast changes were apparent in images obtained during and after the injection period (Figure 2.5*A*). We quantified MRI changes across multiple trials in striatal regions of interest (ROIs) that were reliably (though inhomogeneously) filled by convective spread of the contrast agent from the cannula tips (~1.5 mm radius). Consistent with results obtained *in vitro*, addition of dopamine dampened the observed MRI intensity enhancement by approximately 50% (Figure 2.5*B*); the effect was significant (t-test, $P = 0.003$, $n = 7$). We performed the same paired infusion procedure with WT BM3h, which has very low affinity for dopamine ($K_d \sim 1$ mM). As expected, the time course of the MRI signal during and after the WT BM3h injection period (Figure 2.5*C*) was not significantly affected by the presence or absence of dopamine (t-test, $P = 0.8$, $n = 5$), indicating that the dopamine-dependent signal differences shown in Figure 2.5*B* require the presence of a micromolar-affinity dopamine sensor and cannot be explained by physiological or biochemical effects of dopamine itself. Moreover, infusion of 500 μM dopamine alone into the brain produced no noticeable signal changes in an equivalent experiment (data not shown). Histological analysis showed minimal evidence of toxicity due to these procedures (Supplementary Figure 2.9). Using relaxivity values measured for BM3h-8C8 *in vitro*, we estimated maximal concentrations of 89 ± 19 μM BM3h-8C8 and 75 ± 28 μM dopamine from the data of Figure 2.5*B*, averaged across the striatal ROIs. The ability to quantify BM3h-8C8 concentration on the basis of its T_1 enhancement in the absence of elevated dopamine represents an advantage of this sensor's "turn-off" mechanism.

To test whether BM3h-8C8 could detect release of endogenous neurotransmitters in the rat brain, we acquired MRI data during co-infusion of the dopamine sensor with elevated concentrations of K^+, a depolarizing chemical stimulus shown previously to release large amounts of dopamine into the striatum [124, 125]. We chose K^+ over pharmacological stimuli to obviate potential solubility- or viscosity-related artifacts in the experimental paradigm. K^+ itself had no effect on r_1 of the BM3h variants (data not shown). In the stimulation experiments, three 5-min blocks of high-K^+ (153 mM) infusion alternated with 10 min 'rest' periods during which we administered a low-K^+ solution (3 mM, osmotically balanced with Na^+). Both high- and low-K^+ solutions were delivered at a rate of 0.2 μl min^{-1} and also contained 500 μM BM3h-8C8, ensuring that a relatively constant concentration of dopamine sensor was present throughout the procedure. We acquired T_1-weighted MRI scans continuously as for the exogenous dopamine infusion experiments. To control for effects unrelated to neurotransmitter sensing by the contrast agent (potentially including K^+-induced edema or hemodynamic responses incompletely suppressed by the T_1-weighted spin echo pulse sequence), we paired each striatal injection of BM3h-8C8 with an injection of WT BM3h into the opposite hemisphere, following the same blocked K^+ stimulation paradigm for both injections. As in conventional 'block design' fMRI, we performed a *t*-test analysis to evaluate the correspondence of each voxel's intensity time course with the alternating periods of low and high K^+. We determined an appropriate temporal shift for the stimulus-related analysis windows with respect to infusion buffer switches by observing the time courses of similarly switched mock infusions into 0.6% agarose phantoms [126] and by comparing these with statistical results as a function of offset (Supplementary Figure 2.10 and Supporting Material). As additional controls for MRI effects unrelated to dopamine sensing, we examined MRI signal change in response to K^+ stimulation and again in response to dopamine infusion, both in the absence of contrast agents (data not shown). We also continuously monitored blood oxygen levels and heart rate. In no case were stimulus-associated changes observed.

Figure 2.5*D* shows the distribution of voxels with significant (*t*-test, $P < 0.01$) MRI signal decreases in response to K^+ stimulation in a single rat. We performed a group analysis by combining data from all subjects ($n = 6$) over geometrically defined ROIs centered around the injection cannula tips in each animal. In three slices spanning the infusion site, seven voxels within 0.75 mm of the BM3h-8C8 injection cannula, but only one voxel near the WT cannula, showed strong correlation ($P < 0.01$) with the stimulus. We mapped mean signal decreases over 2.7-mm-diameter ROIs corresponding to the BM3h-8C8 and WT BM3h injection sites in the group analysis (Figure 2.5*E*). Again, dopamine sensor-dependent responses were apparent. The signal difference between low- and high-K^+ periods averaged across the entire BM3h-8C8 ROI (all voxels within a 2.7-mm-diameter by 3-mm-long cylinder, regardless of modulation by K^+) was 0.07%, whereas the signal difference averaged across the control ROI was 0.02% (Figure 2.5*F*). The high- versus low-K^+ signal difference observed near the BM3h-8C8 infusion site was significant (*t*-test, $P = 0.0008$) and consistent with the expected suppression of MRI signal by dopamine release under high-K^+ conditions.

The mean time course of all stimulus-correlated voxels ($P < 0.05$) showing K^+-induced MRI signal changes near the BM3h-8C8 injection site, averaged over animals, is shown in Figure 2.5*G*. Discernable signal decreases of up to 3% were produced during each K^+ stimulation block. The first K^+ block evoked the largest response (presumably because of partial dopamine depletion over subsequent blocks [116]) and elicited a clear spatiotemporal pattern of mean MRI signal change from baseline over the course of the stimulation period (Figure 2.5*G*, top panels).

2.4 Discussion

These results demonstrate the feasibility of developing molecular-level fMRI sensors and serve as a proof of principle that BM3h-based probes can be used to monitor dopamine signaling processes *in vivo*. With the experimental conditions and estimated sensor concentrations (34 ± 4 μM) used for our K^+ stimulation experiments, MRI signal changes of 3% would be evoked by the rewarding brain stimuli reported in previous studies to release large amounts of dopamine [115, 116]. This amplitude is reasonably large by functional imaging standards, and it could be used in the near term to map phasic dopamine release at high resolution across the striatum, or more generally to study mesolimbic dopamine dynamics in animal models of reward processing and neurological conditions that can be probed with strong stimuli.

Sensitivity gains will be possible using repeated stimulation and statistical analysis techniques, as in conventional fMRI, and by optimizing the imaging approach itself. For instance, higher-field scanners and faster alternatives to the T_1-weighted spin echo pulse sequences we used here may offer improved signal-to-noise ratios. Directed evolution or rational modification of BM3h variants for substantially higher relaxivity is possible as well (unpublished data). Sensors with higher relaxivity will produce larger MRI signal changes, and could have the added benefit of reducing the potential for dopamine buffering, because they may be used at lower concentrations *in vivo*: with 35 μM sensor and 35 μM total dopamine present, for example, $\sim$60% of the dopamine would be bound to the sensor, but with 15 μM sensor present, only $\sim$30% dopamine would be sequestered. Protein engineering techniques could also be used to improve the dopamine affinity and specificity of the first-generation sensors described here.

Our method for producing dopamine sensors represents a general paradigm for the development of molecular probes for MRI. Sensors may be evolved for targets inside or outside the brain; the diversity of potential targets is exemplified by the contrast between WT BM3h, which produces MRI signal changes in response to long-chain fatty acids, and BM3h-8C8 and BM3h-B7, which respond to a catecholamine. Contrast agents engineered to detect dopamine and other signaling molecules in the brain will permit functional neuroimaging based on direct detection of neuronal events rather than hemodynamic changes. Exogenous delivery of macromolecules such as BM3h to large regions of animal brains should be possible using a variety of techniques [127]. Because BM3h is a protein, it might also be possible to deliver variants via expression from transfected cells *in vivo* or in transgenic subjects. Preliminary evidence that BM3h can be expressed to 1% protein content in mammalian cells supports the feasibility of this approach (Supplementary Material).

Because of their small size, BM3h-based dopamine sensors might sample synaptic dopamine better than voltammetry or microdialysis probes, and with appropriate targeting could potentially become synapse specific. Dopamine sensor-dependent MRI would offer a combination of spatial coverage and precision inaccessible to other methods and uniquely suited to studies of dopaminergic function in systems neuroscience research.

2.5 Methods

2.5.1 Animal care

We performed all experiments involving vertebrate animals with approval of the Massachusetts Institute of Technology Committee on Animal Care.

2.5.2 Library construction

We constructed BM3h mutant libraries in accordance with a previously published protocol [98]. The starting parent for evolution was the WT heme domain of BM3 with a C-terminal hexahistidine tag, housed in the pCWori vector [128]. We produced mutant libraries through error-prone PCR using the primers 5'-GAAACAGGATCCATCGATGCTTAGGAGGTCAT-3' (forward) and 5'-GCTCATGTTTGACAGCTTATCATCG-3' (reverse) and Taq polymerase (AmpliTaq, Applied Biosystems) with 25 μM $MnCl_2$, producing ~1-2 mutations per gene. Between the fourth and fifth rounds of evolution, we introduced the mutation I366V into BM3h-8C8 via overlap extension PCR to improve protein thermostability [30].

2.5.3 Protein expression and high-throughput screening

We inoculated mutant colonies into deep-well 96-well plates containing 0.4 ml Luria broth (LB) medium and grew them overnight. On each plate, we included the parent clone and up to three previous parents in triplicate. We then transferred 0.1 ml of each culture to new plates containing 1.2 ml fresh terrific broth (TB) medium per well, supplemented with 100 μg ml^{-1} ampicillin, 0.2

mM isopropyl β-D-1-thiogalactopyranoside (IPTG), and 0.5 mM δ-aminolevulinic acid (ALA). We stored remaining LB cultures with glycerol at 80 °C. After 20–30 hours of protein expression at 30 °C, we pelleted cultures and lysed the pellets in 0.65 ml PBS containing 0.75 mg ml^{-1} hen egg lysozyme (Sigma-Aldrich) and 5 μg ml^{-1} DNase I (Sigma-Aldrich). We recorded absorbance spectra of 200 μl of cleared lysate from each well in a multiwell plate reader (Spectramax Plus, Molecular Devices) before and after addition of successively more concentrated dopamine or arachidonic acid. We analyzed the resulting absorbance spectra in MATLAB (Mathworks) using a custom routine that calculated the absorbance difference spectra for each acquisition relative to ligand-free lysate, computed the difference between maximum and minimum of each difference spectrum, plotted each value as a function of ligand concentration and, for each well, fitted a non-ligand-depleting bimolecular association function to estimate the corresponding K_d. We subsequently compared mutant K_d values to those of the parents within each plate and chose eight to ten mutants showing the greatest decrease in K_d for dopamine and/or the greatest increase in K_d for arachidonic acid for bulk expression and analysis.

2.5.4 Bulk expression and titrations

To produce selected proteins in bulk, we began by inoculating frozen LB cultures of candidate mutants into 30 ml TB medium containing 100 μg ml^{-1} ampicillin. We induced the cultures at log phase with 0.6 mM IPTG, supplemented them with 0.5 mM ALA and 50 μg ml^{-1} thiamine, and shook them for an additional 20–25 hours to express protein. We then lysed pelleted cells with Bug-Buster and Lysonase (EMD Chemicals) and purified BM3h mutants overNi-NTA agarose (Qiagen). We exchanged buffer to PBS over PD-10 desalting columns (GE Healthcare), and measured protein concentration using a carbon monoxide binding assay [129]. To characterize ligand affinities of the purified variants, we titrated protein samples with dopamine, arachidonic acid, serotonin, norepinephrine, pyrocatechol, 3,4-dihydroxy-L-phenylalanine (DOPA), 3,4-dihydroxyphenylacetic acid (DOPAC), homovanillic acid (HVA), 3-methoxytyramine (3MT), acetylcholine, glutamate, glycine, GABA, and epinephrine (all from Sigma Aldrich) and analyzed the results using Matlab as described above. We performed all measurements at room temperature (~ 21 °C). 3MT had a K_d of 73 ± 13 μM for BM3h-B7 and 183 ± 28 μM for BM3h-8C8. HVA and pyrocatechol showed no measurable affinity.

2.5.5 *In vitro* magnetic resonance imaging

To assess magnetic relaxation behavior of the proteins, we arrayed BM3h samples (60–100 μl) into microtiter plates and placed them in a 40-cm-bore Bruker Avance 4.7 T MRI scanner, equipped with a 10-cm-inner-diameter birdcage resonator radiofrequency coil and 26 G cm^{-1} triple-axis gradients. We filled unused wells of the microtiter plates with PBS and performed imaging at $\sim$21 °C on a 2 mm slice through the sample. We used a T_1-weighted spin echo pulse sequence; echo time (TE) was 10 ms, and repetition times (TR) were 73, 116, 186, 298, 477, 763 ms, 1.221, 1.953, 3.125 and 5.000 s. Data matrices consisted of 512 × 128 points, zero-filled to 1024 × 512 points, where the second dimension corresponds to the phase-encoding direction; the field of view (FOV) was 16 × 8 cm. We reconstructed and analyzed images using custom routines running in MATLAB and adjusted contrast to optimize MRI images presented in the figures. We calculated relaxation rates by exponential fitting to the image data, using an equation of the form $I = k[1 - \exp(TR/T_1)]$, where I was the observed MRI signal intensity and k was a constant of proportionality. We then determined values of r_1 by linear fitting to a plot of R_1 against protein concentration for six to eight BM3h concentrations in the range from 0 to 240 mM. We also performed low-field relaxivity measurements using benchtop spectrometers operating at 21 °C with proton resonance frequencies of 20 MHz and 60 MHz (Bruker Minispec NMS120 and mq60). Samples of 150 μL containing 50–100 μM BM3h-8C8 in PBS in the absence or presence of 500 μM dopamine yielded 20 MHz relaxivity measurements of 1.0 or 0.25 mM^{-1} s^{-1}, respectively, and 60 MHz relaxivities of 1.1 or 0.23 mM^{-1} s^{-1}, respectively.

2.5.6 Dopamine release from PC12 cells

We grew PC12 cells in suspension in F-12K medium supplemented with 15% (vol/vol) horse serum and 2.5% (vol/vol) FBS (ATCC). In preparation for dopamine release experiments, we incubated 50 ml cell cultures for 1 h in medium supplemented with 1 mM dopamine and 1 mM ascorbic acid, pelleted the cells and washed them twice with Locke's buffer (154 mM NaCl, 5.6 mM KCl, 3.6 mM NaHCO3, 2.3 mM CaCl2, 5.6 mM D-glucose, and 5 mM HEPES pH 7.4). To Locke's buffer missing 54 mM NaCl and containing or not containing the sensor, we added a 1:50 dilution of 2.7 M KCl (stimulus) or NaCl (control). We resuspended the washed PC12 cell pellets in 200 μl of either K^+- or Na^+-supplemented buffer, with or without sensor. After 30–60 min incubation at ∼21 °C, we pelleted cells and imaged the supernatant in an MRI scanner as described above. We estimated dopamine release by calculating sensor saturation level from observed r_1, then solving the quadratic equation describing bimolecular equilibrium binding with a known K_d, and assuming 32 μM of sensor for ligand concentration. We made independent measurements of dopamine release using the Dopamine EIA Kit (LDN).

2.5.7 Brain injection of sensors with exogenous dopamine

For injection experiments testing the effect of exogenous dopamine on BM3h-8C8 and WT BM3h (Figure 2.5*A*–*C*), we anesthetized adult male Lewis rats with 1–2% isoflurane. We stereotaxically inserted plastic guide cannulae (Plastics One) bilaterally into the striatum and secured them in place with dental cement (coordinates with respect to bregma: +0.7 mm anterior, 3 mm lateral, 6 mm below the surface of the skull). We connected tubing filled with silicone oil to an MRI-compatible dual channel syringe pump (Harvard Apparatus), attached it to internal cannulae and back-filled the cannulae with contrast agent solution, 500 μM BM3h-8C8 or WT BM3h, with or without equimolar dopamine (chemicals from Sigma-Aldrich). We ran the pump in infusion mode for a few seconds to ensure that no air entered the system and then lowered the internal cannulae (connected to the pump) into the implanted guide cannulae and fixed them in place with dental cement. After the implantation was complete, we transferred the animal to a plastic positioning device (Ekam Imaging) for imaging and placed it into a 4.7-T Bruker Avance scanner. We acquired fast spin echo (FSE) MRI scans (TE/TR 14/277 ms, 8.9 s per scan, $0.3 \times 0.3 \times 1.0$ mm resolution, 3.8×3.8 cm FOV, data matrix 128×128) before infusion (seven scans) and continuously during and after bilateral infusion of paired solutions, each injected for 20 min at 0.5 μl min^{-1}. We monitored heart rate continuously during the infusions using a Nonin Medical 8600V pulse oximeter equipped with a nonmagnetic sensor. We digitized raw oximetry readings using a National Instruments USB-6008 interface and converted them to heart rate using a MATLAB code. Values were stable at around 350 beats per minute ± 40 (s.d.).

We analyzed MRI data from these experiments using custom routines running in MATLAB. We detrended image signal time courses, converted them to percent change with respect to the preinjection baseline and averaged them over striatal ROIs. ROIs were chosen to approximate the maximal volumes reliably filled with the contrast agent, and were defined by a five-voxel in-plane radius (2.7 mm diameter) around the cannula tips over three image slices centered rostrocaudally around the implantation position, excluding voxels with notable signal dropout due to the cannulae themselves. We produced data for group analyses by combining data from ROIs defined separately with respect to the injection cannulae tips in each individual. We computed maximal MRI signal changes and signal change maps by averaging the image intensity at the end of the injection period (scans 121–140) and subtracting and normalizing it to the pre-injection intensity (scans 1–7).

2.5.8 Protein and dopamine quantification based on *in vivo* imaging data

We estimated absolute concentrations of contrast agent and dopamine under the assumptions that minimal endogenous dopamine was present, that the relaxivity and dopamine affinity of BM3h variants were the same *in vivo* and *in vitro*, and that the MRI acquisition procedure satisfied a strong T_1-weighting requirement, where TR $<<$ T_1. Under these assumptions, the fractional MRI signal change $\Delta I/I_0$ is approximately equal to $\Delta R_1/R_{10}$, the fractional change in R_1 (equal to $1/T_1$). R_{10} is the basal value of R_1, measured as 0.55 ± 0.01 s^{-1} from curve fitting to multiple FSE images obtained with different TR values. We estimated the maximal total concentration of

BM3h-8C8 by determining the corresponding ΔR_1 averaged over multiple injections of BM3h-8C8 in the absence of dopamine and dividing it by the relaxivity of the unliganded protein. We determined total dopamine concentration from the value of ΔR_1 observed during injection of BM3h-8C8 plus dopamine, the relaxivities of liganded and unliganded BM3h-8C8, the previously determined BM3h-8C8 concentration and the mass action relationships governing binding of the sensor to dopamine. In *in vitro* measurements, BM3h-8C8 had a T_2 relaxivity of 4 mM^{-1} s^{-1}; addition of 1 mM dopamine did not noticeably perturb this value significantly, suggesting that T_2 effects in conjunction with appropriate imaging methods might be able to provide a basis for protein quantification similar to the approach we describe here.

2.5.9 Histological analysis

After MRI contrast agent injection experiments using the paradigm described above, we placed rats under terminal anesthesia with ketamine and xylazine and transcardially perfused them with phosphate buffer containing heparin (Hospira) and then with 4% wt/vol paraformaldehyde (Sigma-Aldrich). We removed brains and obtained coronal cryosections of 10 μm thickness at 100 μm intervals across a range extending ~1 mm anterior and posterior to the injection cannula insertion site. We used standard protocols for hematoxylin and eosin staining. Terminal deoxynucleotidyl transferase-mediated dUTP nick-end labeling (TUNEL) staining was performed using the DeadEnd Colorimetric TUNEL system from Promega with visualization enhanced by the DAB Substrate Kit from Vector Laboratories. Histological procedures were implemented by Wax-it Histology Services.

2.5.10 *In vivo* potassium stimulation experiments

For *in vivo* K$^+$ stimulation experiments (Figure 2.5*D*–*G*), we used isoflurane-anesthetized male Lewis rats. We fitted internal cannulae with Y-connectors and positioned them through bilateral guide cannulae at coordinates 0.8 mm anterior to bregma, 2.8 mm lateral to the midline, and 7.8 mm below the skull surface. Each two-channel cannula delivered a given BM3h variant (BM3h-8C8 or WT, paired on opposite hemispheres of the brain; sides were randomized). On each two-channel injection cannula, we loaded one arm with protein in standard artificial cerebrospinal fluid (aCSF, containing 150 mM Na^+ and 3 mM K^+) and the other arm with protein in high-K^+ modified aCSF containing no Na^+ and 153 mM K^+. Two infusion pumps (Harvard Apparatus) drove the infusions; one pump controlled the standard aCSF (low-K^+) infusions on both BM3h-8C8 and WT control sides, and the other pump controlled high-K^+ infusions on both sides. We programmed the two pumps and synchronized them with the MRI experiment so as to acquire a preinfusion image baseline for 2 min, followed by continuous scanning over three stimulation cycles consisting of 10 min low K^+ alternating with 5 min high K^+, followed by a further 10 min of low K^+, followed by up to 30 min of post-injection signal acquisition. During these experiments, we continuously recorded heart rate and found it to be 355 beats per minute ± 45 (s.d.); blood oxygen saturation levels were 94.1 ± 5.8%.

We acquired T_1-weighted multislice MRI scan series as for the dopamine injection experiments described above. We imported raw data into MATLAB, processed them with spatial smoothing over nearest neighbors (in-plane) and converted them to percent signal change with respect to a fitted third-order polynomial baseline. Scans from the initial protein-only injection period (< 15 min) were excluded from the analysis. To statistically analyze data acquired during the three cycles of K^+ stimulation, we used a procedure analogous to classical fMRI methods, by performing a t-test on intensity values associated with high and low K^+ conditions. We considered voxels showing lower signal during the 5-minute intervals corresponding to K^+ stimulation to be consistent with the expected effect of K^+-evoked dopamine release on MRI signal in the presence of BM3h-8C8. We estimated that the delay between infusion pump switching and actual changes to K^+ concentration in the brain was roughly 8–9 min. We derived this estimate by recording the time required for spreading of Trypan blue to a radius of 0.75 mm (comparable to ROIs used for most of the analyses presented) from an injection cannula embedded in 0.6% agarose, in a switched injection paradigm equivalent to the K^+ stimulation paradigm applied *in vivo*. We chose a delay of 9 min for analyses presented in the text, but delays ranging from 7 to 12 min produced qualitatively similar statistical results, all with elevated numbers of voxels near the BM3h-8C8 infusion site showing the expected

MRI signal decrease upon K^+ stimulation and far fewer (if any) voxels near the WT BM3h control cannula showing significant ($P < 0.01$) effects (Supplementary Figure 2.10). We performed ROI-wide computations on cylindrical regions of 1.5 or 2.7 mm diameter in-plane extending over three (1 mm thick) image slices, centered about the BM3h-8C8 and WT BM3h infusion cannula tips, excluding from the calculations voxels showing substantial signal dropout due to the cannulae themselves. We performed group analyses by combining data from ROIs defined separately with respect to injection cannulae in each animal, without further anatomical coregistration.

2.6 Supplementary Material

2.6.1 DA binding to BM3h-B7 and BM3h-8C8 is stable and reversible

The stability of dopamine (DA) binding to BM3h-B7 and BM3h-8C8 was tested by incubating each protein with various amounts of DA at room temperature and measuring the absorbance difference between 430 nm and 410 nm over two hours (Supplementary Figure 2.6). During this time, a decline of less than 5% was observed when 1 μM sensor was incubated in the presence of excess DA (800 μM). Optical changes were greater when sensor was incubated with subsaturating concentrations of DA (up to 22% signal change for 1.3 μM DA incubated with BM3h-B7), consistent with the predicted effects of DA oxidation (known to take place under ambient conditions) on the partial saturation of the sensor. To test the reversibility of DA binding to BM3h-B7 and BM3h-8C8, we acquired absorbance spectra of the proteins alone and with 400 μM DA before and after filtering the solutions through a 30 kDa cutoff centrifugal filter. Successive steps of filtering and resuspension restored the original, ligand-free spectrum (Supplementary Figure 2.7). Rate constants for binding and unbinding of DA to BM3h-8C8 were estimated to be 3×10^3 $M^{-1}s^{-1}$ and 0.02 s^{-1}, respectively, at room temperature in phosphate buffered saline (PBS). Association was measured by rapid 1:1 mixing of protein (4.8 μM) and DA solutions (0.6–2.5 mM), followed by absorbance spectroscopy in a Hi-Tech KinetAsyst SF-61DX2 stopped-flow spectrometer (TgK Scientific, Wiltshire, UK). Dissociation was measured by fast 1:100 dilution of a solution containing 100 μM BM3h-8C8 and

50 μM DA, followed by absorbance recording in a Spectramax M2e spectrophotometer (Molecular Devices, Sunnyvale, CA).

2.6.2 Thermostabilization of BM3h-8C8

Mutations in the amino acid sequence of BM3h have been shown to reduce the proteins thermostability. This was observed during directed evolution of BM3h: the melting temperature (T_m, the midpoint temperature for thermal denaturation after 10 min) of WT BM3h was approximately 58 °C, while the T_m for BM3h-8C8 was approximately 48 °C. While this change did not significantly affect the proteins stability in physiologic buffer at room temperature, it did apparently reduce the yield of our bulk purification procedure. To improve thermostability of BM3h-8C8 before performing the fifth round of evolution, we introduced the mutation I366V, which has been shown previously to improve stability [14]. For BM3h-8C8 this improvement corresponded to a Tm increase by approximately 4 °C.

2.6.3 Histological analysis of injected rat brains

Histological analysis was performed on some of the experimental animals, following injection with contrast agents and MRI scanning (Supplementary Figure 2.9). Brains were extracted, and frozen sections were obtained from regions near the cannula placement sites. Hematoxylin and eosin (H&E) staining was performed to evaluate tissue architecture, and terminal deoxynucleotidyl transferase-mediated dUTP nick end labeling (TUNEL) staining was performed to assay cell death in brain regions near the injection sites. Some mechanical disruption occurred in areas immediately surrounding the cannula implantation tracts, but the tissue appeared normal in other respects. The tissue was intact in brain regions that did not include the cannula tracts. To obtain quantitative information about cell death near the injection site, TUNEL positive and negative nuclei were tallied in a region of 0.5 mm × 0.5 mm near the cannula tip placement site in a representative histological image; this site had been administered an injection solution containing 500 μM BM3h-8C8 and 500 μM DA. Fewer than 4% of nuclei appeared to be TUNEL positive, indicating that widespread cell death was not induced by introduction of the agents used in our experiments.

2.6.4 BM3h expression in mammalian cells

As a preliminary test of the feasibility of applying BM3h-based sensors as genetically encoded reporters in mammalian cells and animals, we created mammalian codon-optimized versions of the BM3h and BM3h-8C8 genes using Gene Designer software (DNA2.0, Menlo Park, CA). The resulting sequences were synthesized by Blue Heron (Bothell, WA), cloned into a PCMV-Sport vector (Invitrogen, Carlesbad, CA), and transfected into the HEK293 cell line adapted for the Invitrogen Freestyle293 expression system. One day after transfection, cells were supplemented with 0–40 μM hemin, to ensure an adequate supply of heme for expression and folding of BM3h. Cells were lysed after four days of expression. Proteins were purified from lysates by nickel affinity chromatography and analyzed by optical spectroscopy. Spectra of mammalian cell-expressed and recombinant bacterial BM3h were virtually identical, indicating correct folding of the proteins. Absolute BM3h expression levels were measured by performing CO assays on clarified HEK293 lysates, and the results were compared with BCA assay (Pierce, Rockford, IL) results indicating the total amount of protein present. BM3h accounted for 0.4% to 1.9% of total cell protein depending on the level of hemin supplementation during growth. Assuming a typical protein density of 150 mg/mL in living tissue, 1% of total protein would correspond to a tissue concentration of 25 μM. This figure indicates cytosolic expression levels, but robust expression of secreted BM3h, appropriate for extracellular DA sensing, should also be feasible. High level expression of secreted proteins from mammalian cells has been described in the literature. One study reported yields of 5 mg/L over a two-hour period from monolayer cultures containing approximately 10^5 cells per 1.5 mL well [130], a figure likely to constitute well over 10% of total protein in the samples, and therefore significantly greater than the intracellular yields we obtained from BM3h expression in HEK293 cells. In mice, comparably high expression levels of several mg/L in serum of a secreted enzymatic reporter protein have been observed following adenoviral infection [131] or genetically-modified xenograft implantation [132]. Another point of reference related to attainable secreted protein expression levels is the protein content of the brains extracellular space, which has been estimated at 1% by weight [133], equivalent to 200 μM of proteins averaging 50 kD.

2.7 Figures

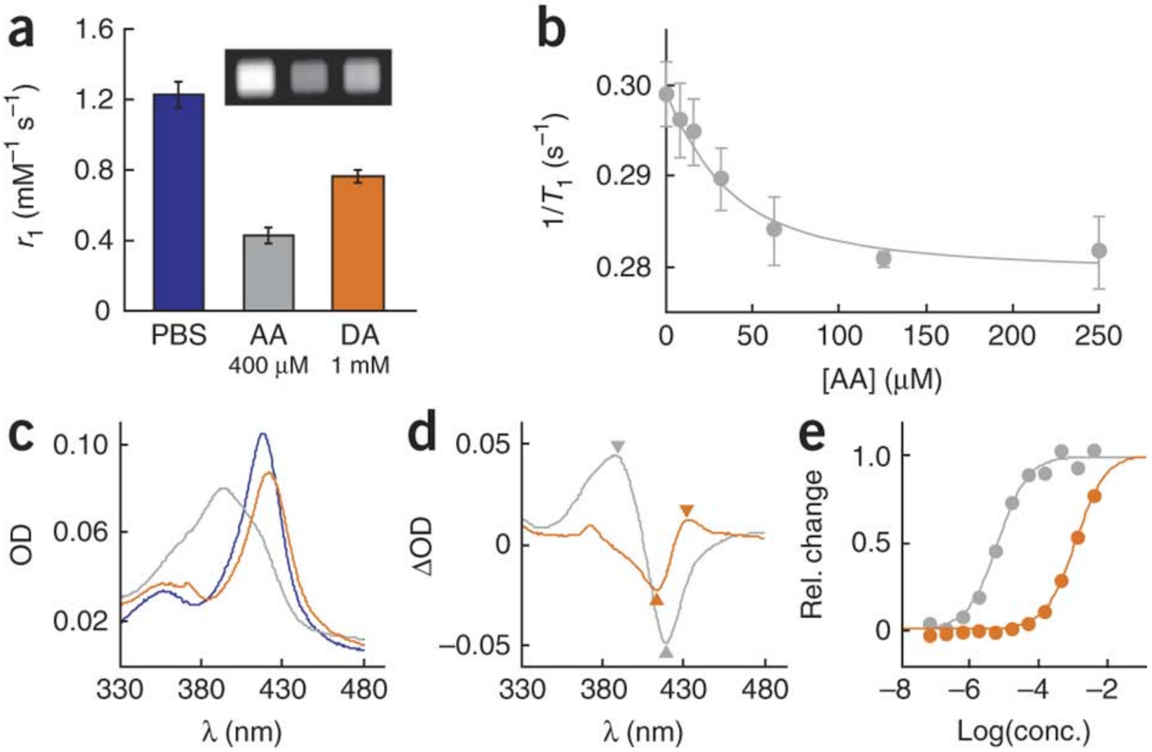


Figure 2.1: Ligand binding to the BM3 heme domain changes MRI contrast and optical absorption in a concentration-dependent manner. (A) T_1 relaxivity (r_1) of BM3h in PBS solution and in the presence of 400 μM arachidonic acid (AA) or 1 mM dopamine (DA); inset shows T_1-weighted spin echo MRI image intensity (TE/TR = 10/477 ms) of microtiter plate wells containing 240 μM BM3h in PBS alone (left) or in the presence of 400 μM arachidonic acid (middle) or 1 mM dopamine (right). (B) T_1 relaxation rates ($1/T_1$) measured from solutions of 28.5 μM BM3h incubated with 0–250 μM arachidonic acid. (C) Optical absorbance spectra of 1 μM BM3h measured alone (blue) and after addition of 400 μM arachidonic acid (gray) or 1 mM dopamine (orange). OD, optical density. (D) Difference spectra showing the change in BM3h absorbance as a function of wavelength upon addition of 400 μM arachidonic acid (gray) or 1 mM dopamine (orange). (E) Normalized titration curves showing binding of BM3h to arachidonic acid (gray) or dopamine (orange). We computed the optical signals used for titration analysis by subtracting the minimum from the maximum of difference spectra (arrowheads in D) under each set of conditions. Error bars in A, B, and E reflect s.e.m. of three independent measurements (errors in E were smaller than the symbols).

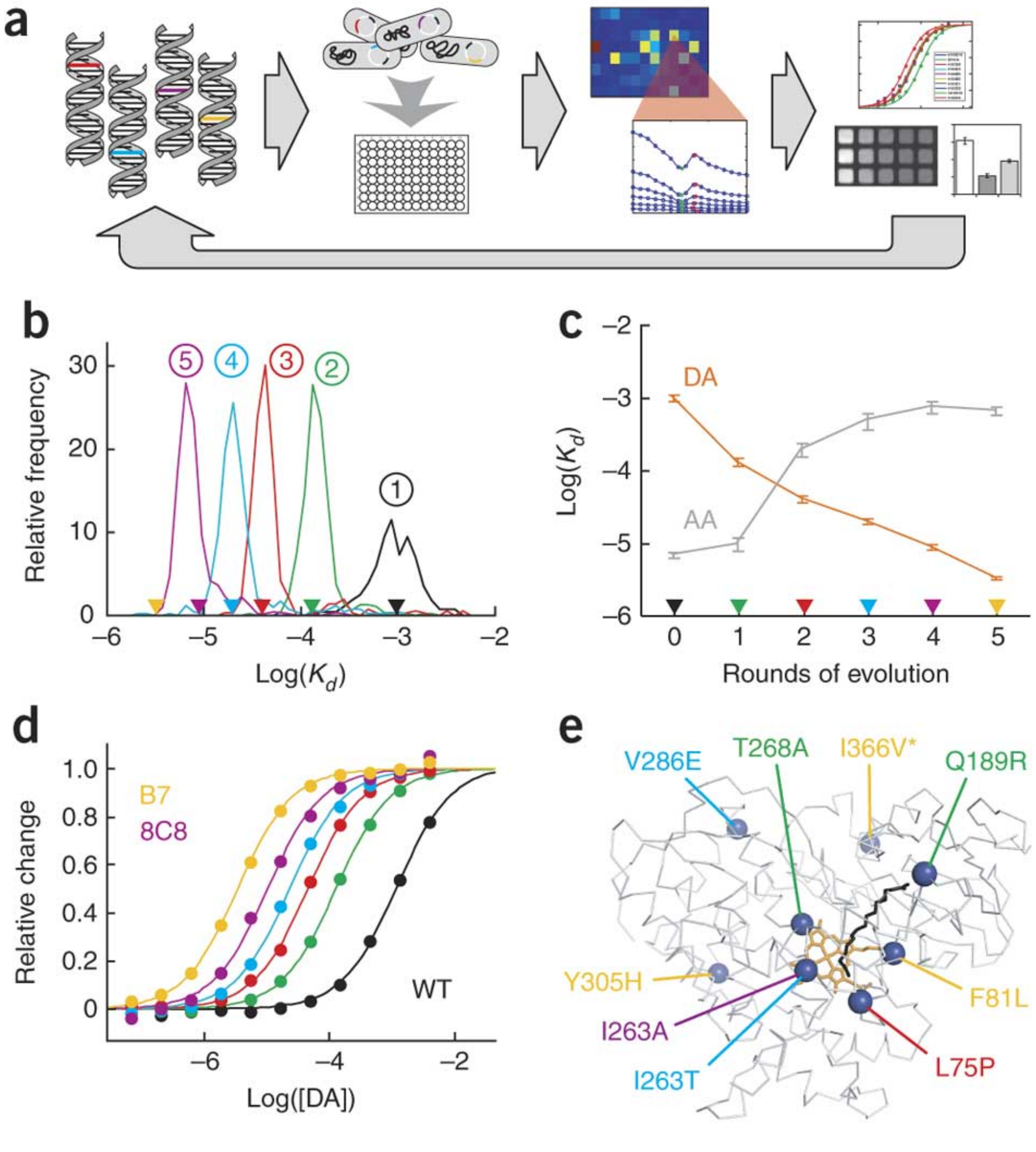

a
b
Relative frequency
30
20
10
0
–6
–5
–4
–3
–2
Log(K_d)
5
4
3
2
1
c
Log(K_d)
–2
–3
–4
–5
–6
DA
AA
0
1
2
3
4
5
Rounds of evolution
d
Relative change
1.0
0.8
0.6
0.4
0.2
0
–6
–4
–2
Log([DA])
B7
8C8
WT
e
V286E
T268A
I366V*
Q189R
Y305H
F81L
I263A
I263T
L75P

Figure 2.2

Figure 2.2: Screen-based isolation of BM3h mutants with enhanced dopamine affinity. (*A*) Schematic of the directed evolution approach, including (left to right) generation of a mutant DNA library, transformation into *E. coli* and growth in multiwell plate format, spectroscopic analysis of each mutant's ligand binding affinities, and detailed MRI and optical characterization of selected mutant proteins. (*B*) Histograms of mutant dopamine dissociation constants determined during each round of directed evolution, comparing each mutant protein's relative dopamine affinity (measured in plate format) to the K_d of the parent protein (measured in bulk). K_d distributions for screening rounds 1 (black), 2 (green), 3 (red), 4 (cyan), and 5 (purple) are labeled with numbers in circles. Color-coded arrowheads indicate the measured K_ds of parent proteins used to create the library of mutants at each round; yellow arrowhead indicates the K_d of the mutant protein selected after round 5. (*C*) Dissociation constants for dopamine (DA; orange) and arachidonic acid (AA; gray) for WT BM3h and mutant BM3h variants isolated at each round of screening; progressive increases in dopamine affinity and attenuation of arachidonic acid affinity are evident. Colored arrowheads indicate correspondence with data in *B*. Error bars denote s.e.m. of three independent measurements. (*D*) Titration analysis of dopamine binding to WT BM3h and to proteins selected after each round of directed evolution (colored as in *B*). Mutant proteins identified by rounds 4 (8C8) and 5 (B7) were considered to be end products of the screening procedure. (*E*) X-ray crystal structure [134] of WT BM3h (gray; heme group shown in orange) bound to palmitoleic acid (black), indicating the locations of amino acid substitutions accumulated during directed evolution of enhanced dopamine binding affinity. Each mutation's location is marked with a blue sphere and a label color-coded according to the parent protein in which the substitution was first identified (see legend for *B*). The previously characterized I366V mutation (asterisk) was incorporated between screening rounds 4 and 5 to improve the thermostability of the engineered proteins.

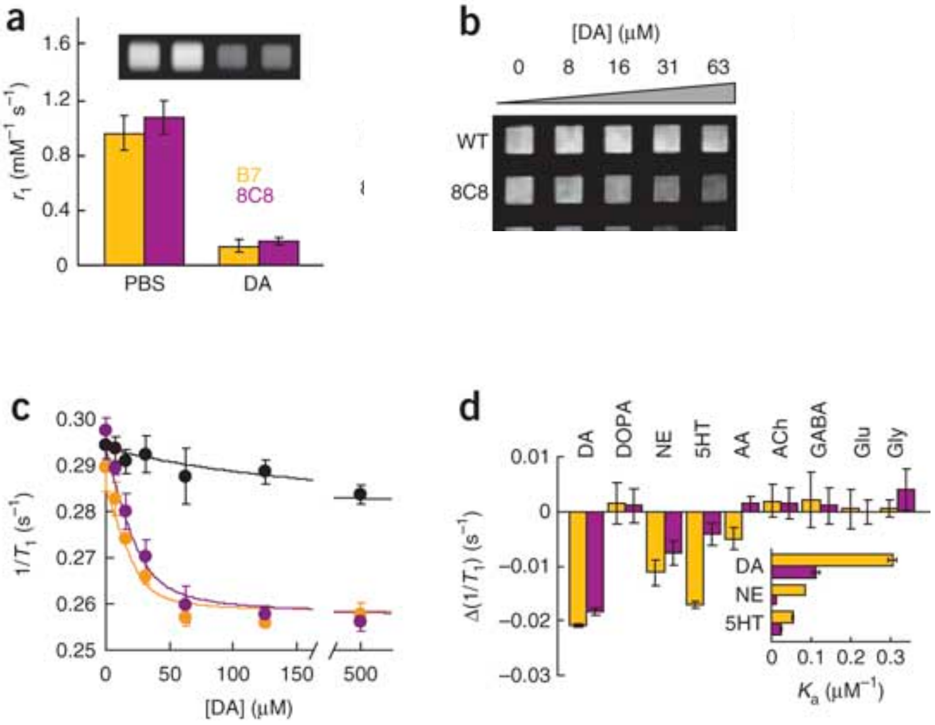


Figure 2.3: Selected sensor proteins produce strong and specific MRI signal changes in response to dopamine. (*A*) Relaxivity values measured from BM3h-B7 (yellow bars) and BM3h-8C8 (purple bars) in PBS alone or in the presence of 400 μM dopamine (DA). Inset, T_1-weighted MRI signal (TE/TR = 10/477 ms) obtained from 195 μM BM3h-B7 or BM3h-8C8, each incubated in microtiter plate wells with or without 400 μM dopamine (wells ordered left to right as in the bar graph). (*B*) MRI image showing signal amplitudes measured from wells containing 28.5 μM WT BM3h, BM3h-8C8 or BM3h-B7, each incubated with increasing dopamine concentrations (0–63 μM, left to right). The image was obtained using a T_1-weighted pulse sequence (TE/TR = 10/477 ms). (*C*) Relaxation rates ($1/T_1$ values) measured from solutions of 28.5 μM WT BM3h (black), BM3h-B7 (yellow), or BM3h-8C8 (purple), as a function of total dopamine concentration. Curves were fit using a ligand-depleting bimolecular association model. (*D*) Changes in $1/T_1$ relative to ligand-free protein for 28.5 μM BM3h-B7 (yellow) or BM3h-8C8 (purple) incubated with 30 μM dopamine, serotonin (5HT), norepinephrine (NE), DOPA, arachidonic acid (AA), acetylcholine (ACh), GABA, glutamate, or glycine. Inset, spectroscopically determined affinities ($K_a = 1/K_d$) of BM3h-B7 and BM3h-8C8 for dopamine, serotonin, and norepinephrine. Error bars in panels *A, C*, and *D* denote s.e.m. of three independent measurements.

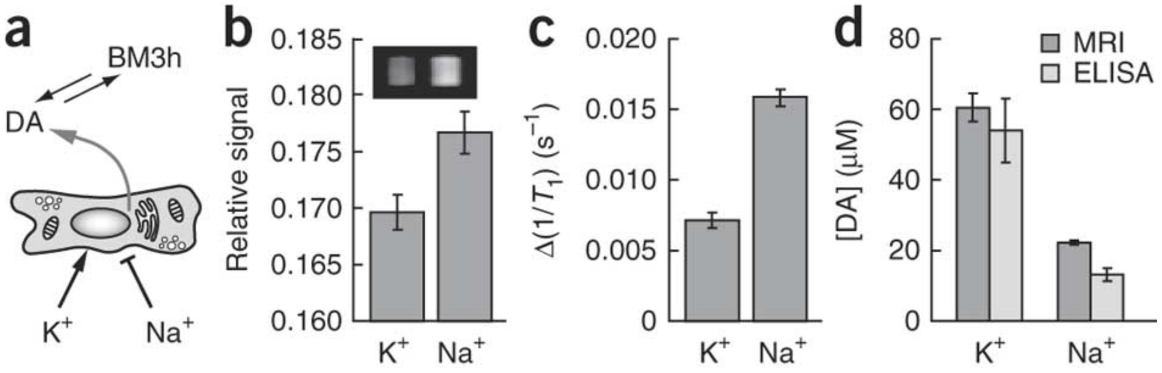


Figure 2.4: BM3h-based sensors measure dopamine release in cell culture. (*A*) PC12 cells depolarized by addition of 54 mM K$^+$ were stimulated to release dopamine (DA) into supernatants containing a BM3h-based sensor; cells did not release dopamine after addition of 54 mM Na$^+$. (*B*) T_1-weighted spin echo MRI signal amplitudes (TE/TR = 10/477 ms) measured from the supernatants of PC12 cells incubated with 32 μM BM3h-B7 in the presence of K$^+$ (stimulus) or Na$^+$ (control). Inset, MRI image of microtiter wells under corresponding conditions. (*C*) Relaxation rates measured from the samples in *B*, minus the relaxation rate of buffer not containing BM3h-based sensors. Given the approximate concentration of BM3h variants in these samples, the $\Delta(1/T_1)$ values presented here can be converted to apparent relaxivities of 0.23 and 0.50 mM^{-1} s^{-1} in K$^+$ and Na$^+$ incubation conditions, respectively. (*D*) Data from *C* were used to estimate the concentrations of dopamine present in samples treated with K$^+$ and Na$^+$ (dark bars). We independently measured the concentrations of dopamine under equivalent conditions using ELISA (light bars).

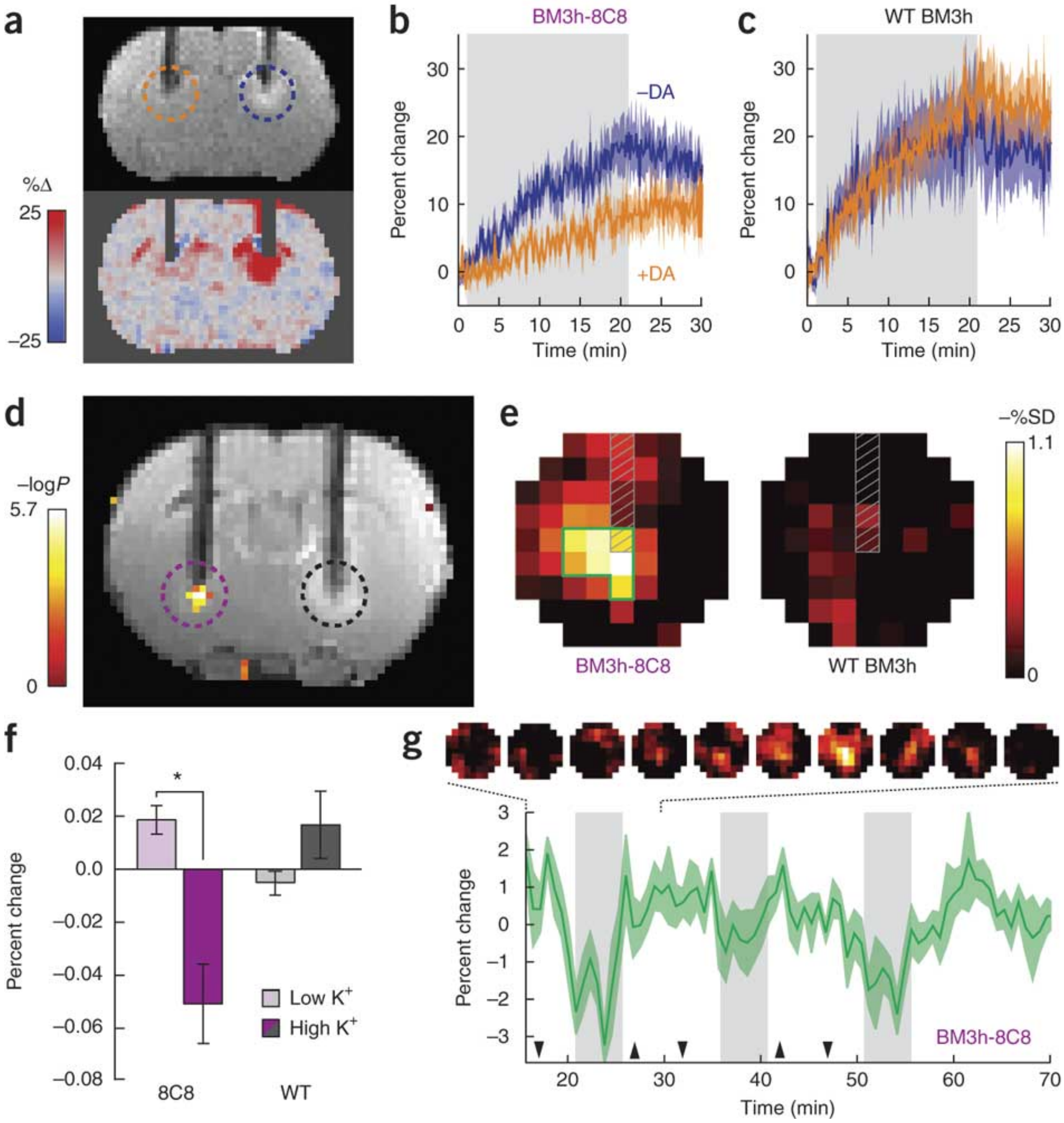
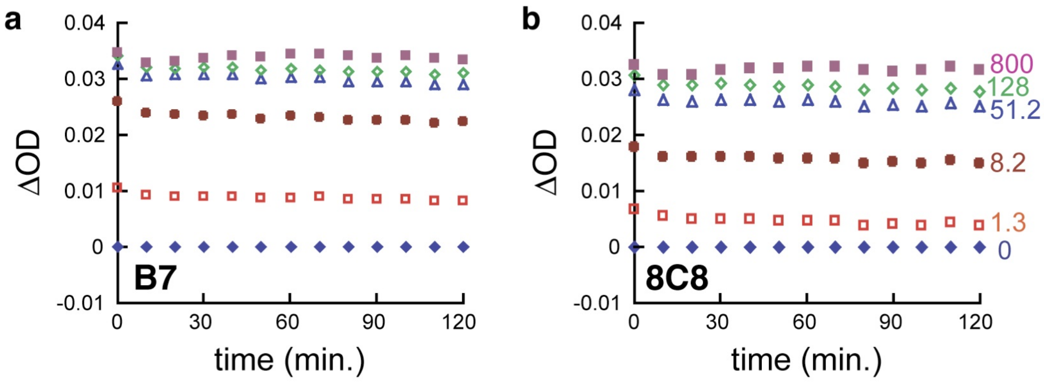
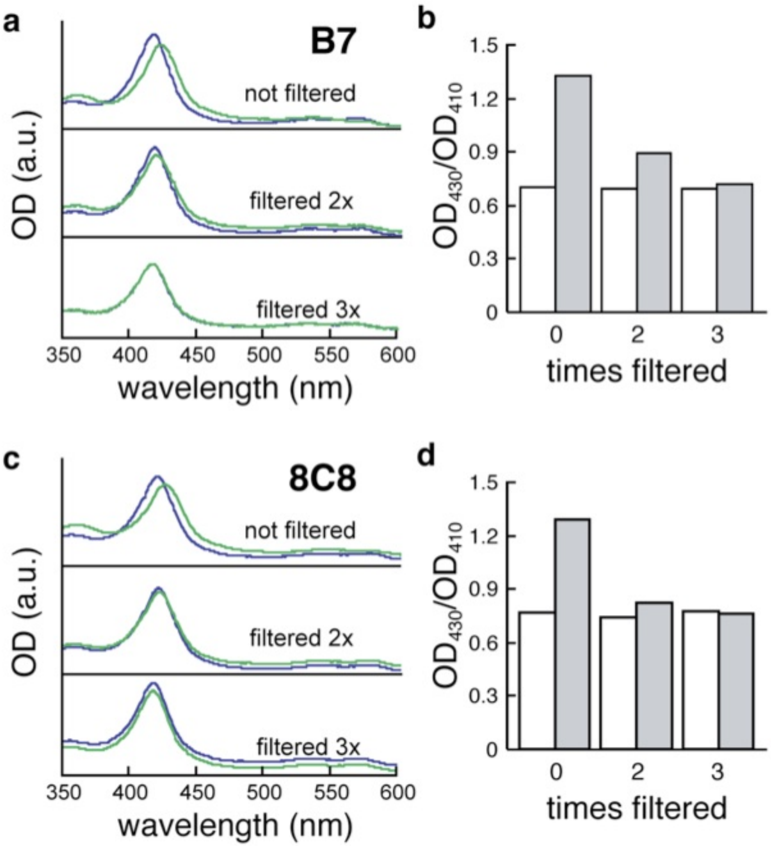

a
%Δ
25
–25
b
BM3h-8C8
–DA
+DA
Percent change
Time (min)
c
WT BM3h
d
–log$P$
5.7
0
e
–%SD
1.1
0
BM3h-8C8
WT BM3h
f
0.04
0.02
0.0
–0.02
–0.04
–0.06
–0.08
*
Low K$^+$
High K$^+$
8C8
WT
g
BM3h-8C8
Time (min)

Figure 2.5

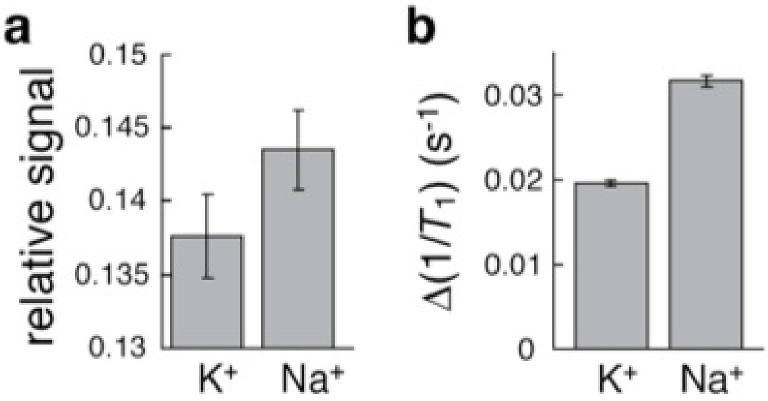
Figure 2.5: BM3h-8C8 reports dopamine in injected rat brains. (*A*) Top, coronal MRI image (0.7 mm anterior to bregma, averaged over the injection period) from a single rat injected with 500 μM BM3h-8C8 in the presence (orange dashed circle) or absence (blue dashed circle) of equimolar dopamine; the image contrast was linearly adjusted for display. MRI hyperintensity is noticeable near the tip of the dopamine-free cannula. The circles indicate approximate ROIs ($\sim$1.5 mm around cannula tips) over which image intensity was averaged for quantitative analyses. Bottom, map of percent signal change (%Δ) for the same animal, computed by comparing pre- and post-injection MRI signal. Areas corresponding to both high- and low-dopamine co-injections (+DA and DA) are delineated by apparent signal changes, but the strong difference between the two conditions is clear. (*B*) Time courses of relative signal change observed during injection of BM3h-8C8 DA (blue) or +DA (orange), averaged over multiple animals ($n = 7$) in ROIs denoted in *A*. Gray shading denotes the 20 min injection period. (*C*) Corresponding time courses of a control injection in which WT BM3h was introduced instead of the dopamine sensor ($n = 5$). (*D*) Statistical parametric map of *t*-test significance values (color scale) for correlation of MRI intensity with low- and high-K$^+$ conditions in an individual rat, overlaid on a corresponding T_1-weighted coronal slice (grayscale) showing injection cannulae used for BM3h-8C8 infusion (left, purple dashed circle) and WT BM3h control infusion (right, black dashed circle). (*E*) Maps of percent signal difference (SD) between high- and low-K$^+$ conditions observed in 2.7-mm-diameter ROIs centered around BM3h-8C8 sensor (left) and WT BM3h control (right) injection sites, after spatial coregistration and averaging across multiple animals ($n = 6$); ROIs correspond approximately to the color-coded circles in d. Voxels outlined in green are those that showed the most significant correlation with the K$^+$ stimulus regressor in the group analysis (Student's *t*-test, $P < 0.01$); these generally showed $\sim$1% mean signal change. Gray cross-hatching indicates approximate locations of the infusion cannulae. (*F*) Mean MRI signal change from baseline observed during high-K$^+$ (dark bars) and low-K$^+$ (light bars) periods in ROIs centered around infusion sites for BM3h-8C8 (purple) and WT BM3h (gray) proteins. ROIs were cylinders 2.7 mm in diameter and extending over 3-mm-thick slices registered around the infusion sites; signal was averaged in unbiased fashion over all voxels, regardless of correlation with the stimulus. The signal difference in the presence of BM3h-8C8 was statistically significant ($P = 0.0008$, asterisk). (*G*) Graph shows the mean time course of MRI signal in voxels within the BM3h-8C8-infused ROI and identified as correlated ($P < 0.05$) with the stimulus, averaged over animals and binned over 1.5 min intervals (shaded area denotes s.e.m., $n = 6$; individual traces are shown in Supplementary Figure 2.11). Gray vertical bars denote periods associated with highest K$^+$ stimulation, accounting for delays due to convective spreading of K$^+$ from the cannulae tips and the dead time of the injection apparatus. Arrowheads indicate the timing of pump switches associated with transitions from low to high (up) and from high to low (down) K$^+$ infusion conditions. Panels above the graph depict 'snapshots' of signal change spaced throughout the first K$^+$ stimulation cycle, as indicated by the dotted lines. The ROI corresponds to the left side of *E*, and the color scale denotes 0% (black) to 3% (yellow) signal change from baseline at each voxel and time.



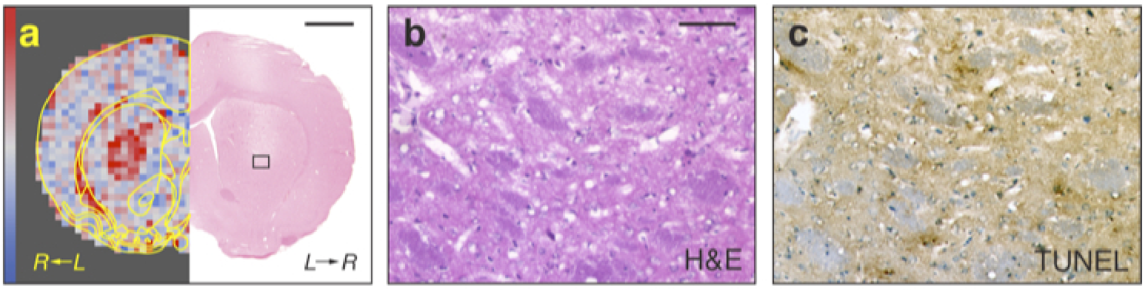
Supplementary Figure 2.6: Stability of BM3h complexes. Optical signatures of DA binding to BM3h-B7 (A) and BM3h-8C8 (B) are stable over two hours. Absorbance at 430 nm minus 410 nm collected over 2 hours for 1 μM sensor incubated in the presence of 0-800 μM DA (labels in color).



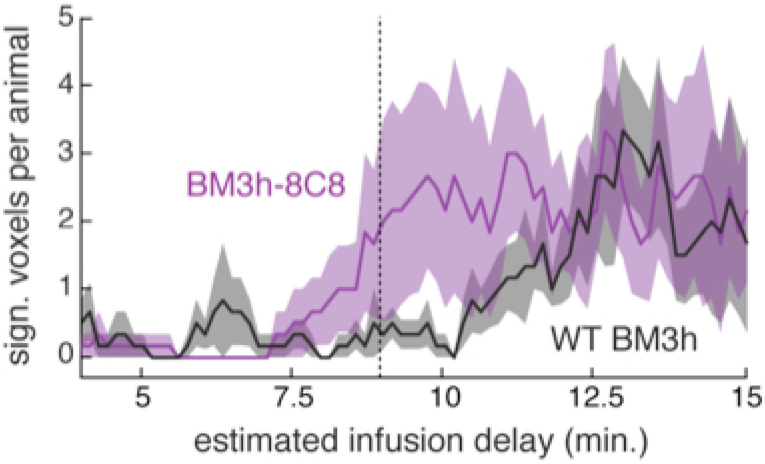
Supplementary Figure 2.7: DA binding to BM3h-B7 and BM3h-8C8 is reversible. (A) Absorbance spectra of 1 μM BM3h-B7 alone (blue) or incubated with 400 μM DA (green) before filtering (top), and after filtering twice (middle) or three times (bottom) through a 30 kDa cutoff filter. (B) Ratios of absorbance at 430 nm to 410 nm corresponding to the DA-free (white bars) and DA-incubated (gray bars) spectra in A. Panels C and D display equivalent data for sensor variant BM3h-8C8.



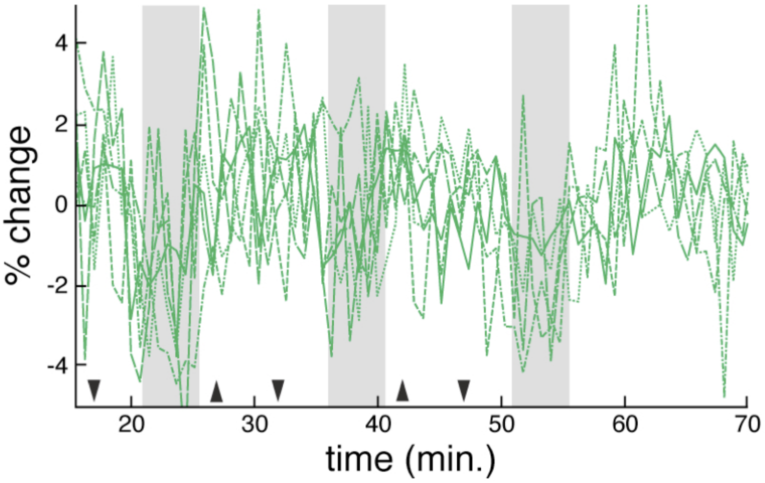
Supplementary Figure 2.8: BM3h-8C8 reports DA release from PC12 cells. Plots of signal change (A) and $\Delta(1/T1)$ (B) are analogous to Figure 2.4, and were obtained using identical experimental procedures.



Supplementary Figure 2.9: Histological data from a rat injected with 500 μM BM3h-8C8 and 500 μM DA. Panel A shows a comparison of MRI and hematoxylin and eosin (H&E) stained brain hemi-slices, taken from near the injection site (∼1 mm anterior to the cannula position). The pseudocolored MRI image (left) maps the percent change in MRI signal (relative to preinjection baseline) observed in this animal; the color scale (left edge) ranges from -25% to +25%. The superimposed yellow brain atlas diagram [135] demonstrates that the region of greatest signal change following this injection falls within the caudate nucleus of the striatum. The right side of panel A shows a corresponding H&E section from the same animal (scale bar = 2 mm) (H&E and MRI data are shown as mirror images, to facilitate comparison). A closeup view (20x) of the H&E stained section in the region of greatest contrast agent accumulation shows normal cellular staining (B). The scale bar denotes 100 μm, and the field of view shown in B corresponds to the small rectangle in the right half of panel A. TUNEL staining was performed to assay cell viability in the injected ROI (C). Over 95% of visualized nuclei were judged to be TUNEL negative, indicating the overall health of the tissue.



Supplementary Figure 2.10: Statistical analysis of K$^+$ stimulation data with variable delay between infusion switching and estimated periods of high and low K$^+$ in the brain. Data acquired during three cycles of potassium stimulation were analyzed over a range of delays, and the number of voxels showing significant (t-test $p < 0.01$) signal decreases during modeled periods of high K$^+$ was tallied in 0.75 mm radius ROIs extending over three 1 mm slices around the BM3h-8C8 (purple) and WT BM3h control (black) infusion sites. The plot shows the average (solid line) and standard error (shading) across the six individual animals that contributed to the analysis. The horizontal scale spans a range from the dead time of the injection cannulae (4 min.) to the duration of an entire stimulus cycle (15 min.). The vertical dotted line specifies the delay (9 min.) used to generate figures and data cited in the main text.



Supplementary Figure 2.11: Time courses of MRI signal in the six individual animals contributing to the average in Figure 2.5G. For each animal (represented by different line styles), signal was averaged over all voxels within ROIs of 2.7 mm diameter and three 1 mm slice thickness around BM3h-8C8 infusion sites, and determined to be correlated ($p < 0.05$) with K$^+$ stimulation blocks. As in Figure 2.5G, gray vertical bars denote periods associated with highest K$^+$ stimulation, and arrowheads label timing of pump switches associated with transitions from low to high (up) and high to low (down) K$^+$ infusion conditions.

Chapter 3

SCHEMA-designed chimeras of Human Arginase I and II reveal sequence elements important to stability and catalysis

3.1 Abstract

Arginases catalyze the divalent cation-dependent hydrolysis of L-arginine to urea and L-ornithine. The divalent metal cluster of arginase is not only integral to catalysis, but also contributes to the stability of the enzyme. There is much interest in using arginase as a therapeutic anti-neogenic agent against L-arginine auxotrophic tumors and in enzyme replacement therapy for treating hyperargininemia. Both therapeutic applications require enzymes with sufficient stability under physiological conditions. Using SCHEMA structure-guided recombination, we designed and synthesized a diverse set of active chimeric arginases that are composed of sequence fragments from the two human isozymes Arginase I and II. Within this data set, linear regression was used to identify the sequence elements that contribute to an arginase's long-term stability under physiological conditions. Our data revealed: (i) a striking correlation between an arginase's isoelectric point and its long-term stability and (ii) a moderate correlation between an arginase's metal affinity and catalytic efficiency.

3.2 Introduction

Humans produce two arginase isozymes (EC 3.5.3.1) that catalyze the hydrolysis of L-arginine (L-Arg) to urea and L-ornithine (L-Orn). The Arginase I (hArgI) gene is located on chromosome 6 (6q.23), is highly expressed in the cytosol of hepatocytes, and functions in nitrogen removal as the final step of the urea cycle. The Arginase II (hArgII) gene is found on chromosome 14 (14q.24.1). Arginase II is mitochondrially located in tissues such as kidney, brain, and skeletal muscle, where it is thought to provide a supply of L-Orn for proline and polyamine biosynthesis [136]. Both enzymes, which share 61% sequence identity, adopt a homo-trimeric structure composed of an α/β fold of a parallel eight-stranded β-sheet surrounded by several helices. These enzymes contain a di-nuclear metal cluster that generates a hydroxide for nucleophilic attack on the guanidinium carbon of L-arginine [137, 138]. In eukaryotes and most prokaryotes, the native metal cofactor in arginase is Mn^{2+}.

There is significant interest in applying arginases as cancer chemotherapeutic agents. A number of high morbidity tumors such as hepatocellular carcinomas (HCC), melanomas, renal cell, and prostate carcinomas [139, 140, 141] are deficient in the urea cycle enzyme, argininosuccinate synthase (ASS), and thus are sensitive to L-arginine (L-Arg) depletion. Non-malignant cells typically enter into quiescence (G0) when deprived of L-Arg and remain viable for several weeks. However, ASS-deficient tumor cells experience cell cycle defects that lead to the re-initiation of DNA synthesis even though protein synthesis is inhibited, in turn resulting in major imbalances that lead to rapid cell death [142, 143]. The selective toxicity of L-Arg depletion for HCC, melanoma and other urea-cycle enzyme deficient cancer cells has been extensively demonstrated *in vitro*, in xenograft animal models and in clinical trials [139, 144, 142, 140]. Arginase has also been explored as an enzyme replacement therapy to treat hyperargininemia. Although rare, autosomal recessive errors in hArgI can result in hyperargininemia, which clinically presents with hyperammonemia, spasticity, seizures, and failure to thrive [145]. Dietary management in combination with oral phenylbutyrate is often successful in controlling hyperammonemia, but the underlying hyperargininemia can persist, which can result in L-arginine-associated neurotoxicity [146]. Red blood cell replacement, which provides supplemental hArgI, has shown promise in treating hyperargininemia as evidenced by reduced serum L-Arg levels and improved clinical outcomes [147, 148].

To function as a therapeutic, arginase must efficiently degrade L-Arg under physiological conditions (∼100 μM L-Arg, 37 °C, and pH 7.4) and not produce adverse immunological responses. For these reasons, the human arginase isozymes are logical starting points for the development of a therapeutic enzyme. Unfortunately, both hArgI and hArgII display low enzymatic activity at physiological pH and are rapidly inactivated in blood serum, with half-lives of only a few hours. We recently reported that Co^{2+}-substituted hArgI (Co-hArgI) displays a dramatically reduced K_m for L-Arg relative to the native Mn^{+2}-containing enzyme, which leads to a twelvefold increase in k_{cat}/K_m. More importantly, Co-hArgI is significantly more stable in blood serum, with an inactivation half-life of over 30 hours [149]. With these enhanced properties, Co-hArgI displays potent tumor cytoxicity against numerous cancer cell lines *in vitro*, and inhibits the growth of HCC and pancreatic carcinomas in mouse xenograft models [149, 150].

Here, we designed a highly informative set of engineered arginases and used this set of sequences to study the biophysical properties of arginase enzymes. We start by designing a SCHEMA structure-guided recombination library of arginase chimeras composed of sequence fragments from the human arginases hArgI and hArgII. Past SCHEMA libraries have contained a large number of functional and diverse sequences [80, 151], providing ideal data sets for investigating the properties of sequences within a protein family. Next, we select a maximally informative subset of the arginase SCHEMA library using a novel active learning algorithm which identifies sequences that are both functional and highly informative. This diverse set of engineered arginases is used to study the long-term stability and catalytic efficiency of arginases under physiological conditions. From this investigation, we propose a mechanism of arginase inactivation and suggest future strategies for engineering arginases with exceptional long-term stability.

3.3 Materials and Methods

3.3.1 Active learning algorithm

The active learning algorithm consists of a two-step experimental design. The first step involves finding an informative set of chimeras for a logistic regression model of functional status, that is whether a chimeric sequence is folded and has arginase activity. Here, we would like to find the set of sequences S which maximize the mutual information between the chosen set of chimeras and the remainder of the library

$$I(S; L\backslash S) = H(L\backslash S) - H(L\backslash S|S), \tag{3.1}$$

where $H(L\backslash S)$ is the Shannon entropy of library L excluding the chimeras in subset S and $H(L\backslash S|S)$ is the entropy of the same sequences after the chimeras in S have been observed. We approximate the intractable entropy of the Bayesian logistic regression model by replacing the logistic response with a Gaussian likelihood. Gaussian mutual information is a submodular set function [152] and therefore can be maximized efficiently using a greedy approximation algorithm [153]. The functional status of the resulting sequences was then used to train a Bayesian logistic regression model, which can predict the probability a chimera will function for all sequences in the library.

The second step of the algorithm consists of finding a highly informative set of functional chimeric arginases. Here, we want to find the set of chimeras S which maximize the expected value of the mutual information

$$\mathrm{E}[I(S; L\backslash S)] = \sum_{A \in \mathcal{P}(S)} \left[I(A; L\backslash A) \prod_{\mathbf{c} \in A} p_{\mathbf{C}} \prod_{\mathbf{c} \in (S\backslash A)} 1 - p_{\mathbf{C}} \right], \tag{3.2}$$

where the sum is over all subsets A in the power set of S, and $p_{\mathbf{c}}$ is the predicted probability of being functional for chimera $\mathbf{c}$ from the logistic regression model. This objective is chosen to simultaneously find sequences that are informative and have a high probability of being functional. Since submodular functions are closed under positive linear combinations, the expected value of the Gaussian mutual information is also submodular and therefore greedy maximization provides strong performance guarantees. The covariance between sequences was calculated using the chimera-block coding scheme described in the regression analysis section (below). All experimental designs were performed with the Submodular Function Optimization Matlab Toolbox [154].

3.3.2 Gene synthesis and cloning

Genes encoding the SCHEMA designed arginase chimeras were synthesized from oligonucleotides as described previously [155]. In brief, long DNA oligonucleotides (99 bases) were synthesized in-house and assembled into two 560-base pair fragments using inside-out PCR. These primary fragments were combined without purification in a secondary overlap-extension reaction that formed the final desired 1086-base pair product. Custom software directed the assembly schemes and the efficient re-use of oligonucleotides across multiple related sequences. 32-base-pair overlaps were designed between adjacent oligonucleotides and a 35-base-pair overlap was designed between the two primary fragments. Genes were synthesized with an N-terminal 6x His tag followed by a tobacco etch virus protease cleavage site and NcoI and EcoRI restriction sites as described previously [149]. These genes were cloned into a pET28a expression vector and the sequences were verified using DNA sequencing.

3.3.3 Expression and Purification

E. coli BL21(DE3) cells expressing arginase variants were grown at 37 °C in minimal media to an OD600 of 0.8–1. Cells were collected by centrifugation, re-suspended in fresh minimal media containing 0.5 mM IPTG and 100 μM $MnSO_4$, and incubated for an additional 8–12 hours at 37 °C with shaking. After protein expression, cells were collected by centrifugation, lysed by a French pressure cell, and centrifuged at 14,000 g for 20 min at 4 °C. The clarified lysate was applied to a nickel IMAC column, washed with 10–20 column volumes of IMAC buffer and the purified arginases were eluted with IMAC elution buffer (50 mM $NaPO_4$, 250 mM imidazole, 300 mM NaCl, pH 8). The purified arginases were buffer exchanged several times into PBS, 10 % glycerol, pH 7.4 using a 10,000 MWCO centrifugal filter device (Amicon). Aliquots of purified arginase variants were then flash frozen in liquid nitrogen and stored at -80 °C.

3.3.4 Enzyme Kinetics

Michaelis-Menten kinetics for L-Arg hydrolysis were determined in 100 mM HEPES buffer at 37 °C, pH 7.4 as previously described [149].

3.3.5 Long-term stability

The long-term stability of the arginase chimeras was measured in 100 mM HEPES buffer, pH 7.4 at 37 °C, with or without 500 mM NaCl. Proteins were diluted to 2 μM with 100 mM HEPES, pH 7.4 and placed at 37 °C. Aliquots of 30 − 50 μL were taken at different time points (typically t = 0, 0.5, 3, 24, 48, and 72 hours). The activity at each time point was immediately measured using 1 mM L-Arg, as described previously [149]. The data were plotted as percent activity as a function of time, and the area under this inactivation curve (AUC) was calculated using Kaleidagraph.

3.3.6 Thermal stability

Arginase variants (20–40 μM) in PBS, pH 7.4 with or without EDTA (10 mM final concentration) were incubated in 96-well low-profile PCR plates (Fisher Scientific, Rockford, IL) on ice for 30 minutes. SYPRO orange dye (Sigma Aldrich) was added into each well immediately before placing the plate in an RT (reverse transcription)-PCR machine (LightCycler 480, Roche). The temperature dependence of protein unfolding was measured from 20–95 °C, and all chimeric arginases unfolded in a single cooperative transition. The thermal transition midpoint (T_m) was determined by fitting the unfolding curves to a modified logistic equation. All measurements were performed in at least duplicate.

3.3.7 Regression analysis

For both regression models, the independent variable corresponds to the sequence of chimeric arginases and is represented with a binary vector $\mathbf{x}$, where x_i indicates the parent identity at block i. Since we have limited data, we use Bayesian parameter estimation, which outperforms maximum likelihood estimation on small data sets.

A chimera's binary functional status was modeled with a Bayesian logistic regression model, which contains a Bernoulli likelihood function and a zero-mean, isotropic Gaussian prior on coefficients [156]. The resulting posterior distribution was approximated using Laplace's method and prior variance was estimated from the data by maximizing the marginal likelihood function. Using Newton's method, we found the maximum a posteriori (MAP) estimates for each chimera block's contribution to functionality. A chimera's probability of being functional was estimated by applying these MAP parameter estimates to the logistic model.

The logarithm of a chimera's long-term stability (AUC) was modeled with a Bayesian linear regression model, which consists of a Gaussian likelihood function with a zero-mean, isotropic Gaussian prior on coefficients [156]. The measurement noise and prior variance were estimated from the data by maximizing the marginal likelihood function. With these hyperparameters, MAP estimates for each block's contribution to long-term chimera stability were found in closed-form.

3.4 Results

3.4.1 SCHEMA library design

Human Arginase I (hArgI) and human Arginase II (hArgII) were chosen to be the parent sequences for the recombination library. These human isozymes, which share 61% sequence identity, were chosen in order to minimize the immunogenicity of the resulting chimeras, as required for therapeutic applications. The combinatorial library was chosen to have seven recombination sites (eight sequence blocks), resulting in a total of 256 (2^8) chimeric arginases. The SCHEMA disruption of a chimeric protein is based on a residue-residue contact map representation of a protein structure [73]. The trimeric structure of hArgII (PDB ID: 1PQ3) was used to prepare this contact map, which included both intra- and inter-subunit contacts. The RASPP algorithm was used to identify libraries that minimized the average SCHEMA disruption at various levels of mutation [157], and we selected a library that balanced deleterious interactions with sequence diversity. Based on the modest average SCHEMA disruption ($\langle E \rangle = 16$) and results from past libraries, we expected approximately half the chimeras within this library to be functional arginases. The sequences within this library were diverse: on average chimeras differed by 60 mutations (as low as 6 and as high as 120). These chimeras were also novel: the average mutational distance from the parental arginase sequences is 40.3.

The chimera blocks chosen for the arginase recombination library are illustrated in Figure 3.1. The arginase superfamily is characterized by a trimeric quaternary structure composed of 3-layer ($\alpha\beta\alpha$) sandwich subunits. Residues within chimera blocks 5, 6, and 8 form the trimer interface, and in each subunit's central parallel β-sheet seven of eight strands come from different blocks. Substrate recognition is achieved by several loops that flank the active site in addition to numerous water-mediated hydrogen bonds [158]. Within the library, each of these 'specificity' loops is located in different blocks. The residues that coordinate the catalytic binuclear manganese cluster are conserved in the parents, but the surrounding, second-shell residues come from chimera blocks 3, 4, 7, and 8. When these sequence fragments are shuffled within the library, new residue combinations will possibly contribute to the diversification of multiple arginase properties.

3.4.2 Generation of an informative set of chimeric arginases

Studying sequence-function relationships within recombination libraries using randomly selected chimeras requires constructing and characterizing a large number of sequences [86]. Systematically chosen chimera sets are more effective than randomly chosen ones, but still lack efficiency due to the significant proportion of nonfunctional sequences which provide little information about sequence properties [151]. Nonfunctional sequences can be avoided by one-factor-at-a-time experimental designs, which avoid disruptive interactions, but these designs result in highly skewed data sets [159]. Here, we present a two-step active learning algorithm that efficiently identifies an informative set of functional chimeras by first training a functional status classifier and then using this classifier to guide an experimental design.

The first step of the algorithm involves finding an informative set of chimeras for a logistic regression functional status model. Here, we chose a set of eight chimeras that maximized the mutual information between the chosen set and the remainder of the library (see Methods). These eight chimeras were synthesized and expressed (see Methods). As expected, only half produced functional arginases. With this functional status information, we trained a Bayesian logistic regression model to predict the probability a chimera will function for all sequences within the library.

The second step of the algorithm consists of finding a highly informative set of functional chimeric arginases. With the predictions from the logistic regression model, we can select sequences that maximize the expected value of the mutual information between the chosen set and the remainder of the library (see Methods). This objective was chosen to simultaneously find sequences that are both informative and have a high probability of being functional. We selected a set of four chimeras that maximized the expected value of the mutual information. These sequences were synthesized, expressed, and all yielded functional enzymes.

Our final data set (Table 3.1) consists of a highly informative set of 13 functional arginases (2 parents and 11 chimeras). Within this set of sequences, each parent at each block was typically observed multiple times, and 103 of the 112 possible block pairs are observed. Some blocks, such as block 4 parent 1, were under-represented because they contributed to loss of function and were therefore avoided in the second step of the sequence selection algorithm. Most importantly, the set comprised a full-rank parameter covariance matrix, allowing for accurate parameter estimation with linear regression. In the following section we use chimera-block linear regression to explore sequence-function relationships within the arginase library.

3.4.3 Regression model for long-term stability

Many of the chimeric arginases exhibited a biphasic loss of activity with time, which can be attributed to kinetic differences between the two metal-binding sites [149, 160]. To account for all enzymatically active forms of the arginases, we quantify the long-term stability as the area under the normalized inactivation curve (AUC). The long-term stability of all ten enzymes within the designed set of sequences was measured (see Methods) and is presented in Table 3.1. A Bayesian linear regression model was used to relate sequence fragments to the experimentally measured AUC (see Methods).

This model fits the experimental data well ($r = 0.97$) as seen in Figure 3.2*A*. The block regression parameters are given in Table 3.2. To validate the linear regression model, we designed two additional chimeric arginases (SCHEMA O and P) that were predicted to have enhanced long-term stability. These sequences were synthesized and characterized. The regression model showed good predictive ability (Figure 3.2*A*) and both sequences are more stable than 80% of the other chimeric arginases.

From the regression analysis, the most stabilizing sequence element is block 3, where substituting parent 1 for parent 2 is estimated to increase a chimera's AUC by 66%. Closer inspection of the amino acid sequences of this important chimera block revealed an abundance of charged residues, which led us to consider how a chimera's isoelectric point may contribute to long-term stability. The chimera's estimated isoelectric point [161, 162] shows a striking negative correlation ($r = -0.90,\ p < 0.001$) with AUC, Figure 3.2*B*. Here, chimeras with greatest net charge under the assay conditions (pH 7.4 and 37 °C) were the most stable while those closer to their isoelectric point exhibited faster inactivation.

3.4.4 Factors contributing to arginase inactivation

Using our diverse set of chimeric arginases, we explored how additional factors, besides isoelectric point, contribute to arginase inactivation under physiological conditions. The melting temperature for all sequences was measured (see Methods) and is shown in Table 3.1. These melting temperatures show no significant correlation with long-term stability ($r = -0.05,\ p = 0.87$). This suggests that arginase inactivation is minimally influenced by thermodynamic stability since the melting temperature of a protein is closely related to its Gibbs free energy of unfolding [163].

In another set of experiments, we measured the long-term stability of the chimeric arginases in the presence of excess manganese (500 μM $MnCl_2$) (Table 3.1). Surprisingly, this excess manganese typically increases a chimera's long-term stability more than twofold, indicating that arginases inactivate significantly more slowly under these conditions. This suggests that the manganese-free form of arginase is irreversibly inactivated, or on route to irreversibly inactivated states (see Discussion).

3.4.5 Correlation of relative metal affinity and catalytic efficiency

For all chimeric arginases, we performed Michaelis-Menten kinetic measurements (see Methods) and estimated their catalytic efficiency (k_{cat}/K_m) (Table 3.1). From these data, we found a compelling correlation between an arginase's relative metal affinity and its catalytic efficiency ($r = 0.82$ and $p = 0.015$), where enzymes with low metal affinity tend to have the greatest catalytic efficiency for L-Arg hydrolysis (Figure 3.3). Here, we estimate the relative metal affinity of arginases as the difference between the long-term stability in the presence of excess manganese (AUC_{Mn}) and the long-term stability at physiological conditions (AUC). Arginases with a high affinity for metal will have a small difference ($AUC_{\mathrm{Mn}} - AUC$) because excess manganese has little affect on their inactivation rate. A similar correlation has been observed within a set of Cu^{2+} complexes [164]. In this study, the authors found the stability of a Cu^{2+} complex to be inversely related to its rate of glycine methyl ester hydrolysis, indicating that more stable complexes lower the Lewis acidity of the Cu^{2+} ion. Likewise, arginases that bind Mn^{2+} more tightly may have reduced Lewis acidity in coordinating substrate or water ligands, and therefore diminished catalytic efficiency.

3.5 Discussion

The combination of structure-guided SCHEMA recombination and an efficient active learning procedure has generated a highly informative set of chimeric arginases. The high level of sequence diversity within this set of sequences translates into functional diversity: many of the measured properties are outside the range displayed by the two parents (Table 3.1). Site-directed recombination libraries provide unique data sets for studying sequence-function relationships, offering distinct advantages over sets of point mutants or naturally existing proteins. The effects of point mutations are frequently too small to resolve experimentally, and the large numbers of neutral mutations in naturally existing proteins make it difficult to pinpoint the basis of functional differences. In contrast, libraries of chimeric proteins contain an intermediate level of sequence diversity and mutational changes are observed in multiple sequence backgrounds. Additionally, the additive structure of the recombinational landscape allows linear regression models to efficiently identify sequence features of interest.

Within this data set, a linear regression model helped identify the strong negative relationship between a chimeric arginase's isoelectric point and its long-term stability ($r = -0.90,\ p < 0.001$). Recombination of hArgI and hArgII ($pI = 6.8$ and 5.7, respectively) generated a set of functional chimeras with isoelectric points ranging from 5.5 to 7.5. Since the long-term stability experiments were performed at physiological pH (7.4), chimeras with the greatest net charge (low pI) displayed the greatest stability. This relationship between a protein's net charge and its stability has been observed numerous times. A large survey across multiple protein families found many proteins to be less stable near their isoelectric point [165]. Similarly, engineered ribonuclease variants show decreased solubility and increased aggregation near their isoelectric point [166, 167]. Our proposed mechanism of arginase inactivation at physiological conditions is depicted in Figure 3.4. Based on the correlation between a chimera's isoelectric point and long-term stability (Figure 3.2*B*) and the frequently observed aggregation during purification and characterization (not shown), we believe that the irreversible loss of catalytically active arginase at physiological conditions is dominated by protein aggregation, rather than thermodynamic stability. More negatively charged arginases resist the tendency to aggregate by electrostatic repulsion, providing longer enzyme lifetimes. Additionally, we find the rate of this aggregation process to be strongly metal-dependent: excess manganese typically increases the long-term stability more than twofold. This suggests the metal-free arginases are more prone to aggregation and therefore are the primary route to inactivation. Based on this mechanism of inactivation, future arginase engineering efforts should be focused on increasing metal affinity and preventing aggregation. However, increasing an arginase's metal affinity may come at the cost of reduced catalytic efficiency as described above (Figure 3.3). 'Supercharging' proteins by replacing numerous surface residues with charged residues [168], is an engineering method for preventing aggregation and may provide a simple strategy for designing arginases with exceptional long-term serum stability.

Hyperargininemia patients have elevated serum L-Arg levels that range from 600–900 μM [169], in contrast to normal reference values of 50–150 μM [170]. Arginase replacement therapy is a potential treatment modality for patients suffering from the neurotoxicities associated with hArgI deficiencies. Since hArgI has been under investigation as an antineoplastic agent, its serum retention time has been pharmacologically optimized via PEGylation, resulting in dose dependent L-Arg depletion in rats for up to days at a time [171]. However, the desired kinetic parameters for treating cancer (very low L-Arg) are different than those required for treating hyperargininemia patients, where the goal is long-term reduction of L-Arg to non-pathological levels. For hyperargininemia treatment, the most crucial pharmacological parameter is an arginase's long-term serum stability, which determines the patient's dosage intervals. The SCHEMA K variant, identified in this study, has a stable linear decay rate of only 1% per hour when loaded with Mn^{2+}, which is ideal for introducing long-term basal arginase activity. A simple model suggests that a single dose of SCHEMA K could maintain L-Arg levels in hyperargininemia patients within the normal range for five days longer than a single dose of the hyperactive (but less stable) Co^{2+}-loaded hArgI.

The ability to design enzymes that are customized to specific reaction conditions is of significant interest to biomedical science. SCHEMA recombination provides a diverse sampling of the protein fitness landscape, revealing features that cannot be observed by traditional biochemical methods. These data sets provide a unique opportunity to explore the relationships between protein sequence and protein function, providing principles that can be used to engineer highly-optimized protein sequences.

3.6 Figures and Tables

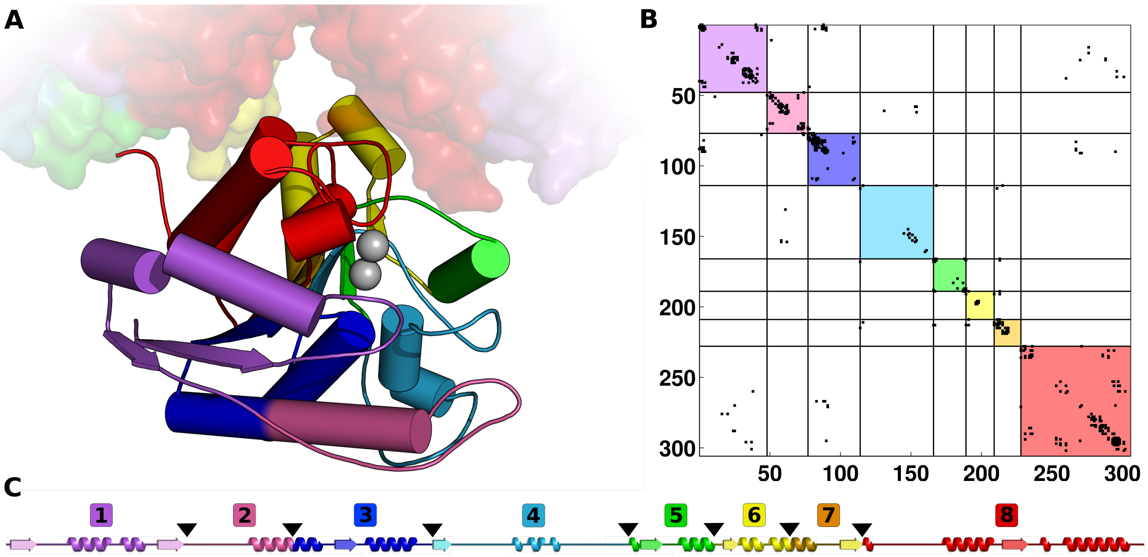


Figure 3.1: Arginase chimera library block boundaries. (*A*) Arginase three-dimensional structure with SCHEMA blocks represented by different colors. The trimer interface is shown as a transparent surface. (*B*) Contact map displaying residue-residue contacts that could be broken upon recombination. The colored squares correspond to the block divisions of the library. (*C*) Secondary structure diagram showing the chimera library block divisions.

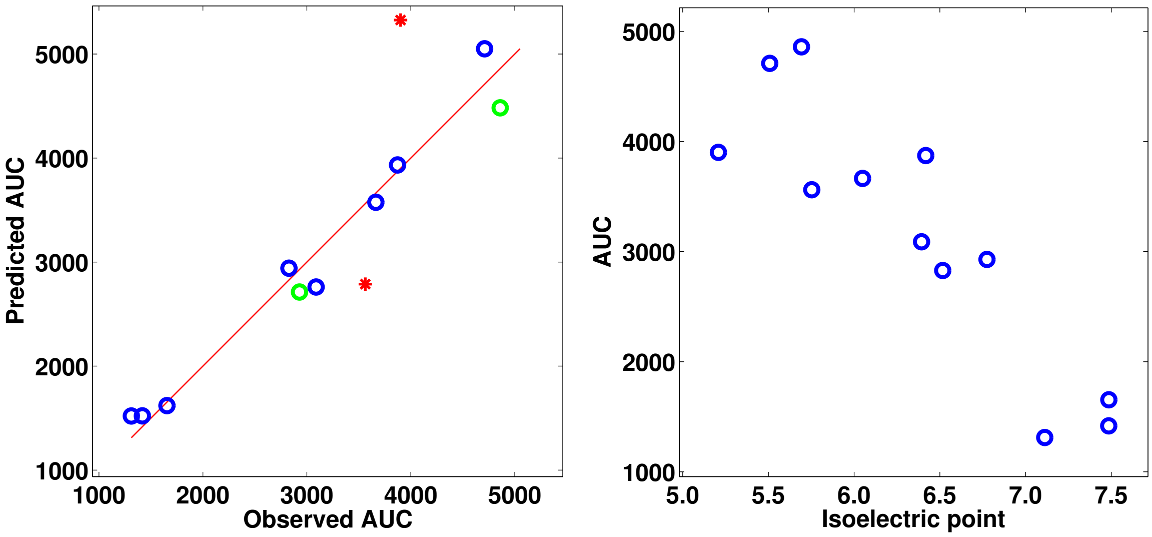


Figure 3.2: Arginase long-term stability. (A) Bayesian linear regression model for AUC. Green and blue circles correspond to the parents and chimeras (respectively) within the initial data set ($r = 0.97$ and $p < 0.001$). Red stars represent the model's predictions on the validation set. (B) Correlation between isoelectric point and AUC ($r = -0.90$ and $p < 0.001$).

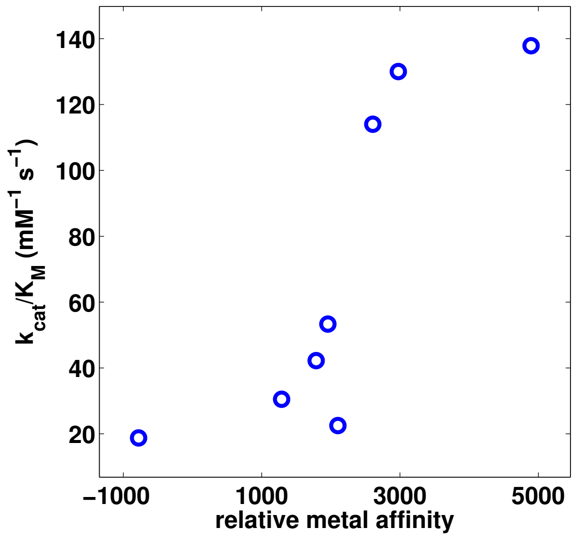


Figure 3.3: Correlation between relative metal affinity ($AUC_{\text{Mn}} - AUC$) and catalytic efficiency (k_{cat}/K_m), $r = 0.81$ and $p = 0.015$.

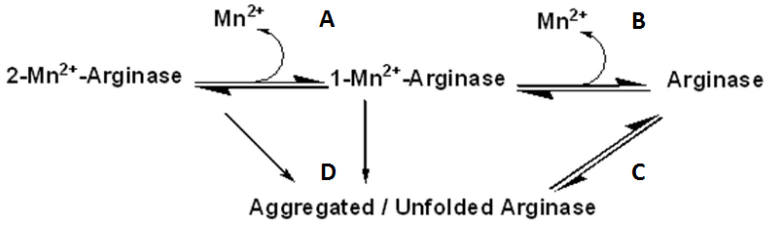


Figure 3.4: Schematic of potential arginase inactivation mechanisms. A) Loss of first equivalent of bound metal and decrease of some activity, B) loss of second equivalent of bound metal and loss of all activity, C) equilibrium between folded and unfolded states, D) irreversible precipitation/aggregation.

	Chimera				
Name	Blocks	*AUC*	AUC_{Mn}	T_m °C	k_{cat}/K_m mM^{-1}s^{-1}
hArgI	11111111	2929	5326	81.0	130 ± 20
hArgII	22222222	4042		80.6	114 ± 18
SCHEMA A*	11112122				
SCHEMA B	12122211	3664	2927	81.2	19 ± 7
SCHEMA C	11221221	3872	5838	68.1	53 ± 10
SCHEMA D*	12211212				
SCHEMA E*	21121221				
SCHEMA F*	21212211				
SCHEMA G*	22111221				
SCHEMA H	22221111	3089	4109	68.5	31 ± 7
SCHEMA I	11122222	994	5890	82.5	138 ± 19
SCHEMA J	21122121	1654		67.5	27 ± 11
SCHEMA K	11222112	4710	6188	70.7	42 ± 8
SCHEMA L	22121121	1311	3408	78.9	23 ± 10
SCHEMA M	21222111	2828		70.5	19 ± 7
SCHEMA N	21121111	1417		74.3	39 ± 10
SCHEMA O	12122122	4005		71.2	45 ± 11
SCHEMA P	12222112	3901		70.6	36 ± 5

Table 3.1: Chimeric arginase data. SCHEMA A-N comprise the designed set of highly informative sequences, and SCHEMA O and P are the sequences used to validate the regression model. Asterisked sequences had no detectable protein expression.

parameter name	$\log(AUC)$
Reference	7.96
Block 1	-0.27
Block 2	0.12
Block 3	**0.50**
Block 4	-0.38
Block 5	0.19
Block 6	0.31
Block 7	-0.14
Block 8	0.18

Table 3.2: Parameters for the arginase long-term stability regression model. The parameters represent the effect of substituting parent 2 for parent 1 at a given block on the logarithm of the AUC. The most significant substitution occurs at block 3, which is shown in bold.

Chapter 4

Random field model of the protein recombinational landscape

4.1 Abstract

Intragenic recombination events contribute to the evolution of natural genomes, yet it is unclear what advantage this molecular diversification mechanism provides over mutation. Experimental results from libraries of proteins made by recombination have revealed the extreme tolerance of proteins to recombination with homologous sequences and that sequence fragments make largely additive contributions to a protein's biophysical properties. Here, we develop a random field model to describe the statistical features of the subset of protein space accessible by recombination, which we refer to as the recombinational landscape. This model shows quantitative agreement with experimental results compiled from nine recombination libraries. We use the random field model to understand the origin of a protein's tolerance to recombination and the additive effects of sequence fragments. The results reveal a recombinational landscape that is enriched in functional sequences, with properties dominated by a large-scale additive structure. Intragenic recombination explores a unique subset of sequence space that may promote rapid molecular adaptation.

4.2 Introduction

The ubiquity of sex and recombination suggests these mechanisms must play a significant role in evolution, yet their benefit is still highly debated [172, 173]. Intragenic recombination events generate chimeric proteins, which are believed to make important contributions to allelic diversity in natural populations [174, 175, 176, 177]. In the laboratory, the benefits of intragenic recombination of homologous proteins have been clear: it provides a powerful method for engineering new proteins that are functionally diverse while still having a high probability of functioning [75, 178]. The resulting chimeric proteins have been useful for understanding protein thermostability [179, 151], divergence of enzymatic function [180], and cellular signaling responses [181], shedding light on how these properties may evolve in Nature and how best to engineer them in the laboratory.

Previously, we have developed techniques for the design, construction, and characterization of large libraries of chimeric proteins [73, 157, 182]. Briefly, libraries are designed (crossover sites are selected) to minimize the number of novel residue contacts generated upon recombination (SCHEMA disruption), which tend to be deleterious to protein function. To date, the Arnold lab has tested nine such site-directed recombination libraries: a library of chimeric bacterial β-lactamases (βlac13 and βlac), bacterial cytochrome P450s (P450), bacterial family 9 cellulases (Cel9), fungal family 5 cellulases (Cel5), bacterial family 48 cellulases (Cel48), fungal class I cellobiohydrolases (CBHI), fungal class II cellobiohydrolases (CBHII), and human arginases (Arg) (Table 4.1). Each library, which typically consists of thousands of new sequences, provides a glimpse of the protein fitness landscape and can reveal important aspects of its structure. Here, we refer to the subset of the protein fitness landscape that is accessible by recombination as the recombinational landscape.

These recombination libraries have highlighted the enrichment of functional sequences within the recombinational landscape: most libraries contain a significant proportion (~20–50%) of functional sequences. For comparison, random mutation libraries with the same number of mutations are estimated to contain 10–20 orders of magnitude fewer functional sequences [37, 36, 43]. The accumulation of random mutations decreases the probability of functional sequences exponentially, whereas mutations generated by recombination always move towards other functional sequences and are therefore significantly more conservative [56]. For this reason, intragenic recombination effectively explores a functional ridge through the mostly nonfunctional protein sequence space.

These libraries have also revealed significant variation in thermostability [86, 151] and other properties [85] within the recombinational landscape. We have observed that a majority of this variation can be explained by additive effects [86, 151, 159]. This additivity has been used to efficiently engineer highly-optimized chimeric proteins for a variety of applications. The additive structure, or lack of epistasis, within the recombinational landscape may contribute to the ability of intragenic recombination to traverse the protein fitness landscape.

We would like to understand the features of the recombinational landscape that contribute to its extreme enrichment in functional sequences and its additive structure. Since the details of the true protein recombinational landscape are unknown, we develop a random field model which captures its statistical properties. Random field models are effective at describing statistical features of uncertain, spatially-organized functions, with applications ranging from geostatistics to image analysis [183, 184, 185]. This versatile class of models has also been used to describe fitness landscapes [186], the best known example being Kauffman's NK-model [187]. Our random field model for the recombinational landscape is based on a physics-inspired energy function and parametrized with experimental data. Using this model, we derive approximations for the expected value of the proportion of functional sequences within a recombination library and the degree of landscape additivity, and relate these quantities back to experimental observations. We discuss how the structure of the recombinational landscape may contribute to the utility of intragenic recombination as a evolutionary mechanism and the implications for the design of future recombination libraries.

4.3 Results

4.3.1 Random field model for the protein recombinational landscape

Consider a pairwise, residue-level energy function to describe the large number of intramolecular interactions that stabilize protein structures. These simplified contact potentials have been used in the past for protein folding simulations and structure prediction [188, 189, 190]. Assuming a fixed structure (set of contacts), the energy of any sequence can be determined from the sum of energy terms associated with the sequence's specific residue combinations at every pair of contacting residues. For chimeric proteins we distinguish between two types of contacts: parental (P) contacts, which are residue pairs observed in at least one of the parents, and novel (N) contacts, which are not. With this model, the energy of any chimeric protein **c** is given by

$$E_{\mathbf{c}} = \sum_i a^i_{\mathbf{c},P}\ \varepsilon^i_P + \sum_i a^i_{\mathbf{c},N}\ \varepsilon^i_N, \tag{4.1}$$

where ε^i_P is the energy term associated with parental contact i, ε^i_N is the energy term associated with novel contact i, and $a^i_{\mathbf{c},P}$ and $a^i_{\mathbf{c},N}$ are binary variables which indicate the specific residue pairs for each contact i in chimeric protein **c**. Since the specific values of ε^i_P and ε^i_N are unknown, we define the independent and identically distributed random numbers P_i and N_i, distributed with means and variances

$$P_i \sim \mu_P, \sigma^2_P \tag{4.2}$$

$$N_i \sim \mu_N, \sigma^2_N. \tag{4.3}$$

Substituting these random variables into equation 4.1 defines a random energy function associated with any chimeric protein **c**

$$\mathcal{E}_{\mathbf{c}} = \sum_i a^i_{\mathbf{c},P}\ P_i + \sum_i a^i_{\mathbf{c},N}\ N_i. \tag{4.4}$$

This random energy function is defined over the parental subspace $\mathbb{S}_p$, the set of all sequences that can be generated by recombining the parent sequences, which specifies the random field

$$\{\mathcal{E}_{\mathbf{c}} : \mathbf{c} \in \mathbb{S}_p\}. \tag{4.5}$$

The expected value of the random field at chimeric protein **c** is

$$\mathrm{E}[\mathcal{E}_{\mathbf{c}}] = \mu_P \sum_i a^i_{\mathbf{c},P} + \mu_N \sum_i a^i_{\mathbf{c},N}, \tag{4.6}$$

and the covariance between any two sequences is

$$\mathrm{Cov}[\mathcal{E}_{\mathbf{c1}}, \mathcal{E}_{\mathbf{c2}}] = \sigma_P^2 \sum_i a^i_{\mathbf{c1},P}\ a^i_{\mathbf{c2},P} + \sigma_N^2 \sum_i a^i_{\mathbf{c1},N}\ a^i_{\mathbf{c2},N}. \tag{4.7}$$

This random field model provides a statistical description of the recombinational landscape. Most importantly, the covariance structure captures our intuitive notion of protein similarity: proteins with similar sequences have similar structures and therefore similar properties.

To parametrize the random field model, we must determine the mean energy μ_P and variance σ_P^2 of parental contacts and the equivalent parameters μ_N and σ_N^2 for novel contacts. Using a large binary functional status data set from the cytochrome P450 recombination library [80], these four parameters were estimated by maximizing a marginalized likelihood function (see Methods). If we assume the functional status data depends on a sequence's Gibbs free energy difference from the nonfunctional state, these estimated parameters can be interpreted as Gibbs free energy differences (in arbitrary units), see Supplementary Material. As expected, parental contacts are slightly stabilizing (μ_P = -0.66 AU), novel contacts are significantly destabilizing (μ_N = 52.06 AU), and both classes of contacts show significant variation relative to their means (σ_P = 51.94 AU and σ_N = 58.33 AU). Estimating these parameters on recombination data from other protein families yields qualitatively similar results (Supplementary Material). This is not surprising considering that most proteins are marginally stable [33] and mutations (novel contacts) tend to be deleterious to protein function [37, 36, 43]. In the following sections, this parametrized random field model is used to interpret experimental observations from protein recombination libraries.

4.3.2 Effect of homologous substitutions on protein function

Previously, we compared the effects of random versus homologous amino acid substitutions [56]. Whereas the fraction of functional sequences declines exponentially with increasing random mutations [37, 36], that fraction varies log-parabolically with the number of substitutions taken from another functional parent. For two parents, the log-parabolic behavior appears because accumulating homologous substitutions must eventually convert one functional parent sequence into another functional parent sequence. Random mutagenesis carried out on β-lactamase measured a probability that a single random mutation will preserve function (neutrality) of ∼0.54, whereas recombination experiments on the same enzyme indicated the probability a homologous substitution will preserve function (recombinational tolerance) is ∼0.79 [56]. Considering the relatively high recombinational tolerance and the log-parabolic dependence, homologous substitutions are significantly more conservative than random mutations. Here, we evaluate the effects of homologous substitutions using the random field model and compare the results to our previous analysis.

Analyzing the two-parent, thirteen-crossover β-lactamase library (βlac13) [191], the probability of functioning for each chimera was estimated by evaluating the logistic function at the expected value of the random field. These probabilities were averaged within 15 groups binned by the number of homologous substitutions. The same analysis was also performed on simulated random substitutions, where a novel contact was any residue pair not seen in the two β-lactamase parents. After reanalyzing the chimeric β-lactamase data to account for library construction errors (see Methods), the random field model shows excellent agreement with the experimental substitutions generated by recombination and randomly (Figure 4.1*A*). As observed previously, the fraction of functional sequences shows a steep exponential decline with random mutations, while functionality displays a log-parabolic dependence with homologous substitutions.

With the random field model, we can now explore the influence of recombination parameters such as parent sequence identity and the number of sequence crossovers on the shape of the recombination curve shown in Figure 4.1*A*. As parent sequence identity decreases, the curve stretches to a higher level of mutation and to a lower fraction functional (Figure 4.1*B*), as was shown previously using lattice protein simulations [56]. As the number of sequence crossovers decreases, the log-parabolic curve shifts towards a higher fraction functional (Figure 4.1*C*), necessarily approaching a flat line when there are no crossovers. This highlights an improvement to our previous analysis because it shows how the estimated recombinational tolerance depends on the number of sequence crossovers.

To estimate the effects of homologous substitutions, independent of the number of crossovers, we sampled random homologous substitutions and calculated the average probability of folding at each level of mutation (Figure 4.1C). The effects of random homologous substitutions still follow the log-parabolic curve, although this curve dips over five orders of magnitude lower than the experimentally characterized β-lactamase library [191]. Fitting the log-parabolic equation [56], we estimate the recombinational tolerance of random homologous substitutions to be $\rho = 0.68 \pm 0.01$. The recombinational tolerance is still greater than the neutrality, and thus homologous substitutions are still more conservative than random substitutions, but to a lesser degree than previously estimated. From this analysis, we propose an updated model for the conservative nature of intragenic recombination which includes contributions from homologous substitutions (as shown previously) as well as groups of coevolved residues varying simultaneously. The latter effect is expected to play a major role in natural evolution where the number of intragenic crossover events per generation is likely to be small.

Surprisingly, the random field model for the recombinational landscape also works reasonably well to describe the effects of random mutations. With this model, random mutations will frequently result in a non-parental amino acid and therefore cause deleterious novel interactions with all contacting residues. This simplified model recapitulates the exponential decline in functional sequences which was observed with the experimental β-lactamase data (Figure 4.1A) and other mutational studies [37, 36, 43]. In addition, this model trivially captures the well-known fact that surface mutations tend to be less deleterious than mutations in the protein core, because core residues tend to have many more contacts. With a single model to explain the effects of both random and homologous substitutions, we can understand their differences in terms of residue contacts. The number of deleterious contacts generated by a homologous substitution is less than or equal to the number generated by a random mutation at the same position, with equality rarely being achieved. This explanation is consistent with hypothesis that homologous substitutions are conservative because they have been previously selected to be compatible with the protein fold [56].

4.3.3 Effect of intragenic recombination across protein families

The factors that determine a protein family's tolerance to recombination events are unknown. Table 4.1 reports the fraction of functional sequences compiled from nine recombination libraries, representing protein families of different functions, sizes, and fold classes. Eight of these libraries were designed with the intent of maximizing the fraction of functional sequences, yet there is significant variation (2–3 fold) in this fraction between libraries. While some of this variation is likely due to experimental differences in classifying functional sequences from different enzyme classes, we expect a significant proportion of this variation to arise from differences in parent fold, parent sequence identity, and the specific crossover locations chosen in the library design. Using the random field model, we derive an approximation for the expected value of the fraction of functional sequences within a recombination library and use this to understand the factors that contribute to a protein family's tolerance to recombination.

Consider a recombination library L, generated by recombining sequence fragments from p parental sequences at n crossover sites. We refer to the sequence fragments between crossover sites as 'blocks'; therefore the library is composed of b sequence blocks ($b = n + 1$). Assume a Gaussian distribution of sequence energies within the library, a good approximation when the library size (p^b) is greater than a few hundred. Here, the distribution of sequence energies within recombination library L can be described by its mean

$$M_L = \frac{1}{p^b} \sum_{\mathbf{c} \in L} E_{\mathbf{c}} \tag{4.8}$$

and variance

$$V_L = \frac{1}{p^b} \sum_{\mathbf{c} \in L} (E_{\mathbf{c}} - M_L)^2. \tag{4.9}$$

The fraction of functional sequences within library L is given by evaluating the Gaussian cumulative distribution function at zero, that is, the fraction of sequences having an energy less than 0.

Since the specific energy terms that shape the recombinational landscape are unknown, we use the random field model to calculate the expected value of the fraction of functional sequences by integrating over all possible energy terms ε_P^i and ε_N^i. The expected value of the library mean is given by

$$\mathrm{E}[M_L] = \frac{1}{p^b} \sum_{\mathbf{c} \in L} \mathrm{E}[\mathcal{E}_\mathbf{c}] = \mu_P n_C + (\mu_N - \mu_P) \frac{\sum_\mathbf{c} n_{N,\mathbf{c}}}{p_b} \tag{4.10}$$

where n_C is the total number of contacts and $n_{N,\mathbf{c}}$ is the number of novel contacts in chimera $\mathbf{c}$. The expected value of the library variance is given by

$$\mathrm{E}[V_L] = \frac{1}{p^b} \sum_{\mathbf{c} \in L} \left[(\mathrm{E}[\mathcal{E}_\mathbf{c}] - \mathrm{E}[M_L])^2 + \mathrm{Var}[\mathcal{E}_\mathbf{c}] + \mathrm{Var}[M_L] - 2\,\mathrm{Cov}[\mathcal{E}_\mathbf{c}, M_L] \right] \tag{4.11}$$

More specific details of $\mathrm{Var}[\mathcal{E}_\mathbf{c}]$, $\mathrm{Var}[M_L]$, and $\mathrm{Cov}[\mathcal{E}_\mathbf{c}, M_L]$ are given in the Supplementary Material. With these two expectations, the expected value of the fraction of functional sequences can be approximated with a Taylor series expansion

$$\mathrm{E}[f_F] \approx \frac{1}{2} \left[1 - \mathrm{erf}\left(\frac{\mathrm{E}[M_L]}{\sqrt{2\,\mathrm{E}[V_L]}} \right) \right], \tag{4.12}$$

where erf is the error function and the details of this Taylor series are given in the Supplementary Material. All expectations and approximations were verified with extensive Monte Carlo sampling.

The expected value of the fraction of functional sequences within a library $\mathrm{E}[f_F]$ shows quantitative agreement with the experimentally determined values, as shown in Figure 4.2*A*. The Cel9 library is an outlier, which may be the result of the extreme electrostatic clashes that are present in a significant proportion of the library (discussed in [192]). Within the random field model, both parental and novel contacts contribute to the distribution of sequence energies within a recombination library and therefore the fraction of functional sequences. The highly deleterious novel contacts dictate the mean of the library, while parental contacts, which typically outnumber novel contacts 50–100-fold, dominate the library variance. This suggests recombination events can cause loss of function by two independent mechanisms: (1) by introducing new deleterious interactions between sequence fragments, or (2) by introducing sequence fragments which already contain deleterious interactions.

To better understand the variation in the fraction of functional sequences between the nine recombination libraries, we sampled random libraries, calculated $\mathrm{E}[f_F]$, and estimated the contribution from protein fold, specific breakpoints, and parent sequence identity. For each protein fold, we sampled 100 random two-parent sequence alignments with sequence identity ranging from 10–90%, and for each of these alignments we sampled 100 random 7-crossover libraries, for a total of 90,000 libraries. A three-way analysis of variance shows the protein fold ($p < 0.001$), specific breakpoints ($p < 0.001$), and parent sequence identity ($p < 0.001$) all make significant contributions to the $\mathrm{E}[f_F]$. Estimating the variance components, we find parent sequence identity to be the main determinant of $\mathrm{E}[f_F]$ (contributing 92% of the variance), followed by specific breakpoints (4%), and protein fold (2%). This strong dependence on parent sequence identity is the result of the approximately exponential increase in the number of deleterious novel contacts as parent sequences diverge. Interestingly, when the parent sequence identity is low, most of the nonfunctional chimeric proteins are the result of inactivation mechanism (1), but when the parent sequence identity is high, nonfunctional sequences are usually the result of inactivation mechanism (2). This is consistent with the observation of high mutual information between a chimeric protein's functional status and its number of novel contacts for the β-lactamase library (low parent sequence identity) and the low mutual information observed for the P450 library (high parent sequence identity) [193].

Using the random field model, we would like to evaluate the effectiveness of the library design objective (to minimize the average number of novel contacts $\langle E_{sch} \rangle$) in producing libraries of functional chimeric proteins. At a fixed library variance, the model shows a super-exponential decline in the fraction of functional sequences with increasing $\langle E_{sch} \rangle$, similar to the trends seen in random mutation libraries with an increasing number of mutations [37, 36, 43]. Since the entire range of $\langle E_{sch} \rangle$ is not accessible for a particular set of parent sequences, we would like to understand the effect of minimizing $\langle E_{sch} \rangle$ on f_F over randomly generated libraries with the same parents, number of sequence blocks, and block size distribution. For all nine recombination libraries, we generated 1000 random libraries with similar features and estimated the fraction of functional sequences within each. The model estimates a 1.3–5.8-fold increase in the fraction of functional sequences from $\langle E_{sch} \rangle$-minimized libraries over average (2–19-fold increase over the least functional libraries). As expected, the largest benefit is seen for libraries with more diverged parents (higher $\langle E_{sch} \rangle$), with diminishing returns as the parents become more similar (as $\langle E_{sch} \rangle$ becomes lower). As $\langle E_{sch} \rangle$ goes to zero, the model predicts some non-zero fraction of nonfunctional sequences will remain, typically 10–30%. These nonfunctional sequences are present because some recombination events combine multiple deleterious parental contacts within a single protein.

4.3.4 Additive structure of recombinational landscape

Perhaps the most surprising finding from these recombination experiments has been the additive structure of the recombinational landscape [86, 151, 159]. Linear models are able to explain a majority of variation in thermostability and other properties, suggesting that sequence elements make largely independent, additive contributions to a protein's overall measured properties. In quantitative genetics, this is referred to as additive genetic variance, which according to Fisher's fundamental theorem of natural selection determines a population's response to selection [194, 195]. Additive landscapes are easier for evolving populations to climb because they are not stymied by rugged, epistatic features. This additivity has been especially useful for engineering optimized chimeric proteins because a small sampling of sequences provides sufficient information to make accurate predictions across the entire library [86, 151, 159]. Here, we use the random field model to understand the origin of the additive structure within the recombinational landscape.

Within the recombination library L described in the previous section, the total variance can be partitioned into additive and epistatic components ($V_L = V_A + V_E$). We define the landscape's degree of additivity A as the fraction of the total variance that is explained by additive effects

$$A \equiv \frac{V_A}{V_L}. \tag{4.13}$$

This dimensionless quantity, which ranges from 0 to 1, describes the smoothness of the landscape, and is inversely related to the landscape 'ruggedness' defined in [196]. For four of the recombination libraries, there is sufficient data to calculate the additivity of the thermostability landscape (see Methods) and the results are presented in Table 4.1.

The expected value of a library's additive variance $\mathrm{E}[V_A]$ can be found in a similar way as the total variance (previous section), but only considering the contributions of additive effects (Supplementary Material). The expected value of the additivity can be approximated with a Taylor series expansion about $\mathrm{E}[V_A]$ and $\mathrm{E}[V_L]$

$$\mathrm{E}[A] \approx \frac{\mathrm{E}[V_A]}{\mathrm{E}[V_L]}. \tag{4.14}$$

All expectations and approximations were verified with Monte Carlo sampling.

The expected value of the landscape additivity $\mathrm{E}[A]$ shows close agreement with the experimentally determined values (Figure 4.2*B*). While the correlation is not statistically significant, due to the limited data, all the $\mathrm{E}[A]$s are large and within the experimentally observed ranges. In addition, the five uncharacterized libraries also have large expected additivities (βlac13 = 0.44, βlac = 0.67, Cel5 = 0.65, Cel9 = 0.94, Arg = 0.82), suggesting this additive structure within the recombinational landscape may be quite general. Despite being generated by a purely pairwise energy function, which is by definition epistatic, a majority of the variation within these recombination libraries can be explained by additive effects. This surprising result can be attributed to two factors: sequence conservation among the parents and the partitioning of interactions into structural modules. Nonlinear interactions that are conserved among all parents will not contribute to the variation of any property within the library, and those interactions conserved among some parents will only make minor contributions. Nonlinear interactions that are partitioned into structural modules will vary together, and therefore contribute to only additive variation.

Coincidentally, the recombination library design objective, to minimize a library's average SCHEMA disruption $\langle E_{sch} \rangle$, accounts for sequence conservation and attempts to partition interactions into structural modules. In fact, the interactions that contribute to epistatic variation are the exact same interactions counted in SCHEMA disruption, therefore the library design has the effect of maximizing $\mathrm{E}[A]$. However, sampling 1000 randomly generated libraries with the same parents, number of sequence blocks, and block size distribution shows minimizing $\langle E_{sch} \rangle$ only increases the additivity modestly (0.05–0.2) over average. Therefore we conclude the additive structure observed within recombination libraries is a general result of dividing the sequence into fragments, rather than the specific crossover sites chosen for the library design.

The additivity exhibited by the random field model does not hold for chimeric proteins that adopt alternate structures (as described by a contact map). For example, nonfunctional sequences, which account for a significant proportion of chimeras, will clearly not display additivity in properties involving protein function. For many properties, such as thermostability (loss of enzymatic function at elevated temperatures), we have observed additivity because the experimental measurements require enzymatic activity, which greatly increases the likelihood that a set of enzymes will adopt the same, or very similar structures. The subset sequences that adopt the same structure is referred to as a neutral network [197, 35] and this may define the domain of the additivity within the recombinational landscape.

4.4 Methods

4.4.1 Compiling the chimeric protein data set

The residue-residue contact map for each library was determined by identifying all protein chains within the Protein Data Bank that share at least 50% sequence identity with any parent. Also included were three unpublished P450 structures and two unpublished Cel5 structures, for a total of 88 βlac13, 173 βlac, 91 P450, 39 CBHI, 24 CBHII, 6 Cel5, 16 Cel9, 21 Cel48, and 143 arginase chains. For each chain, a residue pair was considered contacting if they contained any heavy atoms within 4.5 Å. The final contact map for each library is composed of residue pairs that are contacting in more than 50% of all chains.

The number of functional and nonfunctional chimeric proteins was retrieved from previously published results: βlac13 [191], βlac [198], P450 [80], CBHI [159], CBHII [151], Cel5 [199], Cel9 [192], Cel48 [200], Arg [201]. The fraction of functional chimeras was estimated using maximum likelihood and 95% confidence intervals were calculated using the Clopper-Pearson method [202]. We could not accurately estimate the fraction of functional sequences for the CBHI library due to the extreme bias in the chimera sampling [159]. The results from the βlac13 library were reanalyzed to account for library construction errors (see below).

The additivity of the P450, CBHI, CBHII, and Cel48 libraries was calculated using published thermostability data [86, 159, 151, 200]. For each library, a block-based linear regression model [86] was parametrized on all the available data. The resulting predictions are unbiased, so the total variance can be partitioned into explained and residual components. The ratio of the explained variance to total variance is the additivity A, and in this case is identical to the coefficient of determination R^2.

4.4.2 Estimation of parental and novel contact parameters

Given a data set which maps contact information to binary functional status, we want to estimate the mean energy μ_P and variance σ_P^2 of parental contacts and the mean energy μ_N and variance σ_N^2 for novel contacts. The true energy terms ε_P^i and ε_N^i can be integrated out to give the marginalized likelihood function

$$p(\mathbf{y}|\mathbf{A}, \mu_P, \sigma_P^2, \mu_N, \sigma_N^2) = \iint p(\mathbf{y}|\mathbf{A}, \boldsymbol{\varepsilon}_P, \boldsymbol{\varepsilon}_N)\ p(\boldsymbol{\varepsilon}_P|\mu_P, \sigma_P^2)\ p(\boldsymbol{\varepsilon}_N|\mu_N, \sigma_N^2)\ d\boldsymbol{\varepsilon}_P d\boldsymbol{\varepsilon}_N, \tag{4.15}$$

where $\mathbf{y}$ is the binary functional status and for notational simplicity all parental energy terms ε_P^i are combined in the vector $\boldsymbol{\varepsilon}_P$, all novel energy terms ε_N^i are combined in the vector $\boldsymbol{\varepsilon}_N$, and all binary indicator variables ($a_{\mathbf{c},P}^i$ and $a_{\mathbf{c},N}^i$) are combined into the matrix $\mathbf{A}$. The mean and variance of parental and novel contacts can be estimated by maximizing this marginalized likelihood function.

Since $\mathbf{y}$ is composed of binary data, the first term in the integrand is given by the logistic likelihood function

$$p(\mathbf{y}|\mathbf{A}, \boldsymbol{\varepsilon}_P, \boldsymbol{\varepsilon}_N) = \prod_{\mathbf{c}} s\left(\mathbf{a}_{\mathbf{c},P} \cdot \boldsymbol{\varepsilon}_P + \mathbf{a}_{\mathbf{c},N} \cdot \boldsymbol{\varepsilon}_N\right)^{y_{\mathbf{c}}} s\left(-\mathbf{a}_{\mathbf{c},P} \cdot \boldsymbol{\varepsilon}_P - \mathbf{a}_{\mathbf{c},N} \cdot \boldsymbol{\varepsilon}_N\right)^{1-y_{\mathbf{c}}}, \tag{4.16}$$

where s is the logistic sigmoid function given by $s(x) = 1/(1 + \exp(x))$. Assuming the energy components are Gaussian distributed, the second and third terms of the integrand are given by multivariate Gaussian distributions. Since the integral in equation 4.15 is analytically intractable, we can approximate it using Laplace's method [203]. First we approximate the integrand with a multivariate Gaussian about a stationary point and then we evaluate the Gaussian integral to yield

$$p(\mathbf{y}|\mathbf{A}, \mu_P, \sigma_P^2, \mu_N, \sigma_N^2) \simeq p(\mathbf{y}|\mathbf{A}, \boldsymbol{\varepsilon}_{P,0}, \boldsymbol{\varepsilon}_{N,0})\; p(\boldsymbol{\varepsilon}_{P,0}|\mu_P, \sigma_P^2)\; p(\boldsymbol{\varepsilon}_{N,0}|\mu_N, \sigma_N^2)\; \frac{(2\pi)^{M/2}}{\sqrt{|\mathbf{H}|}}, \tag{4.17}$$

where $\boldsymbol{\varepsilon}_{P,0}$ and $\boldsymbol{\varepsilon}_{N,0}$ are the stationary points, M is the fixed number of contacts, and $\mathbf{H}$ is the Hessian matrix evaluated at the stationary points. The stationary points were found using Newton's method and the marginalized likelihood function was maximized using the Nelder-Mead method.

4.4.3 Reanalyzing β-lactamase data to account for library construction errors

The 13-crossover β-lactamase library (βlac13) had a significant amount of construction errors [191]. Sequencing of unselected chimeric genes found 9 of 13 to have frame shift mutations [56], which almost certainly result in inactive proteins. Since a majority of frame shifts are incorporated at the PCR step during library construction, it is likely these errors are present throughout all constructed chimeras [182]. The maximum likelihood estimate for the proportion of correctly constructed chimeras is $4/13 = 0.31$, with 95% confidence intervals between 0.09 and 0.61 using the Clopper-Pearson interval [202]. This sequencing data indicates there may be one to three sequence fragments (chimera blocks) that contain frameshift mutations. Assuming all frame shifts cause inactivation and exhaustive library coverage (over twelvefold sampling), the fraction of functional chimeras can be estimated by the number of functional chimeras divided by the number of correctly constructed chimeras. With these assumptions, we estimate the fraction of functional sequences to be 7×10^{-3} with 95% confidence intervals between 3×10^{-3} and 22×10^{-3} The same modification can be performed on chimeras binned by the number of homologous substitutions (Figure 4.1*A*) because the construction errors display little relation to the level of mutation.

4.5 Discussion

By using a statistical description of the protein recombinational landscape, we can gain insight into the behavior of an astronomical number of sequences, which could not be obtained experimentally or even by homology-based structural modeling. A probabilistic contact potential was used to specify the mean energy of individual chimeric proteins and how the energy of any sequence is expected to co-vary with others (equations 4.6 and 4.7), defining a multivariate probability distribution over all sequences accessible by recombination. While this random field model provides little information about specific sequences, it does reveal the large-scale structure of the recombinational landscape, which we used here to interpret the results from past recombination libraries. Within this random field, the expected values of various library properties show excellent agreement with experimental results across multiple protein families. This striking correspondence may arise because a library's properties depend on a large number of interactions, and the cumulative effect of these interactions converges toward the expected value due to the law of large numbers.

The random field model was used to study the enrichment of functional sequences within the recombinational landscape. As shown previously, we found the tolerance of proteins to recombination events to be influenced by the conservative effects of homologous substitutions, which have been previously selected to be compatible with the protein fold [56]. However, a more significant contribution comes from groups of coevolved residues varying together. This is especially relevant for understanding natural evolution, where the number of crossover events is relatively low. Evaluating the random field model across protein families, we found parent sequence identity to be the primary determinant for tolerance to recombination, while the specific crossover locations (library design) and parent fold make statistically significant, but minor contributions.

Using the random field model, we explored the origins of the additive structure within the recombinational landscape. Both sequence conservation among the parents and the partitioning of nonlinear interactions into structural modules make significant contributions to this additivity. The results presented here are for a random field that describes a protein's free energy difference between the functional and non-functional states, which is closely related to protein stability. However, the results are generally true for any landscape that is generated by higher-order, distance-dependent interactions, which could include numerous biophysical quantities.

Previous studies of protein fitness landscapes have highlighted the abundance of nonfunctional sequences [31, 32] and neutral sequence changes [37, 36, 82], suggesting a surface which is mostly flat and filled with holes [204]. In contrast to this full landscape, the recombinational landscape contains orders of magnitude fewer 'holes' (non-functional sequences). Despite the evidence for neutrality, the functional variation displayed within recombination libraries reveals the large-scale structure of the recombinational landscape, which arises from the cumulative effect of multiple mutations. In addition, most of this functional variation can be explained by additive effects, which are easily selected upon within evolving populations. While these results were observed in SCHEMA-minimized libraries, which tend to be optimized for both functional sequences and additivity, the random field model suggests these properties are generally true for recombination. These SCHEMA-minimized libraries also emphasize the preference for some crossover sites over others, which could explain the presence of recombination hotspots in natural genes [205, 177, 206]. The picture of the recombinational landscape that has emerged from the random field model is a surface enriched in functional sequences, which displays locally-epistatic behavior but has an overall additive structure.

The evolutionary benefit of intragenic recombination may arise because mutation and recombination effectively traverse different landscapes. While climbing the landscape by point mutations, evolution encounters a large number of nonfunctional sequences in addition to epistatic landscape features. In contrast, recombination explores sequences which are much more likely to be functional within a landscape containing an abundance of adaptive pathways. Recombination can provide faster adaptation than point mutation because it generates functional sequences with a large number of substitutions. Recombination may also find sequences that are inaccessible by adaptive point mutation, by simultaneously incorporating multiple coupled mutations, essentially 'jumping over' epistatic landscape features. A similar effect was recently described for recombination at the genome level [207], where the authors describe how landscapes arising from high epistasis within genes and no epistasis between genes strongly favors recombination. Running simulations on these 'modular' landscapes, the authors found recombination to provide an efficient route to genotypes that were inaccessible by point mutation.

Future recombination libraries could be improved by refining the design objective or pushing the library design to lower $\langle E_{sch} \rangle$ via better optimization. While designing chimeric protein libraries by minimizing $\langle E_{sch} \rangle$ works well, contributing up to a six fold enrichment in functional sequences over the average (or 30-fold relative to the least functional library), the library design objective could be improved with a better classification of favorable/unfavorable contacts. The parental and novel contact parameters estimated above show a significant overlap in density (Supplementary Figure 4.3), suggesting it is relatively common for parental contacts to be as deleterious as novel contacts, and vice versa. A contact classification that maximized the difference between favorable/unfavorable contacts would provide a more robust library design objective, resulting in libraries with even greater numbers of functional sequences. New contact classifications can be explored using the maximum marginal likelihood estimation method described above, with class separation being quantified by a statistical distance metric such as the Kullback-Leibler divergence [208].

When the parent sequence identity is high, designing a library to maximize the number of functional sequences becomes less important because any library will contain a substantial fraction of functional sequences. For example, the random field model predicts a 1.25-fold enrichment in functional sequences for $\langle E_{sch} \rangle$-minimized CBHII (pairwise sequence identity: 65%, 67%, 82%) libraries over randomly generated libraries. For libraries between closely related parents ($> 70\%$ identity), alternative design objectives that seek to maximize the information content of the library, such as the expected value of the variance, may be more useful. Libraries with maximized variance would provide more information about the sequence properties of interest and potentially result in more extreme (highly-optimized) chimeric proteins. Similar ideas could be applied to more specific properties, such as substrate specificity, where the library design objective could be to maximize residue variation within the active site.

The most straightforward route to improving recombination library design comes from more advanced optimization protocols. All libraries discussed above were designed under the requirement that chimera blocks be contiguous in sequence, which is very limiting. By removing this design constraint, there are often libraries with the same parents, number of blocks, and block size distribution that have an $\langle E_{sch} \rangle$ of zero. Libraries with a $\langle E_{sch} \rangle$ of zero are predicted to be between 70–90% functional, depending on parent sequence identity. For non-contiguous block CBHII libraries, for example, we estimate a 2.1-fold enrichment in functional sequences over random libraries, compared to the 1.25-fold increase estimated for contiguous block libraries. In addition, as $\langle E_{sch} \rangle$ approaches zero, the library additivity A goes to one because all quadratic interactions are partitioned into structural modules. This suggests the predictive ability of linear regression models would be even more accurate for these non-contiguous block libraries. On the downside, non-contiguous block libraries cannot be made following standard construction protocols due to the large number of sequence fragments. Typically, these libraries would require total gene synthesis of each desired member, which makes constructing the full libraries prohibitively expensive.

Intragenic recombination is a powerful molecular diversification mechanism. The ubiquity of intragenic recombination in nature and experimental evidence from protein recombination libraries suggest that it provides distinct advantages over point mutation. In naturally evolving populations, these two genetic variation mechanisms work together. Mutation provides new innovation, while recombination efficiently sorts through the large combinatorial space of existing diversity. A better understanding of how to balance mutation and recombination could assist in engineering highly-optimized proteins.

4.6 Supplementary Material

4.6.1 Parameters estimated with the logistic likelihood function are proportional to Gibbs free energy differences

Here, we show the parameters estimated by maximizing the logistic likelihood function can be interpreted as Gibbs free energy differences (in arbitrary units) from the nonfunctional state. For notational simplicity, we combine all energy terms (ε_P^i and ε_N^i) into the vector $\boldsymbol{\varepsilon}$ and all binary indicator variables ($a^i_{\mathbf{c},P}$ and $a^i_{\mathbf{c},N}$) into the vector $\mathbf{a_c}$. With this notation, the energy of chimera $\mathbf{c}$ is $E_\mathbf{c} = \mathbf{a_c} \cdot \boldsymbol{\varepsilon}$.

Within a binary functional status data set D, we only observe if a chimeric protein $\mathbf{c}$ is functional or not, which can be represented by the step function

$$y_\mathbf{c} = \begin{cases} 1 & \text{if } \Delta G_\mathbf{c} < 0 \quad \text{(functional)}, \\ 0 & \text{if } \Delta G_\mathbf{c} \geq 0 \quad \text{(nonfunctional)}, \end{cases} \tag{4.18}$$

where $\Delta G_\mathbf{c}$ is the Gibbs free energy difference between chimera $\mathbf{c}$'s functional and nonfunctional states. The logistic likelihood function of data set D is

$$p(D|\boldsymbol{\varepsilon}) = \prod_{\mathbf{c}\in D} s(\mathbf{a_c} \cdot \boldsymbol{\varepsilon})^{y_\mathbf{c}} s(-\mathbf{a_c} \cdot \boldsymbol{\varepsilon})^{1-y_\mathbf{c}}, \tag{4.19}$$

where the data set D consists of examples of the mapping from $\mathbf{a_c}$ to $y_\mathbf{c}$, and s is the logistic sigmoid function given by $s(x) = 1/(1 + \exp(x))$. Note this logistic function is a reflection about the origin ($x = -x$) of the standard logistic function because negative energy is favorable. Maximizing this likelihood function with respect to $\boldsymbol{\varepsilon}$ finds the parameter set that was most likely to generate the observed data.

This maximum likelihood (ML) parameter estimate can be found by minimizing the negative log likelihood, which is a strictly convex function [209]. This convexity ensures that any parameter set where the gradient of the negative log likelihood function is zero in all directions is a global minimizer. The gradient of the negative log likelihood is given by

$$\nabla - \log p(D|\boldsymbol{\varepsilon}) = \sum_{\mathbf{c}\in D} \mathbf{a_c} \left[s(\mathbf{a_c} \cdot \boldsymbol{\varepsilon}) - y_\mathbf{c} \right]. \tag{4.20}$$

This gradient is clearly equal to zero when $s(\mathbf{a_c}\cdot\boldsymbol{\varepsilon}) = y_\mathbf{c}$ for all $\mathbf{c} \in D$. This occurs when $\mathbf{a_c}\cdot\alpha\boldsymbol{\varepsilon} = \Delta G_\mathbf{c}$ for large α, and is approximately true for all positive α. Assume the Gibbs free energy difference can be decomposed into each contact's free energy contribution $\Delta G_\mathbf{c} = \mathbf{a_c} \cdot \mathbf{g}$. Then the maximum likelihood (ML) parameter estimate is simply a linear scaling these Gibbs free energy terms

$$\varepsilon_{\mathrm{ML}} = \frac{1}{\alpha}\mathbf{g}. \tag{4.21}$$

4.6.2 Estimation of contact parameters on other recombination libraries

The parental and novel contact parameters (μ_P,σ_P^2,μ_N,σ_N^2) were estimated on three additional data sets and all the results are presented in Supplementary Table 4.2. While the absolute values of the parameters are quite different, the random field model is unchanged by linear rescaling (changing units). Therefore, only the relative values of the parameters are important (depicted in Supplementary Figure 4.3). Within all four parameter sets, we see the mean of parental contacts is slightly favorable and novel contacts are significantly deleterious. The means of these two distributions are separated by approximately one standard deviation, indicating it is relatively common for parental contacts to be as deleterious as novel contacts, and vice versa. All conclusions from the random field model depend on only these qualitative relationships between parental and novel contacts.

4.6.3 Expected values of various landscape properties

4.6.3.1 Recombination library properties

Consider a recombination library L, generated by recombining b sequence fragments from p parental sequences. From the definition of the library mean M_L (equation 4.8), the expected value of the library mean within the random field is

$$\mathrm{E}[M_L] = \frac{1}{p^b}\sum_{\mathbf{c}\in L}\mathrm{E}[\mathcal{E}_{\mathbf{c}}] \tag{4.22}$$

and the variance of the library mean is

$$\mathrm{Var}[M_L] = \frac{1}{p^{2b}}\sum_{\mathbf{c1}\in L}\sum_{\mathbf{c2}\in L}\mathrm{Cov}[\mathcal{E}_{\mathbf{c1}},\mathcal{E}_{\mathbf{c2}}], \tag{4.23}$$

where the expected value of the random field $\mathrm{E}[\mathcal{E}_{\mathbf{c}}]$ is defined in equation 4.6, and the covariance within the random field $\mathrm{Cov}[\mathcal{E}_{\mathbf{c1}}, \mathcal{E}_{\mathbf{c2}}]$ is defined in equation 4.7.

Similarly, the expected value of the library variance V_L (equation 4.9) is given by

$$\mathrm{E}[V_L] = \frac{1}{p^b} \sum_{\mathbf{c} \in L} \mathrm{E}\left[(\mathcal{E}_{\mathbf{c}} - M_L)^2\right]) \tag{4.24}$$

which can be expanded to

$$\mathrm{E}[V_L] = \frac{1}{p^b} \sum_{\mathbf{c} \in L} \left[(\mathrm{E}[\mathcal{E}_{\mathbf{c}}] - \mathrm{E}[M_L])^2 + \mathrm{Var}[\mathcal{E}_{\mathbf{c}}] + \mathrm{Var}[M_L] - 2\,\mathrm{Cov}[\mathcal{E}_{\mathbf{c}}, M_L]\right] \tag{4.25}$$

where $\mathrm{Var}[\mathcal{E}_{\mathbf{c}}] = \mathrm{Cov}[\mathcal{E}_{\mathbf{c}}, \mathcal{E}_{\mathbf{c}}]$ and

$$\mathrm{Cov}[\mathcal{E}_{\mathbf{c}}, M_L] = -\,\mathrm{E}[\mathcal{E}_{\mathbf{c}}]\,\mathrm{E}[M_L] + \frac{1}{p^b} \sum_{\mathbf{c2} \in L} \left(\mathrm{E}[\mathcal{E}_{\mathbf{c}}]\,\mathrm{E}[\mathcal{E}_{\mathbf{c2}}] + \mathrm{Cov}[\mathcal{E}_{\mathbf{c}}, \mathcal{E}_{\mathbf{c2}}]\right). \tag{4.26}$$

From this, we can substitute equations 4.22, 4.23, and 4.26 into equation 4.25 to get an expression for the expected value of the library variance.

4.6.3.2 Fraction of functional sequences

Within library L, we assume the distribution of energies is Gaussian, which is a good approximation when the library size p^b is greater than a few hundred sequences. The fraction of functional sequences is given by evaluating the Gaussian cumulative distribution function at zero

$$f_F = \frac{1}{2}\left[1 - \mathrm{erf}\left(\frac{M_L}{\sqrt{2V_L}}\right)\right], \tag{4.27}$$

where erf is the error function.

We can approximate the expected value of the fraction of functional sequences with a first-order Taylor series of f_F about $\mathrm{E}[M_L]$ and $\mathrm{E}[V_L]$

$$\mathrm{E}[f_F] \approx \frac{1}{2}\left[1 - \mathrm{erf}\left(\frac{\mathrm{E}[M_L]}{\sqrt{2\,\mathrm{E}[V_L]}}\right)\right]. \tag{4.28}$$

Monte Carlo sampling was used to verify that this approximation is leading order.

4.6.3.3 Landscape additivity

The variance within library L can be partitioned into additive and epistatic components ($V_L = V_A + V_E$). We can define an energy function that describes only the additive energy

$$E_{A,\mathbf{c}} = \sum_i b^i_{\mathbf{c},P}\ \varepsilon^i_P + \sum_i b^i_{\mathbf{c},N}\ \varepsilon^i_N, \tag{4.29}$$

where $b^i_{\mathbf{c},P}$ and $b^i_{\mathbf{c},N}$ specify how the energy terms ε^i_P and ε^i_N contribute to additive energy of chimera $\mathbf{c}$. The b variables are analogous the the a variables in equation 4.1, however they are no longer binary. Their values are are determined by the contribution that interaction i makes to overall library L.

Each interaction i is specified by a residue-residue contact between positions p^i_1 and p^i_2. Within the parent alignment, if the residues at positions p^i_1 or p^i_2 are conserved, then interaction i only contributes to linear variation within the library. Oppositely, if the residues at positions p^i_1 and p^i_2 are nonconserved, then interaction i contributes to quadratic variation within the library. Merging the two variables $b^i_{\mathbf{c},P}$ and $b^i_{\mathbf{c},N}$ into a single variable, we define

$$b^i_{\mathbf{c}} = \begin{cases} a^i_{\mathbf{c}} & \text{if } p^i_1 \text{ or } p^i_2 \text{ conserved,} \\ \\ p(i|\mathbf{c},p_1) + p(i|\mathbf{c},p_2) - p(i) & \text{if } p^i_1 \text{ and } p^i_2 \text{ nonconserved,} \end{cases} \tag{4.30}$$

where $p(i)$ is the probability of interaction i within the entire library L, $p(i|\mathbf{c},p_1)$ is the probability of interaction i in the subset of library L that has the same residue at position p^i_1 as chimera $\mathbf{c}$, and $p(i|\mathbf{c},p_2)$ is the probability of interaction i in the subset of library L that has the same residue at position p^i_2 as chimera $\mathbf{c}$. At the nonconserved contacts, $b_{\mathbf{c}}$ effectively averages over all quadratic variation that cannot be represented by an additive model.

With this additive energy function, we can define library L's additive mean M_A and variance V_A using the standard formulas (equations 4.8 and 4.9). The expected values $\mathrm{E}[M_A]$ and $\mathrm{E}[V_A]$ can be calculated using the same equations as the ‘Recombination library properties’ section (above). Using the definition of library additivity (equation 4.13), we can approximate the expected value of the additivity with a first-order Taylor series about $\mathrm{E}[V_A]$ and $\mathrm{E}[V_L]$

$$\mathrm{E}[A] \approx \frac{\mathrm{E}[V_A]}{\mathrm{E}[V_L]}. \tag{4.31}$$

Monte Carlo sampling was used to verify that this approximation is leading order.

4.7 Figures and Tables

library name	protein family	fold class	parent kingdom	sequence length	number of parents p	number of crossovers n	pairwise parent identity (%)	SCHEMA disruption $\langle E_{sch} \rangle$	fraction functional f_F	fraction functional 95% CI	additivity A	ref.
βlac13	β-lactamase	alpha+beta	bacteria	290	2	13	39	91.5	0.007	0.003, 0.022		[191]
βlac	β-lactamase	alpha+beta	bacteria	267	3	7	37,42,39	50.2	0.20	0.17, 0.24		[198]
P450	cytochrome P450	all alpha	bacteria	466	3	7	64,65,67	33.4	0.47	0.43, 0.51	0.84	[80]
CBHII	class II cellobiohydrolase	alpha/beta	fungi	361	3	7	65,82,67	15.7	0.48	0.33, 0.63	0.86	[79]
CBHI	class I cellobiohydrolase	all beta	fungi	441	5	7	69,61,72, 64,66,73, 81,64,66,70	23.9	0.78**	0.62, 0.90**	0.97	[159]
Cel48	family 48 cellulase	all alpha	bacteria	736	3	7	71,64,65	40.2	0.53	0.43, 0.63	0.73	[200]
Cel9	family 9 cellulase	all alpha	bacteria	619	3	7	51,63,54	18.7	0.18	0.09, 0.30		[192]
Cel5	family 5 cellulase	alpha/beta	fungi	337	3	5	65,31,34	63.1	0.31	0.19, 0.46		[199]
Arg	arginase	trimeric alpha/beta	animalia	3×306	2	7	61	23.0	0.50	0.18, 0.81		[201]

Table 4.1: Summary of nine protein recombination libraries. The library's fold class was retrieved from the SCOP structural database [210]. The fraction of functional sequences and additivity were calculated as described in Methods. **The fraction functional estimates for the CBHI library are significantly biased due to the chimera sampling protocol [159] and are therefore not included in the analysis.

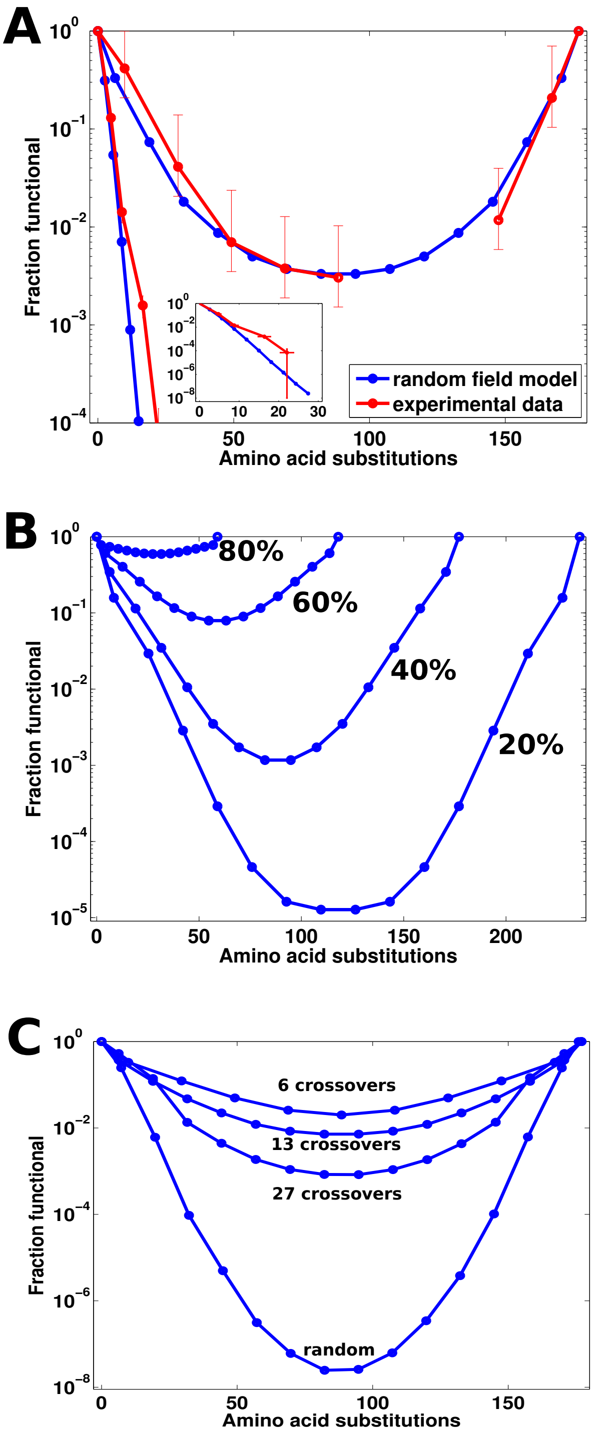


Figure 4.1: Effect of homologous substitutions on the fraction of functional sequences in a library of chimeric β-lactamases. (A) The random field model agrees well with experimental random and homologous substitutions in β-lactamase [56]. The parabolic curve displays the effect of homologous substitutions and the error bars represent the 95% confidence intervals of the fraction of correctly constructed chimeras (see Methods). The steep exponential curves (and inset) show the effect of random mutations and the error bars represent one standard error. (B) As parent sequence identity decreases, the homologous substitution curves stretch to higher levels of mutation and lower fraction functional. Shown are the βlac13 library (crossover locations and contacts) averaged over 100 random parent sequences with sequence identity ranging from 20–80%. (C) As the number of crossovers decreases, the homologous substitution curve shifts towards a higher fraction functional. Shown are the βlac13 library (parents and contacts) averaged over 100 random crossover locations with the number of crossovers varying from 6 to 27. The random homologous substitution curve was generated by averaging over 100 randomly sampled sequences at each level of mutation.

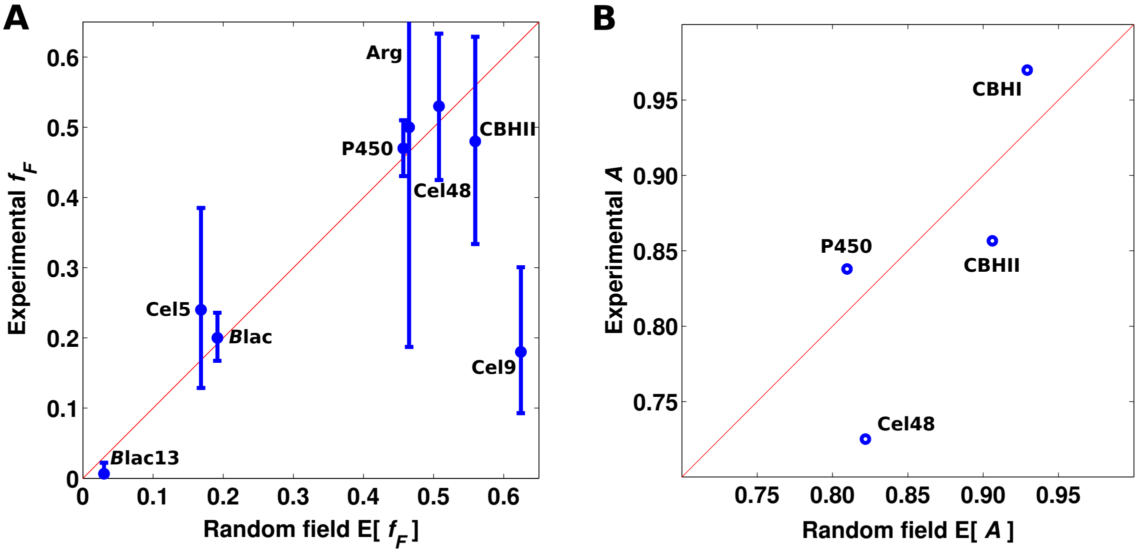
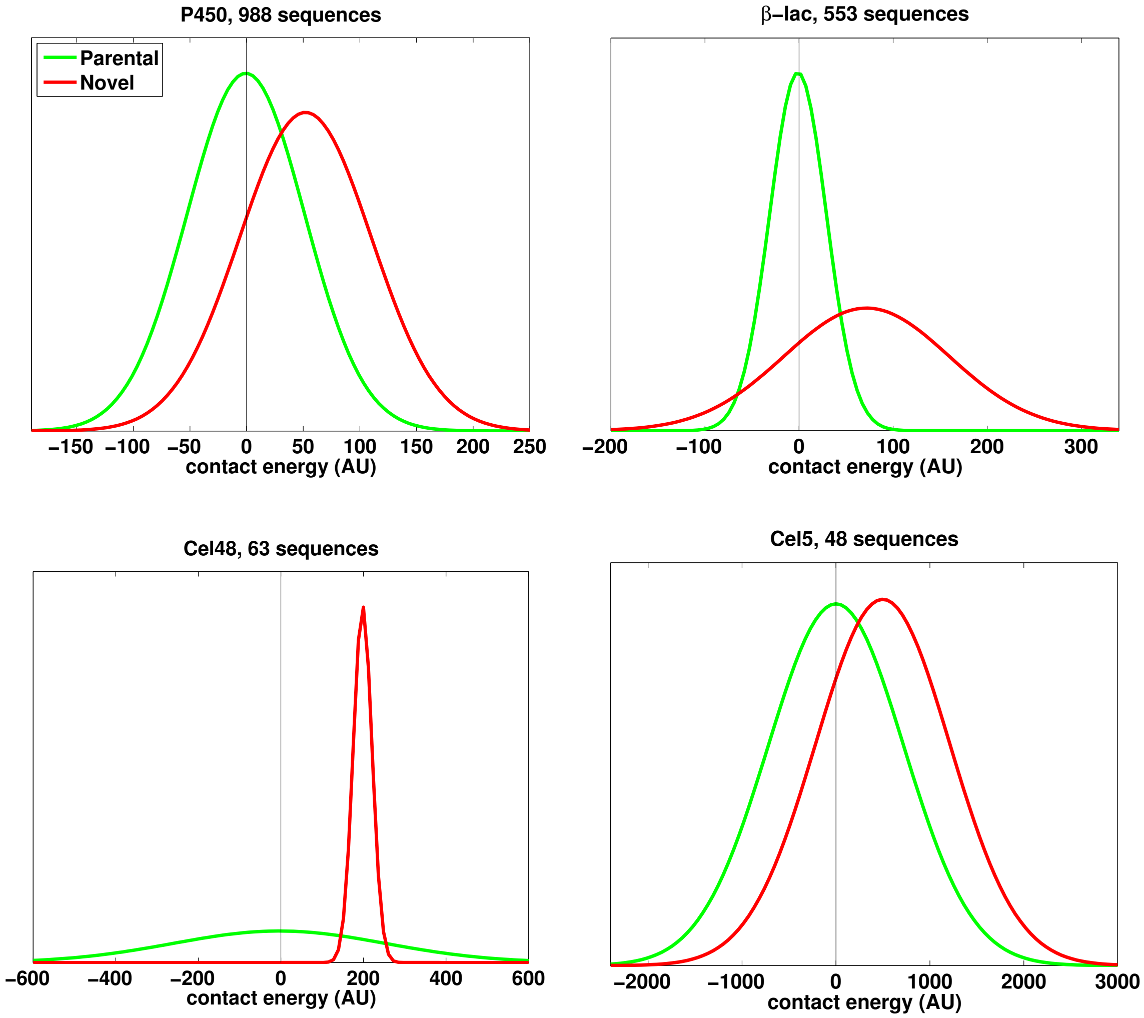


Figure 4.2: Comparison between library properties and their expected values within the random field model. Note diagonal lines represent $x = y$. (A) The random field's expected fraction of functional sequences shows quantitative agreement with experimental results. Error bars represent the binomial 95% confidence intervals calculated using the Clopper-Pearson method [202]. Omitting the Cel9 outlier yields a correlation coefficient of $r = 0.95$ with $p < 0.005$. (B) The expected additivity agrees well with experimentally determined values ($r = 0.78$ with $p = 0.21$). While the small data set limits the statistical significance of this correlation, all E[A]s are large and within the ranges that are observed experimentally.

4.8 Supplementary Figures and Tables

library	number of sequences	μ_P	σ_P	μ_N	σ_N
P450	988	-0.7	51.9	52.0	58.3
βlac	553	-1.1	30.1	71.8	87.9
Cel48	63	-3.9	257.3	198.1	22.7
Cel5	48	-0.2	722.3	497.6	713.4

Supplementary Table 4.2: Contact parameters estimated from four recombination libraries. Note the contact variation is presented as standard deviations, for direct comparison with the means. We see qualitatively similar relationships for all four parameter sets (Supplementary Figure 4.3).



Supplementary Figure 4.3: Comparison of contact parameters estimated on four different recombination libraries. Qualitatively, all parameter sets have slightly favorable parental contacts, significantly deleterious novel contacts, and significant variation in both. Shown are the Gaussian probability density functions with the associated parameters.

Chapter 5

Gaussian process models of the protein fitness landscape

5.1 Abstract

Understanding the map from protein sequence to protein function is important for understanding natural evolution and engineering proteins with new and useful properties. We demonstrate that this protein fitness landscape can be be inferred from experimental data using Gaussian processes, a Bayesian learning technique. These Gaussian process landscapes can model a variety of protein sequence properties including functional status, thermostability, enzymatic activity, and binding affinity. By training on experimental data, these models achieve an unrivaled quantitative accuracy across a large number of sequences. Furthermore, the explicit representation of model uncertainty allows for efficient searches through the massive space of possible protein sequences. We develop and test two protein sequence design algorithms motivated by Bayesian decision theory. The first identifies small sets of protein sequences which are highly-informative about the landscape. The second algorithm identifies optimized protein sequences by iteratively improving the Gaussian process model in regions of the landscape that are predicted to be highly-optimized.

5.2 Introduction

The fitness landscape, which describes the map from genotype to phenotype, is an important concept in evolutionary biology and optimization [26, 211]. At the molecular level, the properties of protein sequences form a high dimensional surface over protein sequence space [1]. This protein fitness landscape could represent a protein's contribution to organismal fitness, but may also describe any biophysical property of protein sequences including thermostability, enzymatic activity, or binding affinity. Understanding the structure of this surface is important for explaining how natural proteins evolve because it describes the spectrum of possible phenotypes and the mutational accessibility among them. This surface is also the objective function for protein engineering, which seeks to identify protein sequences that are highly optimized for a wide variety of applications.

Identifying optimized sequences within the protein fitness landscape is extremely challenging for several reasons. First, the space of possible protein sequences is incomprehensibly large and will never be searched exhaustively by any means, naturally, in the laboratory, or even computationally [25, 212]. Within this massive space of possible sequences, functional proteins are extremely scarce, estimates range from 1 in 10^{11} to as low as 1 in 10^{77} [31, 32]. Of these functional sequences, most are poorly optimized and there is believed to be an exponentially decreasing number of sequences with higher levels of fitness [33, 213]. Within the protein fitness landscape, highly-optimized sequences are vanishingly rare and hidden in the extreme abundance of nonfunctional and mediocre sequences.

Computational protein design uses models of protein function to guide the search through sequence space. These models typically contain an atomic structural representation of a protein and energy-based scoring functions to quantify the target function [214, 6]. While there has been considerable success in recent years using these computational searches to engineer functional proteins, these methods still lack reliability and often result in sequences with properties that are far inferior to natural proteins [6, 8, 215]. Many of these difficulties arise from the use of models which lack even qualitative accuracy, much less the ability to reliably rank the performance of sequences. In general, the factors that make one protein perform better than another are complex and largely unknown. A major challenge for computational protein design is the development of models which accurately describe the mapping from protein sequence to function [216].

Here, we introduce a new class of models for protein function that infer the protein fitness landscape directly from experimental data. With examples of the mapping from sequence to function, the landscape can be learned using Gaussian process regression, a technique that has gained recent popularity in machine learning where it falls into the broader class of kernel methods [217, 218]. The kernel function describes the covariance structure of the fitness landscape by specifying how the properties of pairs of sequences are expected to co-vary. We develop a structure-based kernel function inspired by the simple principle that sequences with similar structures are more likely to have similar properties. These Gaussian process models provide a full probabilistic description of the protein fitness landscape, including the mean and variance of any sequence. Importantly, a sequence's variance provides a measure of the model's uncertainty, which can be used to guide the search through sequence space using Bayesian decision theory.

Using chimeric cytochrome P450s, we demonstrate that Gaussian process landscapes can accurately describe a variety of sequence properties including binary functional status, thermostability, enzymatic activity, and binding affinity. By training directly on experimental data, these models implicitly account for all factors which contribute to a specific property, including those which are unknown. This provides an unrivaled quantitative accuracy across a potentially astronomical number of sequences. Using the Gaussian process model's uncertainty as a guide, we develop two algorithms which efficiently explore the protein fitness landscape. The first is an experimental design algorithm, which identifies the most informative points within the landscape before they are measured. This is used to design a set of 29 highly-informative chimeric P450s. The second algorithm identifies optimized protein sequences by iteratively improving the Gaussian process model in regions of the landscape that are predicted to be highly-optimized. This algorithm is used to design chimeric P450s with thermostabilities beyond what has been achieved by other protein engineering methods.

5.3 Results

5.3.1 Gaussian process model of the protein fitness landscape

Gaussian processes have gained recent attention in supervised machine learning, where they are used for both regression and classification tasks [218]. These nonparametric models use a kernel, or covariance function to define a prior probability distribution over function space. Given examples of the target function, its posterior probability distribution can be inferred using Bayes' theorem. In general, kernel functions represent a notion of similarity between inputs, which allows them to describe many types of complex relationships. This provides Gaussian processes with extreme flexibility, and the ability to learn from structured objects including strings, sets, and graphs.

To model the protein fitness landscape with Gaussian processes, we must define a kernel function which accurately captures the notion of distance between pairs of sequences. While the Hamming distance would be a natural metric, the properties of proteins depend on sequence only though their structure. Therefore, we propose a structure-based distance metric which assumes a fixed structure within a protein family, as defined by a residue-residue contact map. With this contact map, the structural distance between two proteins in the same family is defined as the number of contacting residues with different amino acids. This structure-based distance metric is similar to the Hamming distance, but it describes the effects of mutations more accurately. For example, the properties of sequences that differ by a surface mutation are expected to be more similar than sequences that differ by a core mutation. Most importantly, this distance metric can be represented as an inner-product and therefore satisfies the requirements to be a valid kernel function for Gaussian process learning [218].

Given experimental examples of the mapping from protein sequence to protein function, the full protein fitness landscape can be inferred using Gaussian processes. For regression, the expected value of the landscape f at sequence $\mathbf{s}$ is given by

$$\mathrm{E}[f(\mathbf{s})] = \mathbf{k}^T(K + \sigma_n^2 I)^{-1}\mathbf{y}, \tag{5.1}$$

and the variance of the landscape is

$$\mathrm{Var}[f(\mathbf{s})] = k(\mathbf{s}, \mathbf{s}) - \mathbf{k}^T (K + \sigma_n^2 I)^{-1} \mathbf{k} \tag{5.2}$$

where k is the structure-based kernel function, K is the kernel function evaluated at all pairs of sequences in the training set ($K_{i,j} = k(\mathbf{s}_i, \mathbf{s}_j)$), $\mathbf{k}$ is the kernel function evaluated at sequence $\mathbf{s}$ and all sequences in the training data ($\mathbf{k}_i = k(\mathbf{s}, \mathbf{s}_i)$), σ_n^2 is the variance of the experimental measurements, and $\mathbf{y}_i$ is the experimentally determined property of training set sequence $\mathbf{s}_i$. From equation 5.1, we see a sequence's expected value is simply a linear combination of all the current data $\mathbf{y}$, where the coefficients depend on the structural distance between the sequence and each sequence in the training set. This can be viewed as a spatial interpolation within the protein fitness landscape, where sequences that are close in structure are likely to have similar properties. Interestingly, a nearly identical method has been used for decades in geostatistics to infer the structure of terrestrial landscapes [185]. The variance of a sequence (equation 5.2) can be interpreted as the difference between what was known about the sequence before the experiments and what was learned about the sequence from the experiments. As expected, Gaussian process models have high confidence in regions of the landscape which are well sampled, and low confidence in regions that are not. For the prediction of discrete-valued properties (classification), the Gaussian process posterior does not have a simple, closed-form solution, but can be found using several well-established approximations [219].

The performance of Gaussian process landscape models was tested on a previously published data set of chimeric cytochrome P450 thermostabilities [86]. This Gaussian process model showed excellent predictive ability (cross-validated $r = 0.95$, mean absolute deviation $MAD = 1.4$ °C), as shown in Figure 5.1*A*. Previously, this data set was modeled using a linear regression model that associated weights to individual sequence fragments within the data set [86]. This linear regression model worked well (cross-validated $r = 0.90$, $MAD = 2.0$ °C) and was used to identify highly stabilized chimeric P450s. To compare the predictive performance of the two models, we sampled random sets of training sequences from the data set, trained the linear and Gaussian process models, predicted the thermostability of the remainder of the data set, and quantified each model's predictive ability in terms of the correlation coefficient (r) and the mean absolute deviation (MAD). This was performed with training sets varying from 2 to 60 sequences, and the results for each training set size were averaged over 1000 random samples, Figure 5.1B. The Gaussian process model significantly outperforms the linear regression model, typically explaining 30% more of the variation in thermostability across all training set sizes. On average, the linear regression model trained on all the data (218 sequences for 10-fold C.V.) has the same predictive ability as the Gaussian process model trained on only 40 sequences. These substantial increases in predictive performance can be attributed to the more accurate sequence-sequence covariance specification provided by the structure-based kernel function. The performance of other kernel functions is shown in Supplementary Figure 5.4.

The most significant benefit of the Gaussian process model, over the linear regression model, is predictions are not restricted to sequences composed of a fixed set of fragments (site-directed recombination libraries). In fact, the Gaussian process model can predict the properties of any sequence, however, these predictions will typically contain so much uncertainty (variance) they will be of no use. A more useful prediction domain is the set of all sequences that can be generated by recombining the three parent cytochrome P450s (CYP102A1, CYP102A2, and CYP102A3), which still represents an astronomically large sequence space ($> 10^{75}$ sequences). To test the predictive ability in this new prediction domain, we trained a Gaussian process model on the data set described in the previous paragraph (fixed set of sequence fragments) and predicted the thermostabilities of a data set containing single- and double-crossover chimeric P450s between CYP102A1 and CYP102A2 [220] (no fixed set of sequence fragments). Each of these single- and double-crossover chimeric P450s are composed of different sequence fragments, contain different crossover locations than the training sequences, and on average differ from the closest sequence in the training set by 29.6 mutations (shown schematically in Supplementary Figure 5.5). The Gaussian process model shows good predictive ability ($r = 0.76$, $MAD = 3.0$ °C) on these sequences that cannot be modeled with sequence fragment-based linear regression (Figure 5.1C). While the predictions are less accurate on these sequences, the model is aware these predictions are in an uncertain region of sequence space as indicated by its large confidence intervals (Supplementary Figure 5.5). For these predictions, nearly all experimental measurements (17/19) fall within the model's 95% confidence intervals.

Next, we tested the ability of Gaussian process landscapes to classify the functional status (functional/nonfunctional) of chimeric P450s. Here, a functional P450 must satisfy a number of requirements including sufficient expression, stability, and the ability to properly bind the heme prosthetic group. We trained a Gaussian process model on a large, previously published functional status data set [80] (see Methods). This Gaussian process classifier shows excellent predictive ability, correctly classifying 89% of sequences (10-fold cross-validation). For comparison, a sequence fragment-based logistic regression classifier only achieves 81% accuracy (10-fold cross-validation). Once again, this Gaussian process functional status classifier is applicable to the astronomical number of sequences accessible by recombining the three parent enzymes. By training directly on experimental data, Gaussian process models implicitly capture the numerous and possibly unknown factors which determine if a sequence will form a functional cytochrome P450.

5.3.2 Experimental design on holey landscapes

The utility of Gaussian process models relies on a thorough sampling of the very high dimensional protein fitness landscape. If done inefficiently, this could require an unimaginable amount of experimentation. Fortunately, we can take advantage of the Gaussian process landscape's representation of model uncertainty to select the most informative sequences before they are measured. This is referred to as experimental design, and can significantly reduce the number of experiments required to train a statistical model. There is a well-developed theory for designing informative experiments using Gaussian process models, which has been applied to a number of problems including environmental monitoring and outbreak detection [221, 222]. Experimental design can be posed as a combinatorial optimization problem, where the objective quantifies the informativeness of a set of observations, typically as a function of their covariance matrix. For many of these objective functions, a simple greedy approximation algorithm can achieve near-optimal observation selection for experimental design [152].

Considering the set of all possible sequences in the landscape L and a subset of these sequences S, a logical measure of informativeness is the mutual information $I(S; L)$, that is how much S reduces the uncertainty in L. The set of sequences S that maximize this mutual information, subject to $|S| \leq n$, can be found efficiently using greedy maximization (n is the maximum allowed size of the design). The resulting experimental designs contain sequences which are spaced as far as possible within the fitness landscape.

The primary challenge of performing experimental design on the protein fitness landscape is the abundance of nonfunctional sequences, which provide no information about the protein sequence properties we wish to model. Fortunately, the Gaussian process functional status classifier, which was presented in the previous section, can predict a sequence's probability of functioning with high accuracy. With this knowledge, a better experimental design objective is to maximize the expected value of the mutual information $\mathrm{E}[I(S; L)]$ (see Methods). The set of sequences that maximize this objective are highly-informative, while still having a high probability being functional.

Using a greedy approximation algorithm, we identified a set of 20 sequences which near-optimally maximize the expected value of the mutual information (see Methods). These 20 sequences were constructed and expressed. Seventeen produced functional cytochrome P450s. Building upon this set of sequences, we performed a second experimental design containing 10 sequences, 9 which produced functional cytochrome P450s. These 26 new cytochrome P450s, in addition to the three parent enzymes, provide a highly-informative yet experimentally tractable sampling of the chimeric P450 landscape (shown schematically in Supplementary Figure 5.7). On average, the sequences within this experimental design differ from each other by 106.1 mutations. In the following section this set of sequences is used to train Gaussian process models for enzymatic activity and binding affinity.

5.3.3 Gaussian process landscapes for enzymatic activity and binding affinity

We would like to test if Gaussian process landscapes can model sequence properties besides thermostability and functional status. Each of the 29 sequences (3 parents and 26 chimeras) in the experimentally-designed set was expressed, purified and characterized for enzymatic activity and binding affinity (Supplementary Table 5.2). Enzymatic activity (total substrate turnovers per enzyme) was measured on the following substrates: 2-phenoxyethanol, ethoxybenzene, ethyl phenoxyacetate, propranolol, chlorzoxazone, 11-phenoxyundecanoic acid (see Methods). Binding affinity (K_d) measurements were performed on dopamine and serotonin (see Methods).

Gaussian process regression was used to model the logarithm of the enzymatic activity and binding affinity for each compound. For all of these sequence properties, the Gaussian process models displayed poor cross-validated predictive ability. Suspecting the presence of outliers, we searched for aberrant observations within the data set (see Methods). From this analysis, we identified three strong outliers (ED7, ED9, ED28) and two occasional outliers (ED10, ED12). Looking back at each sequence's absorbance spectra, four of these proteins (ED7, ED9, ED12, ED28) have Sort peaks that are shifted from typical cytochrome P450s and the the remainder of the data set (Supplementary Figure 5.8). ED7, ED12, and ED28 have blue-shifted Soret peaks, indicative a of high-spin heme, which is normally observed with reduced solvent accessibility in the active site. ED9 has a red-shifted Soret peak, which suggests the presence of a distal heme ligand. Regardless of the specific mechanisms involved, these four outliers appear to be adopting conformations that are minimally populated by the other chimeric P450s and therefore cannot be modeled with the remainder of the data set.

Removing outliers from each data set and training the Gaussian process model on the remaining sequences shows good predictive ability, Figure 5.2 and Supplementary Figure 5.9. Many of these sequence properties are minimally correlated with each other (Supplementary Table 5.3) and with thermostability, confirming that Gaussian process landscapes are able to model a wide variety of sequence properties. As a final validation, we trained Gaussian process models on our experimentally designed set of sequences and used these models to predict the enzymatic activity of a previously published data set [85]. These predictions show reasonable agreement with the previously published values (2-phenoxyethanol: $r = 0.72$, ethoxybenzene: $r = 0.61$, ethyl phenoxyacetate: $r = 0.45$, propranolol: $r = 0.13$, chlorzoxazone: $r = 0.46$) and scatter plots are shown in Supplementary

Figure 5.10.

We would like to understand how Gaussian process models are able to capture complex properties such as enzymatic activity and binding affinity. Close inspection of the parent structures reveals that all active site residues are completely conserved and therefore all the chimeric P450s have identical active sites. Furthermore, Poisson-Boltzmann calculations suggest minimal influence of long-range electrostatic interactions (See Methods). It may be that the variation we observe within our data set is arising due to minor differences in the chimeric P450s conformational preferences. The Gaussian process model would be able to capture these differences if the system was dominated by two (or maybe a few) conformational states. If we assume the energy of each conformational state can be represented with a Gaussian process model, then differences between conformational states and therefore conformational preferences can also be represented. By training on experimental data, Gaussian process models can accurately account for these subtle differences. In contrast, modeling these effects with energy-based scoring functions is currently extremely challenging, if not impossible.

5.3.4 Sequence optimization on Gaussian process landscapes

Given the exceptional predictive ability of Gaussian process landscapes, as demonstrated above, it is compelling to use these models to design highly-optimized protein sequences. While these models can predict the properties of an astronomical number of sequences, a majority of these predictions are of little value because the model's uncertainty (variance) is so large. This predictive uncertainty can be reduced by experimentally sampling the landscape in previously uncharted regions. However, the same experimental effort could also be directed towards designing optimized sequences. When optimizing these uncertain functions, one is faced with the decision between trusting the current model and therefore selecting *highly-optimized* sequences, or not trusting the model and selecting *highly-informative* sequences. This is referred to as the exploitation-exploration dilemma, which is challenging because there are no set criteria for how this decision should be made [223]. In general, we would like to make the model good enough to design highly-optimized sequences, but no better.

In theoretical computer science, protein sequence design on Gaussian process landscapes is known as a multi-armed bandit problem because of its similarity to optimally playing slot machines in a casino [224]. These multi-armed bandit problems have been applied to online advertising, clinical trials, and robotic control [225, 226, 227]. The key feature of these problems is the trade-off between acting optimally based on current knowledge or acquiring new knowledge–the exploitation-exploration dilemma. When the objective is modeled as a Gaussian process, or any model which explicitly accounts for uncertainty, optimal solutions can be found efficiently using upper confidence bound (UCB) algorithms [228, 229]. With these iterative algorithms, the data point with the largest upper confidence bound (mean plus multiple of standard deviation) is evaluated, then the model is updated, and this process is repeated until convergence. This simple sampling rule simultaneously chooses points which are predicted to be optimal and uncertain, and implicitly trades off exploitation and exploration. When optimizing Gaussian processes, the UCB algorithm is guaranteed to converge to the optimal solution and displays fast convergence for a wide variety of kernel functions [229].

Theoretically, an upper confidence bound algorithm applied to a Gaussian process landscape model can be used to design optimized protein sequences. Given the current experimental data, a Gaussian process model can be trained and used to design a sequence that maximizes the upper confidence bound. This sequence can be synthesized, experimentally characterized, and used to update the model. This process can be repeated until convergence, or until the desired protein sequence properties are achieved. In early iterations this search will be dominated by exploration of the landscape, but as confidence grows the algorithm will begin to climb the landscape.

We tested the ability of upper confidence bound protein sequence design to optimize the thermostability of chimeric cytochrome P450s. To avoid the high cost of gene synthesis, we restricted our design space to single and double crossover chimeras that could be constructed from currently available chimeric P450s, a set estimated to contain $\sim 10^{10}$ unique sequences. A Gaussian process model was trained on the 261 currently available thermostability measurements. With this model, UCB optimal sequences were found using a Monte Carlo search algorithm and five sequences were chosen using a batch-mode UCB selection criteria [230]. After construction and expression, the thermostabilities of these sequences were measured. In this first round, we identified a sequence (UCBr1c4) with a thermostability (T_{50}) of 65.1 °C, higher than any chimeric P450 characterized to date. The Gaussian process model was then updated with these new data points, and the process was repeated. This UCB sequence optimization was performed for four iterations, the results are shown in Figure 5.3 and the sequences are represented schematically in Supplementary Figure 5.11.

While these 18 new sequences provide a diverse sampling of the thermostability landscape at high elevations (on average 5.1 °C more stable than the most stable parent), none are significantly more stable than the previously identified most stable chimeric P450 [86]. As a check of the current Gaussian process model, we designed a sequence (LCB1) with a maximized lower-confidence bound - a sequence predicted to be stabilized with high certainty. This sequence was constructed, expressed, and characterized, resulting in a thermostability of 67.2 °C. LCB1 is nearly 3 °C more stable than the previously identified most stable chimera and 12 °C more stable than the most stable parent. This sequence differs from the previously identified most stable chimera by 10 mutations.

These results verify the Gaussian process model is working and suggest the upper confidence bound sequence optimization is still in its exploration phase. The relatively slow convergence observed here may be due to restrictions imposed by the sequence construction constraints. During the sequence optimization, it was common to observe sequences in the current generation to be composed of fragments from sequences of the previous generation, an indication that the accessibility of sequences is limiting. This could of course be mitigated with total gene synthesis, but also highlights an interesting feature of constructing new sequences from previously constructed sequences: the number of feasible sequences increases for each iteration of the UCB sequence optimization. The sequence optimization algorithm could possibly be accelerated by considering a sequence's potential to generate UCB optimal sequences in subsequent generations, in addition to its current UCB optimality.

5.4 Materials and Methods

5.4.1 Gaussian process regression and classification

To provide a notion of distance within Gaussian process landscapes, we developed a structure-based kernel function. Here, a protein structure is represented with its residue-residue contact map. The residue contact map for cytochrome P450 was generated using all structures in the Protein Data Bank which have at least 50% sequence identity to one of the parents. Within each of these 91 protein chains, a residue pair was considered contacting if they contained any heavy atoms within 4.5 Å. For the final contact map, a residue pair was considered contacting if the pair was contacting in more than 50% of the P450 chains.

The structure of a specific sequence $\mathbf{s}$ can be described by the amino acids present for each residue-residue contact, and this information can be encoded with a binary indicator vector $\mathbf{x}$. The structure-based kernel function is defined as

$$k(\mathbf{s}_i, \mathbf{s}_j) = \sigma_p \mathbf{x}_i \cdot \mathbf{x}_j, \tag{5.3}$$

where the hyperparameter σ_p corresponds to the prior variance of a single contact, which describes how quickly the landscape is expected to change.

When modeling continuous sequence properties (regression), we used the analytical solutions for the posterior distribution given by equations 5.1 and 5.2 [218]. The hyperparameters σ_p and σ_n were found by maximizing the marginalized likelihood function or cross-validation. When modeling binary sequence properties (classification), we used Laplace's method to approximate the posterior distribution [218]. The kernel hyperparameter σ_p was found by maximizing the marginalized likelihood function.

5.4.2 Experimental design

The experimental design objective was to find the set of sequences S that maximize the expected value of the mutual information $\mathrm{E}[I(S; L)]$. For our Gaussian process model, this is equivalent to maximizing the expected value of the Shannon entropy $\mathrm{E}[H(S)]$, which is given by

$$\mathrm{E}[H(S)] = \sum_{A \in \mathcal{P}(S)} \left[H(A) \prod_{\mathbf{s} \in A} p_{\mathbf{s}} \prod_{\mathbf{s} \in (S \setminus A)} (1 - p_{\mathbf{s}}) \right], \tag{5.4}$$

where $\mathcal{P}(S)$ is the power set of S, $p_{\mathbf{s}}$ is the probability that sequence $\mathbf{s}$ is functional based on the Gaussian process functional status classifier, and the entropy H is calculated from the multivariate Gaussian covariance, which is specified by the kernel function. Unfortunately, the cost of calculating this objective grows exponentially with the number of sequences in the set S (due to the power set). For sets of less than 10 sequences, the objective was calculated exactly. For sets of 10 sequences or more, the objective was approximated by sampling.

To maximize this objective function, we can take advantage of the guaranteed performance of greedy approximation algorithms for the maximization of submodular set functions [153]. The Shannon entropy of Gaussian random fields is a submodular set function [231]. Since submodular functions are closed under non-negative linear combinations [221], the expected value of the entropy is also submodular.

In an effort to minimize sequence construction, we restricted the experimental design to the 4716 sequences that could be easily constructed from existing chimeric P450s (single-crossover overlap extension PCR between the sequences presented in [80, 86] with library-specific primers). For the first experimental design, we conditioned the landscape's covariance matrix on the parent sequences (assuming they had been observed) and selected 20 sequences using an accelerated greedy algorithm [232]. Of these 20 chimeric sequences, 17 produced folded cytochrome P450s. For the second experimental design, we conditioned the landscape's covariance matrix on the parent sequences and the 17 new chimeras, and then selected 10 additional sequences using an accelerated greedy algorithm.

5.4.3 Cloning, expression, and purification of chimeric P450s

All chimeric cytochrome P450 genes were constructed from fragments of previously published chimeric P450s, which were originally constructed from the heme domains of CYP102A1, CYP102A2, and

CYP102A3 [220, 80, 86]. Single- and double-crossover chimeric genes were assembled using overlap extension PCR and cloned into pCWori (P450-specific vector [128]) or pET22b expression vectors containing a C-terminal 6xHis tag. The correct construction of all genes was confirmed by DNA sequencing with forward and reverse primers.

Plasmid DNA was transformed into *E. coli* BL21(DE3), and the resulting transformants were used to inoculate a Luria broth (LB) starter culture supplemented with 100 μg/ml ampicillin. These starter cultures were grown overnight shaking at 37 °C and then diluted 1:100 in fresh terrific broth (TB) containing 100 μg/ml ampicillin and 500 μM δ-aminolevulinic acid. These TB cultures were grown for 3 hours at 37 °C, then protein expression was induced with 500 μM IPTG for 24 hours shaking at 30 °C. After protein expression, the cells were collected by centrifugation and stored at -20 °C.

For the enzymatic activity and binding affinity measurements (chimeric P450s ED1–ED30), frozen cell pellets were thawed and resuspended in 25 mM Tris, 200 mM NaCl, 20 mM imidazole, pH 8.0 containing 0.5 mg/ml lysozyme, and 0.05 mg/ml DNAse I. Clarified cell lysates were prepared by sonification for 2 minutes, followed by centrifugation at 75,000 RCF for 30 minutes. These clarified cell lysates were loaded onto a 5 mL HisTrap HP (high performance) Ni Sepharose column (GE Healthcare) and washed with 50 mL wash buffer (25 mM Tris, 200 mM NaCl, 20 mM imidazole, pH 8.0). The immobilized proteins were eluted with 25 mL elution buffer (25 mM Tris, 200 mM NaCl, 150 mM imidazole, pH 8.0). The peak fractions were pooled and buffer exchanged into 25 mM Tris, pH 8.0. Next, the proteins were loaded onto a 5 mL HiTrap Q HP anion exchange column (GE Healthcare) and washed with 20 mL 25 mM Tris, pH 8.0. The immobilized proteins were eluted with a 50 mL linear gradient of 25 mM Tris, 1 M NaCl, pH 8.0. The peak fractions were pooled, buffer exchanged into phosphate buffered saline, pH 7.4, concentrated to ∼100 μM, flash frozen in liquid nitrogen, and stored at -80 °C.

For the thermostability measurements (chimeric P450s UCBr1cX-UCBr4cX and LCB1), frozen cell pellets were thawed and resuspended in 100 mM potassium phosphate, pH 8.0. Clarified cell lysates were prepared by sonification for 2 minutes, followed by centrifugation at 75,000 RCF for 15 minutes. Thermostability measurements were performed with these freshly prepared cell extracts.

5.4.4 Characterization of P450 enzymatic activity

Purified cytochrome P450s were thawed and diluted into 100 mM EPPS, pH 8.0. Fresh stocks of substrates were prepared in 50% (v/v) DMSO and 50% (v/v) acetone. P450 peroxygenase reactions were performed in 100 mM EPPS, pH 8.0 with a final concentration of 2 μM P450, 4 mM H_2O_2, 1% DMSO, 1% acetone, and varying substrate concentrations. The following final substrate concentrations were chosen based on the compound's solubility: 100 mM 2-phenoxyethanol, 50 mM ethoxybenzene, 10 mM ethyl phenoxyacetate, 4 mM propranolol, 5 mM chlorzoxazone, and 2 mM 11-phenoxyundecanoic acid. Reactions were carried out for two hours at room temperature and then stopped with quench buffer (final concentration of 50 mM NaOH, 2 M urea). Hydroxylation of each substrate, at the appropriate positions, leads to phenolic byproducts. These phenolic compounds can be coupled to 4-aminoantipyrene (4-AAP) to form a red compound, which is detectable at 500 nm [233]. The 'enzymatic activity' values are the raw absorbance increase at 500 nm, which is proportional to the total substrate turnovers per enzyme after two hours. All measurements were performed in triplicate and the median enzymatic activity values are reported.

5.4.5 Characterization of P450 binding affinity

Purified cytochrome P450s were thawed and diluted into 2X phosphate buffered saline (PBS), pH 7.4. Fresh stocks of dopamine and serotonin were also prepared in 2X PBS, pH 7.4. All binding assays were performed in 2X PBS, pH 7.4 with a final concentration of 4 μM P450 and logarithmically-spaced ligand concentrations ranging from 2.8 μM to 500 mM. For each titration, the proportion of bound P450 was determined by the relative shift in the Soret peak [100]. The dissociation constant (K_d) was determined by fitting a two-state binding model to this ligand-binding curve. All binding assays were performed in at least triplicate and the median K_d values are reported.

5.4.6 Characterization of P450 thermostability

The cytochrome P450 concentration within freshly prepared cell extracts was determined using CO-difference spectroscopy [234]. Cell extracts were diluted to 4 μM with 100 mM potassium phosphate, pH 8.0 and arrayed into 96-well PCR plates. Using a gradient thermocycler, the samples were heated over multiple temperatures (typically 55–70 °C) for 10 minutes. The samples were then centrifuged and the remaining P450 was quantified using CO-difference spectroscopy [234]. The T_{50} (temperature where 50% of the protein is inactivated in 10 minutes) was determined by fitting a shifted sigmoid function to the thermal inactivation curves. All measurements were performed in at least triplicate and the median T_{50} values are reported.

5.4.7 Outlier detection

Outlying sequences were identified based on two different criteria. The first was calculated by removing a sequence (or set of sequences) from the data set, training the Gaussian process model on the remainder of the data, and evaluating the predictive likelihood of the omitted data points [235]. Here, outliers are data points that are very unlikely given the remainder of the data set. The second criteria was based on the leave-one-out cross-validated predictive accuracy within the data set when various sequences were removed. By this criteria, outliers are data points that significantly improve the predictive accuracy of the model when they are removed from the data set.

These two criteria were used to detect the presence of outliers in all six enzymatic activity and both binding affinity data sets. ED7, ED9, and ED28 were classified as outliers in all eight of these data sets. In addition, ED12 was an outlier for enzymatic activity on 2-phenoxyethanol, and ED10 was an outlier for enzymatic activity on ethoxybenzene and ethyl phenoxyacetate. Four of these outliers have Sort peaks that are shifted relative to the remainder of the data set (Supplementary Figure 5.8).

5.4.8 Poisson-Boltzmann calculations

In an effort to understand the observed variation in the chimeric P450s properties, we performed Poisson-Boltzmann calculations to estimate the contribution of long-range electrostatic interactions. DelPhi was used to calculate the electrostatic component of the free energy of binding between dopamine and all chimeras within the data set (three parents and ED1-ED30) [236]. The dopamine ligand was used because a crystal structure of a CYP102A1 variant bound to dopamine was available [237]. The results for dopamine should apply to the other ligands and substrates because they are of similar size and net charge.

Using the crystal structure of a CYP102A1 variant bound to dopamine as a template, we modeled the structure of each chimeric P450 using CHOMP [238]. These structural models had fixed backbones with rotamers optimized with respect to the Rosetta energy function. The atomic radii of the heme and dopamine atoms were chosen to match the equivalent atom types in the DelPhi parameter file. The partial charges of the heme and dopamine atoms were calculated with the Electrostatic Potential (ESP) module of NWChem [239]. All Poisson-Boltzmann calculations were run with a 100 mM salt concentration. The binding energy for each chimeric P450 was calculated by taking the sum of the grid energy for individual dopamine and protein molecules and subtracting this from the grid energy of the bound complex.

Across all chimeric P450s within the data set, the standard deviation in the electrostatic component of the binding free energy is calculated to be 0.12 kcal/mol and the total range is 0.42 kcal/mol. Experimentally, we observe the standard deviation of the binding free energy to be 0.66 kcal/mol with a total range of 2.17 kcal/mol. From these calculations, we estimate that long-range electrostatic interactions could be contributing to $\sim$20% of the energetic differences between the chimeric P450s.

5.5 Discussion

We have demonstrated the ability to model the protein fitness landscape with quantitative accuracy using Gaussian process regression and classification. Within the landscape, the relationship between pairs of sequences is specified by a structure-based kernel function, which is derived from the principle that sequences with similar structures are more likely to have similar properties. With this metric over sequence space, a full probabilistic description of the landscape can be inferred from experimental data. These Gaussian process landscapes are able to describe various protein sequence properties including functional status, thermostability, enzymatic activity, and binding affinity. These results suggest Gaussian process models may be applicable to most properties that display significant variation within an experimental data set.

The predictive ability of these Gaussian process landscape models is unprecedented. There are currently no models which can achieve this level of accuracy across such a large and diverse set of sequences. In addition, Gaussian processes explicit representation of model uncertainty provides a valuable guide for knowing when a prediction should be trusted. Gaussian process models improve upon previously developed statistical models [86, 151] by providing increased predictive accuracy across a significantly larger portion of sequence space. Since Gaussian process models must be trained on data from a specific protein family, they are less general than the models traditionally used for protein design. However, this loss of generality comes with substantial increases in accuracy for a wide variety of sequence properties. Many properties are extremely difficult to model accurately with energy-based scoring functions because their origins are unknown or may involve subtle (possibly dynamic) structural changes which are not easily represented with current methods. Since Gaussian process models are trained on experimental data, they capture all the factors which contribute to the sequence property being modeled, whether they are known or not.

The performance of Gaussian process landscapes could possibly be improved by the use of alternate kernel functions. The structure-based kernel function is based on the assumption that a residue-level contact potential is able to describe the properties of protein sequences. While this assumption has a biophysical basis [240], it excludes the possibility of higher-order interactions. The use of polynomial kernels can very easily include interactions up to any order [218]. A promising direction for kernel development might be to make use of any prior knowledge of interactions from current statistical or physical models. If implemented properly, this could significantly reduce the amount of experimental data required to learn or optimize the landscape.

The Bayesian treatment of uncertainty in Gaussian process landscapes allows for efficient explorations in the high-dimensional protein sequence space. We developed an experimental design algorithm which is able to identify the most informative sequences within the landscape. This algorithm was used to generate a set of 29 highly-informative chimeric cytochrome P450s. Gaussian process models can also be used to design optimized protein sequences using upper confidence bound (UCB) algorithms. UCB sequence design iteratively climbs unknown landscapes by deciding when to continue exploring or when to exploit the current model. This algorithm was used to design a chimeric P450 nearly 3 °C more stable than an already highly-optimized chimeric P450, and 12 °C more stable than the most stable parent P450.

Improving the upper confidence bound protein design algorithm is an important area to explore. While fully synthesized genes provide unlimited sequence accessibly, their cost is still prohibitively high. For this reason we constructed new sequences from sequences that had already been constructed. Doing this optimally requires forward-looking sequence selection strategies, which consider a sequence's current utility and its potential to produce high-utility sequences in future generations. This has been done with the UCT (upper confidence bounds applied to trees) algorithm, which estimates how a decision will affect future outcomes by sampling randomly down the decision tree [241]. Even simpler heuristics can probably hasten the sequence optimization. For example, imagine sampling sequences which have lower-confidence bounds greater than the current maxima, or if they do not exist, performing standard UCB optimization. This strategy exploits earlier than UCB, with the intention of building off of any improvements.

By modeling multiple sequence properties, we obtain a more complete description of a protein. Instead of modeling each property individually (as we did above), all properties and their relationships can be modeled simultaneously using multi-task learning [242]. These models provide improved predictive ability because they take advantage of any correlations that might exist between properties. This more holistic description of a protein is important when designing sequences which must satisfy a number of criteria, as most proteins do. These models also highlight the multifaceted nature of the protein fitness landscape by revealing how different properties are related throughout sequence space. This information can help explain the presence of evolutionary trends and constraints. For example, we have previously evolved a cytochrome P450 for high dopamine binding affinity, which inadvertently resulted in high serotonin binding affinity as well [100]. This result is not surprising, given the strong correlation between dopamine and serotonin binding affinity observed within our sampling of the landscape (Supplementary Table 3). However, this is only a trend in the landscape, not a constraint: P450s have been engineered with high affinity and specificity for serotonin or dopamine [237].

While all the results presented here are based on chimeric proteins, Gaussian process models are applicable to any set of sequences which fold into the same three-dimensional structure. Other training sets could include naturally occurring homologs, point mutant libraries, or computationally designed libraries. For example, training these models on large libraries of point mutants would allow prediction of the effect of combinations of mutations, accounting for both additive and pairwise interactions. In general, predictions should be restricted to sequences that contain the same amino acids at each position as observed in the training set, which will minimize the model's uncertainty. This makes chimeric protein libraries particularly desirable training sets because they uniformly sample a massive combinatorial space. On top of this, the sequences within chimera libraries have a high probability of functioning [56] and display significant functional diversity [85, 86].

Protein sequence space is vast, and hidden within it are engineering solutions to a wide-variety of problems and even clues about the evolutionary history of life. To find these things, we must understand the mapping from protein sequence to function, which involves an extraordinarily complex balance of numerous physical interactions. While this mapping is extremely challenging to describe from a physical perspective, statistical models overlook these details and instead learn what the experimental data is telling them. As technology for high-throughput experimentation advances, this class of models could play an increasing role in understanding how proteins function.

5.6 Figures

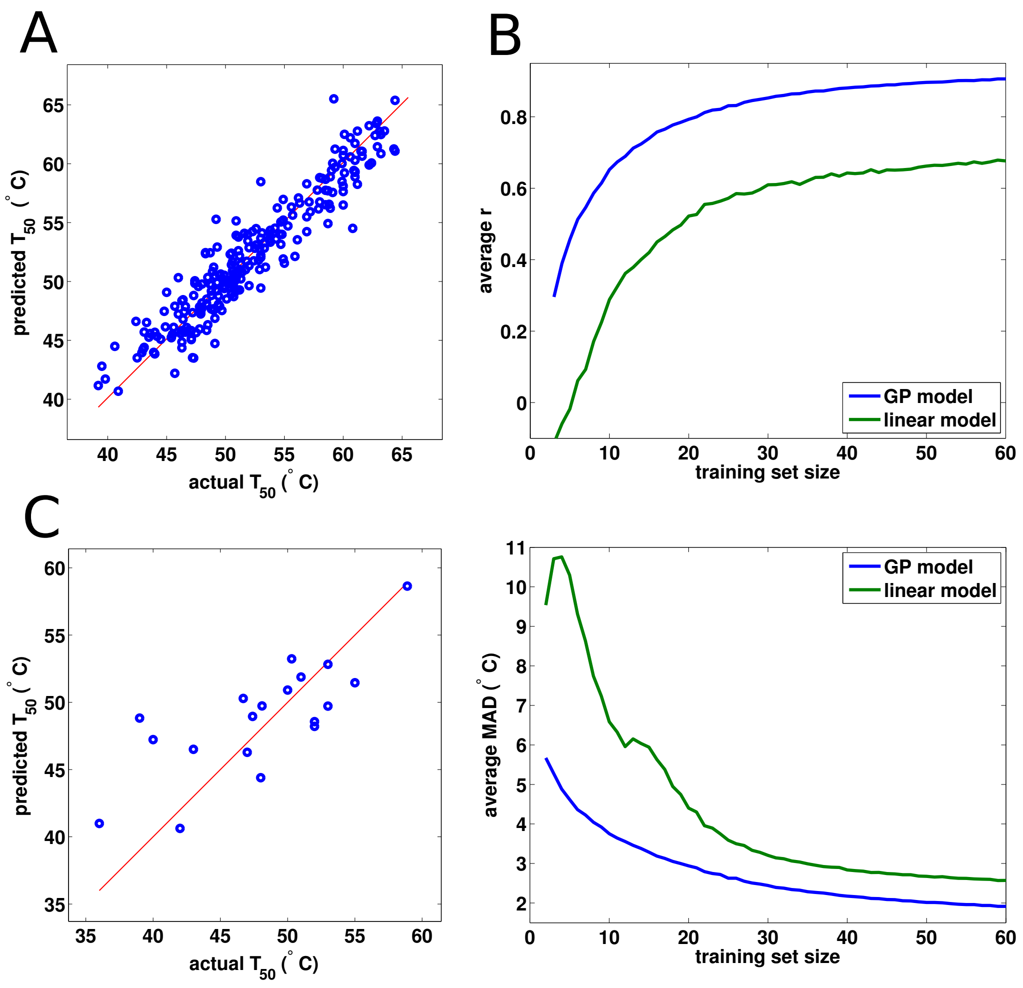


Figure 5.1: Predictive ability of Gaussian process models. (*A*) The Gaussian process model shows excellent predictive ability ($r = 0.95$, $MAD = 1.4$ °C) on a previously published cytochrome P450 data set. Shown are 10-fold cross-validated predictions. (*B*) A comparison of the Gaussian process and linear regression models was made by sampling random training sets of various sizes and evaluating the predictive performance. For each training set size, the results are averaged over 1000 random samples. (*C*) A Gaussian process model was trained on the data set from panel *A* and used to predict the stability of a set of sequences that cannot be represented with the linear regression model. This model shows good predictive ability ($r = 0.76$, $MAD = 3.0$ °C) on these sequences which could not be modeled with previous methods. Confidence intervals for these predictions are shown in Supplementary Figure 5.6

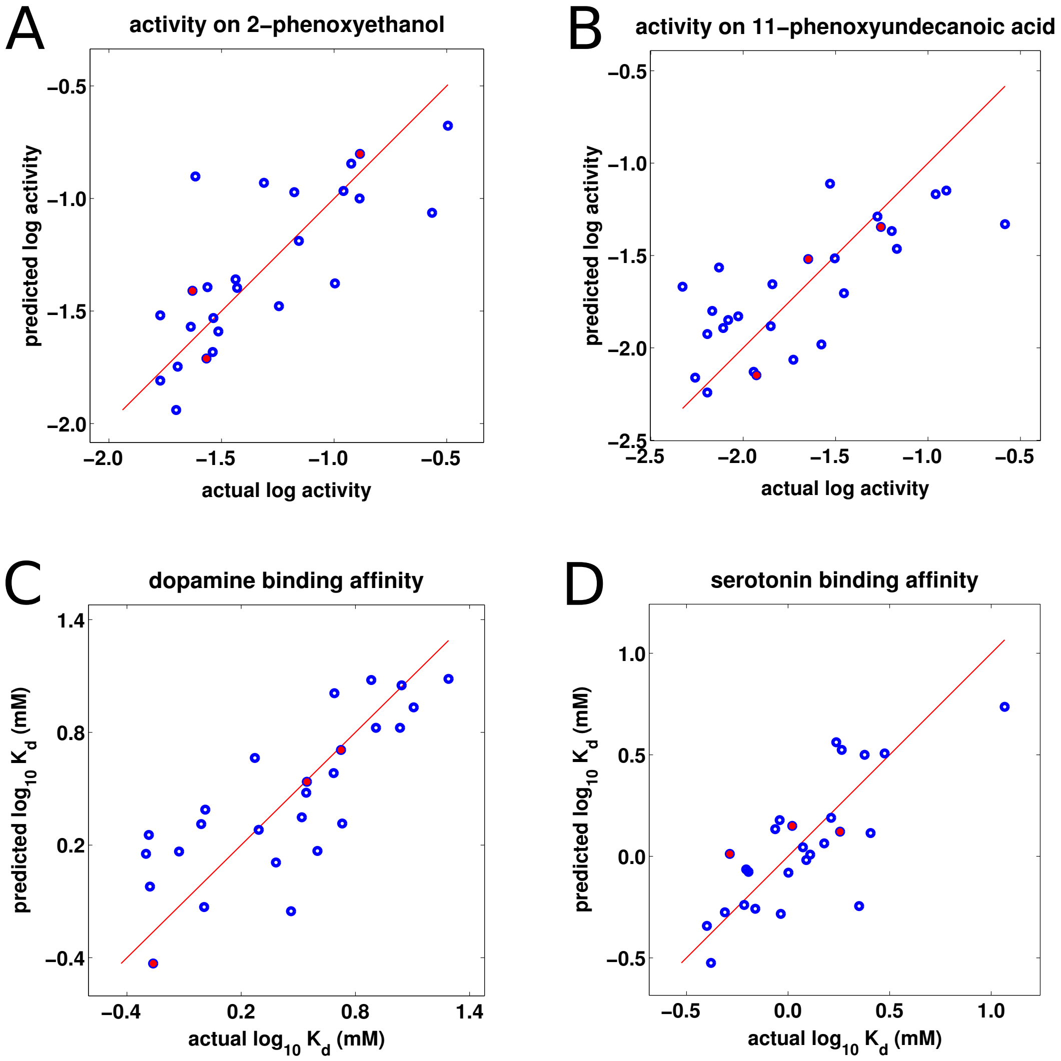


Figure 5.2: Gaussian process models for P450 enzymatic activity and binding affinity. All plots show leave-one-out cross-validated predictions and the red points correspond to the three parent sequences. (*A*) Predictions for enzymatic activity on 2-phenoxyethanol ($r = 0.77$), outliers ED7, ED9, ED12, ED28 were not included. (*B*) Predictions for enzymatic activity on 11-phenoxyundecanoic acid ($r = 0.74$), outliers ED7, ED9, ED28 were not included. (*C*) Predictions for binding affinity on dopamine ($r = 0.79$), outliers ED7, ED9, ED2 were not included. (*D*) Predictions for binding affinity on serotonin ($r = 0.78$), outliers ED7, ED9, ED28 were not included. The correlation coefficients for predictions on the other substrates are: ethoxybenzene: 0.77, ethyl phenoxyacetate: 0.80, propranolol: 0.70, chlorzoxazone: 0.38 (scatter plots are shown in Supplementary Figure 5.9).

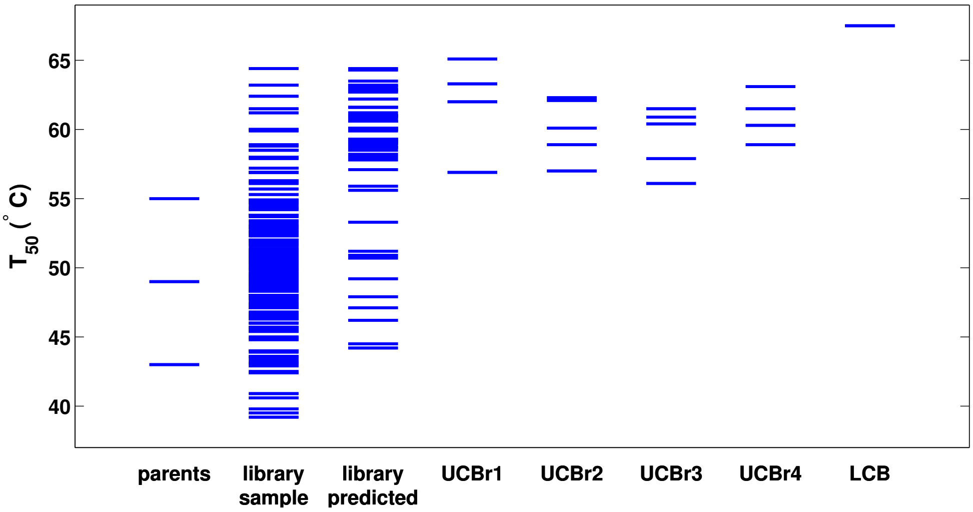
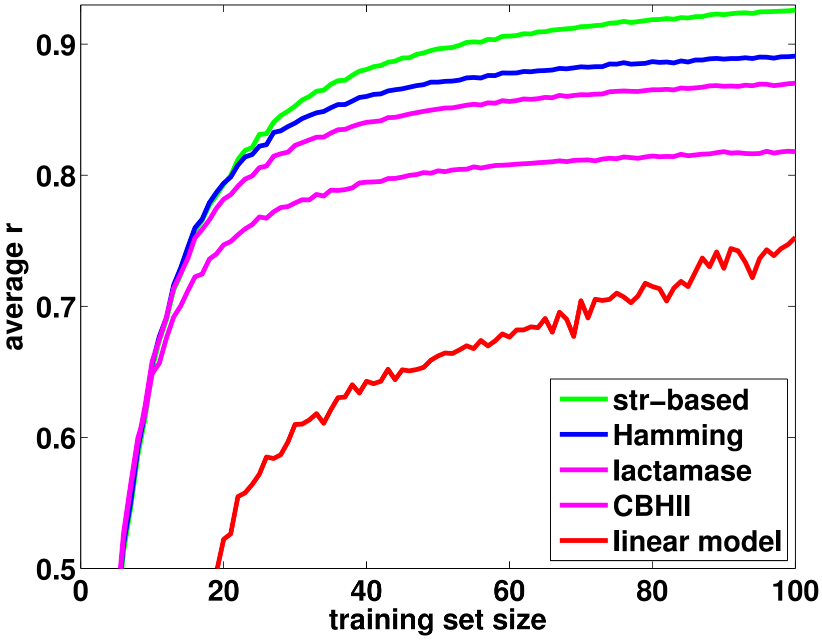
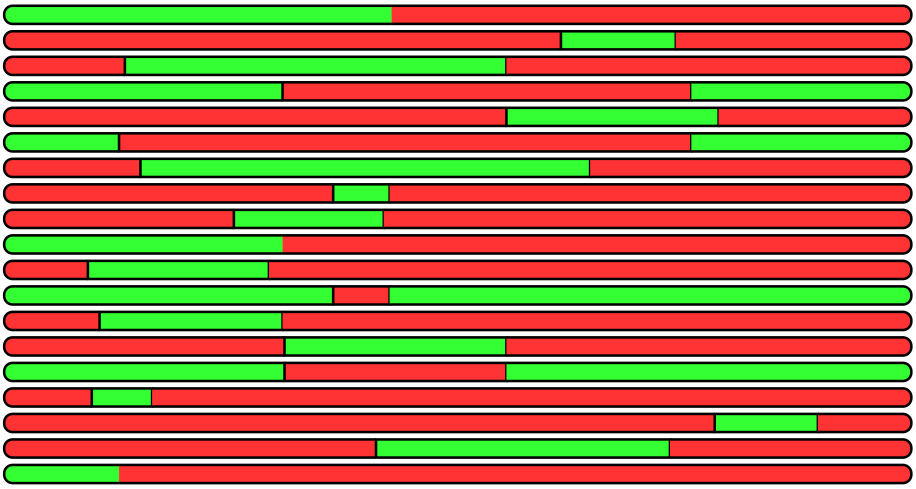


Figure 5.3: Upper confidence bound sequence optimization. The first column shows the thermostabilities of the three parent cytochrome P450s. The next two columns show the results from a large sampling of a P450 recombination library, followed by sequences that were predicted to be stabilized using a linear regression model [86]. The next four columns show four rounds of batch-mode upper confidence bound sequence optimization, providing a diverse sampling of thermostabilized sequences. Note for UCB rounds 1 and 4, one chimeric P450 was not evaluated because of difficulties encountered during the sequence construction. The final column shows the single lower-confidence bound prediction (LCB1). LCB1 has a thermostability of 67.2 °C, which is significantly stabilized relative to all previously identified chimeric P450s. All UCB and LCB sequences are represented schematically in Supplementary Figure 5.11.

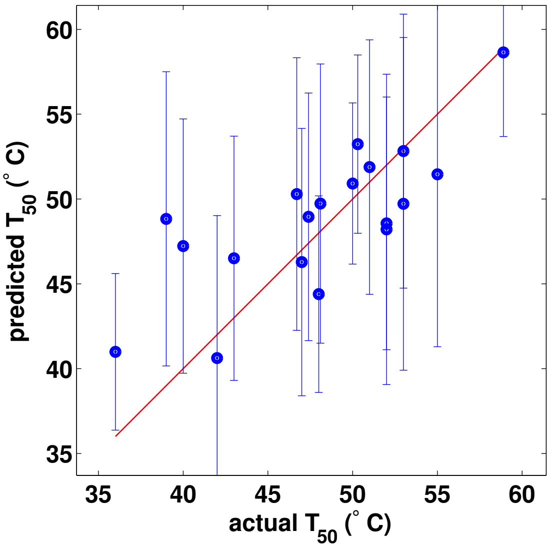
5.7 Supplementary Figures



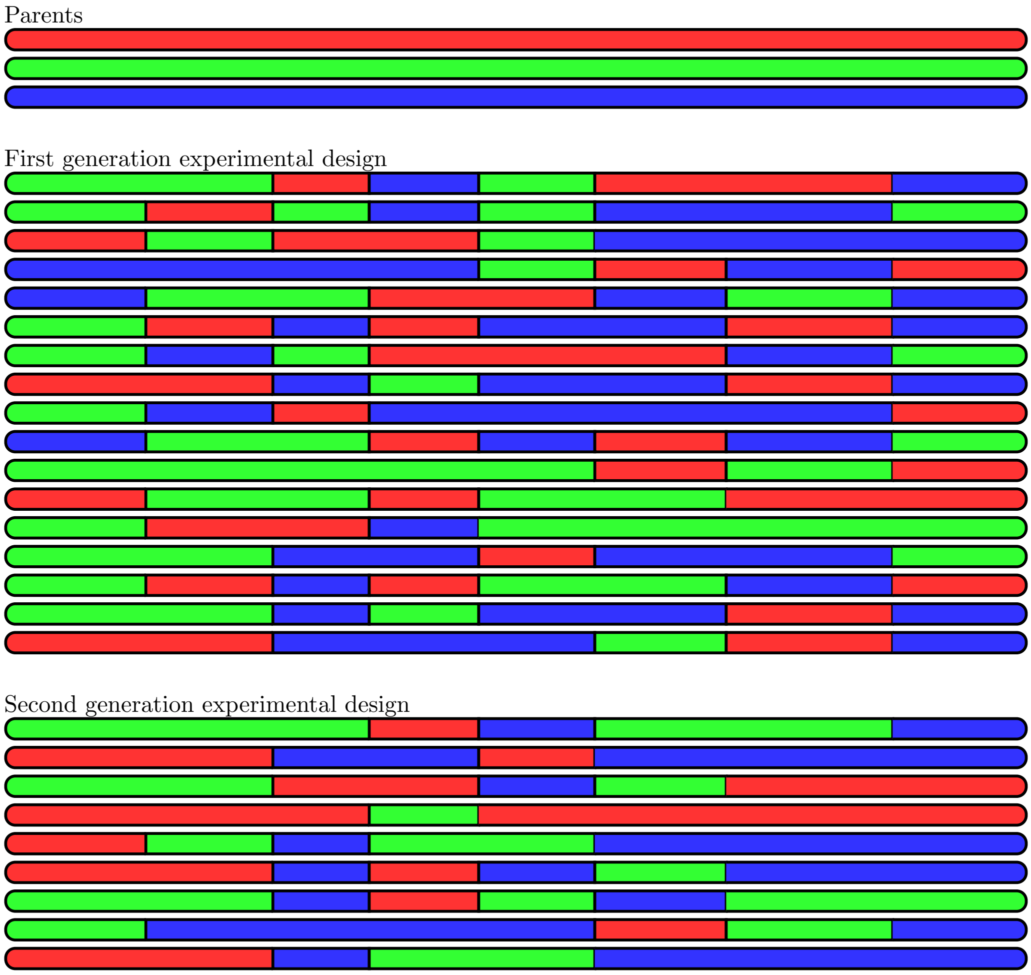
Supplementary Figure 5.4: The correlation coefficient of Gaussian process models as a function of training set size (calculated as in Figure 5.1*B*). The structure-based kernel function (green) outperforms the hamming distance (blue), structure-based kernels with the incorrect structure (magenta), and the linear regression model (red).



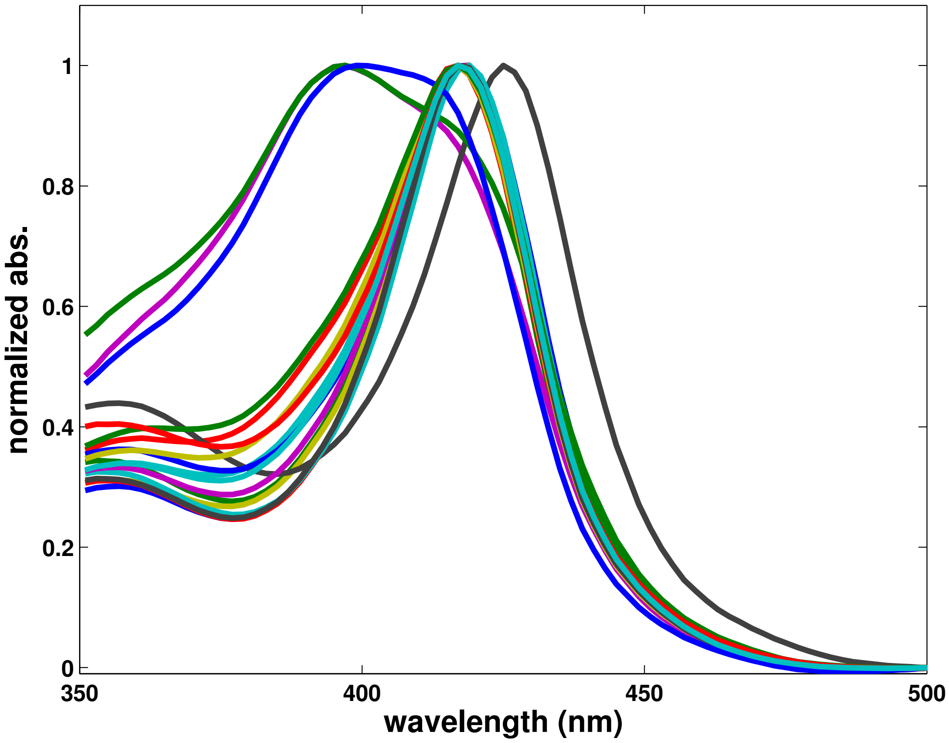
Supplementary Figure 5.5: Schematic representation of the single- and double-crossover chimeric P450s between CYP102A1 and CYP102A2. 14 of these sequences are from [220] and the remaining five sequences are unpublished (presented in Supplementary Table 5.1). Parents CYP102A1 and CYP102A2 are represented with red and green sequence fragments, respectively.



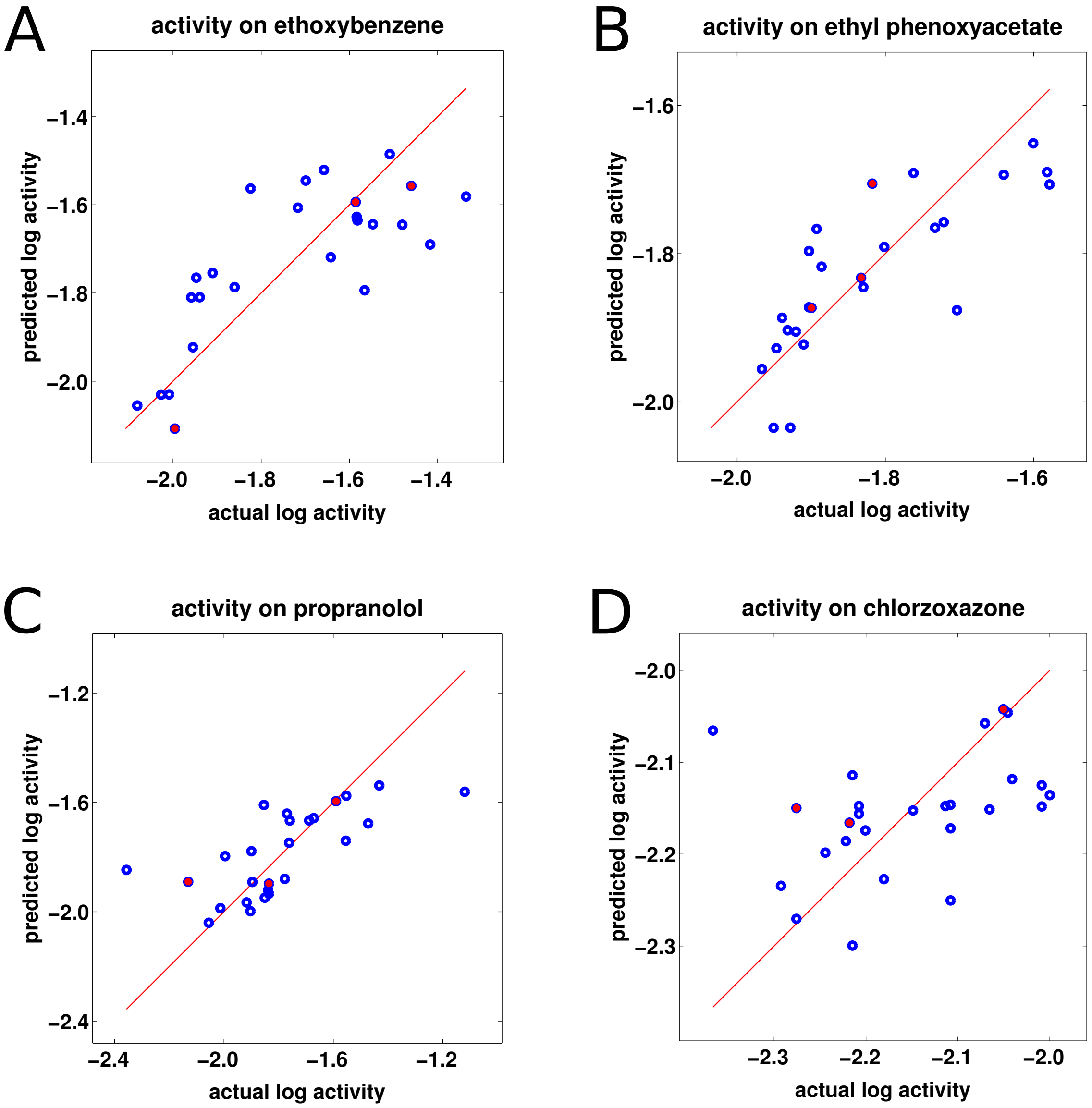
Supplementary Figure 5.6: Uncertainty in Gaussian process predictions. The error bars show the 95% confidence intervals for the predictions shown in Figure 5.1*C*. 17/19 experimental measurements fall within the model's 95% confidence intervals.


Parents
First generation experimental design
Second generation experimental design

Supplementary Figure 5.7: Schematic representation of the folded chimeric P450s within the experimentally designed set of sequences. Parents CYP102A1, CYP102A2, CYP102A3 are represented with red, green, and blue sequence fragments, respectively.



Supplementary Figure 5.8: Absorbance spectra of the 26 chimeric P450s within the experimentally designed set of sequences. Three of the chimeras (ED7, ED12, and ED28) have a blue-shifted soret peak, indicative a of high-spin heme, which is normally associated with reduced solvent accessibility in the active site. ED9 has a red-shifted soret peak, which suggests the presence of a distal heme ligand.

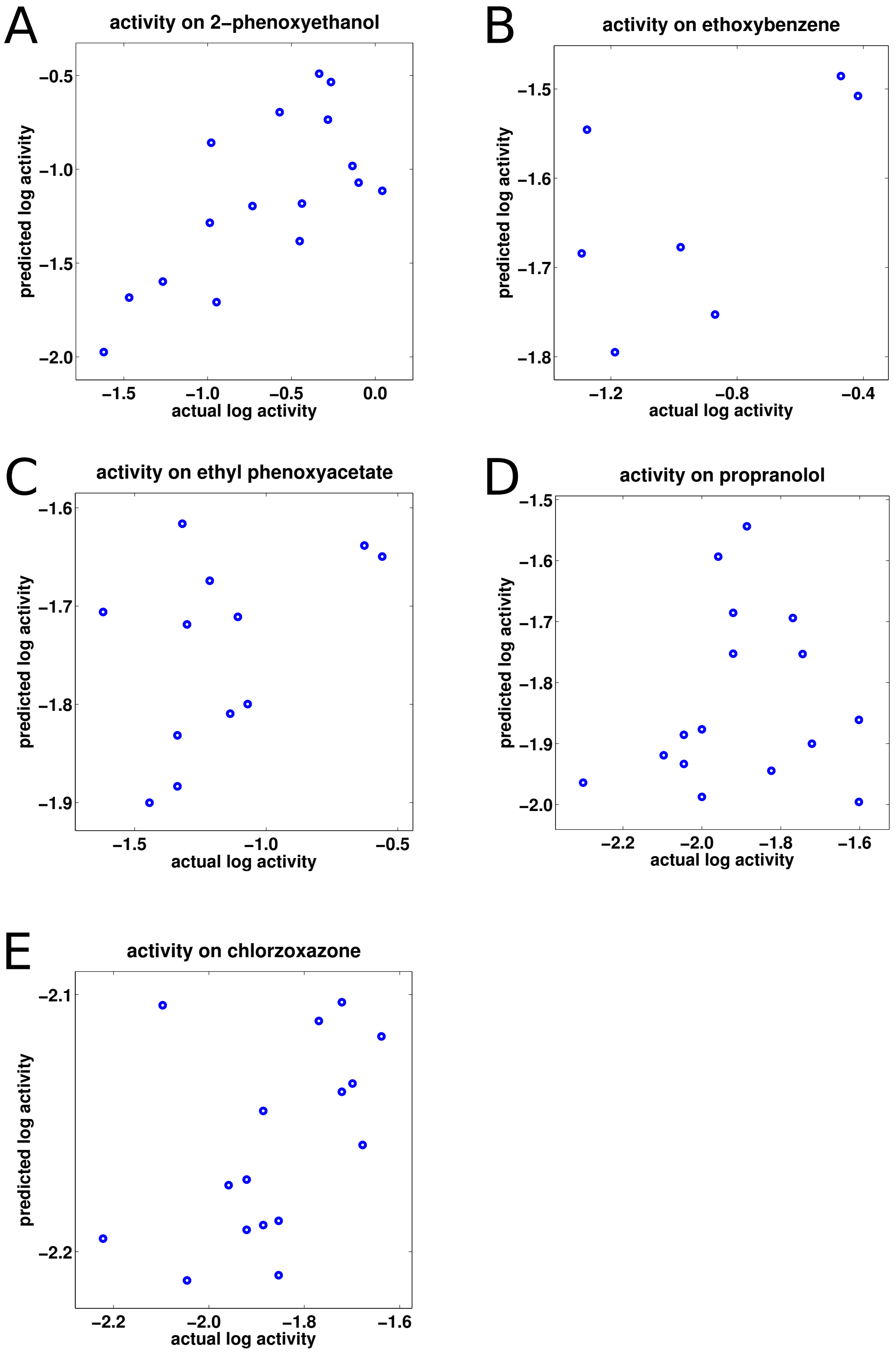


Supplementary Figure 5.9: Additional Gaussian process models for P450 enzymatic activity. All plots show leave-one-out cross-validated predictions and the red points correspond to the three parent sequences (*A*) Predictions for enzymatic activity on ethoxybenzene ($r = 0.77$), outliers ED7, ED9, ED10, ED28 were not included. (*B*) Predictions for enzymatic activity on ethyl phenoxyacetate ($r = 0.80$), outliers ED7, ED9, ED10, ED28 were not included. (*C*) Predictions for enzymatic activity on propranolol ($r = 0.70$), outliers ED7, ED9, ED28 were not included. (*D*) Predictions for enzymatic activity on chlorzoxazone ($r = 0.38$), outliers ED7, ED9, ED28 were not included.

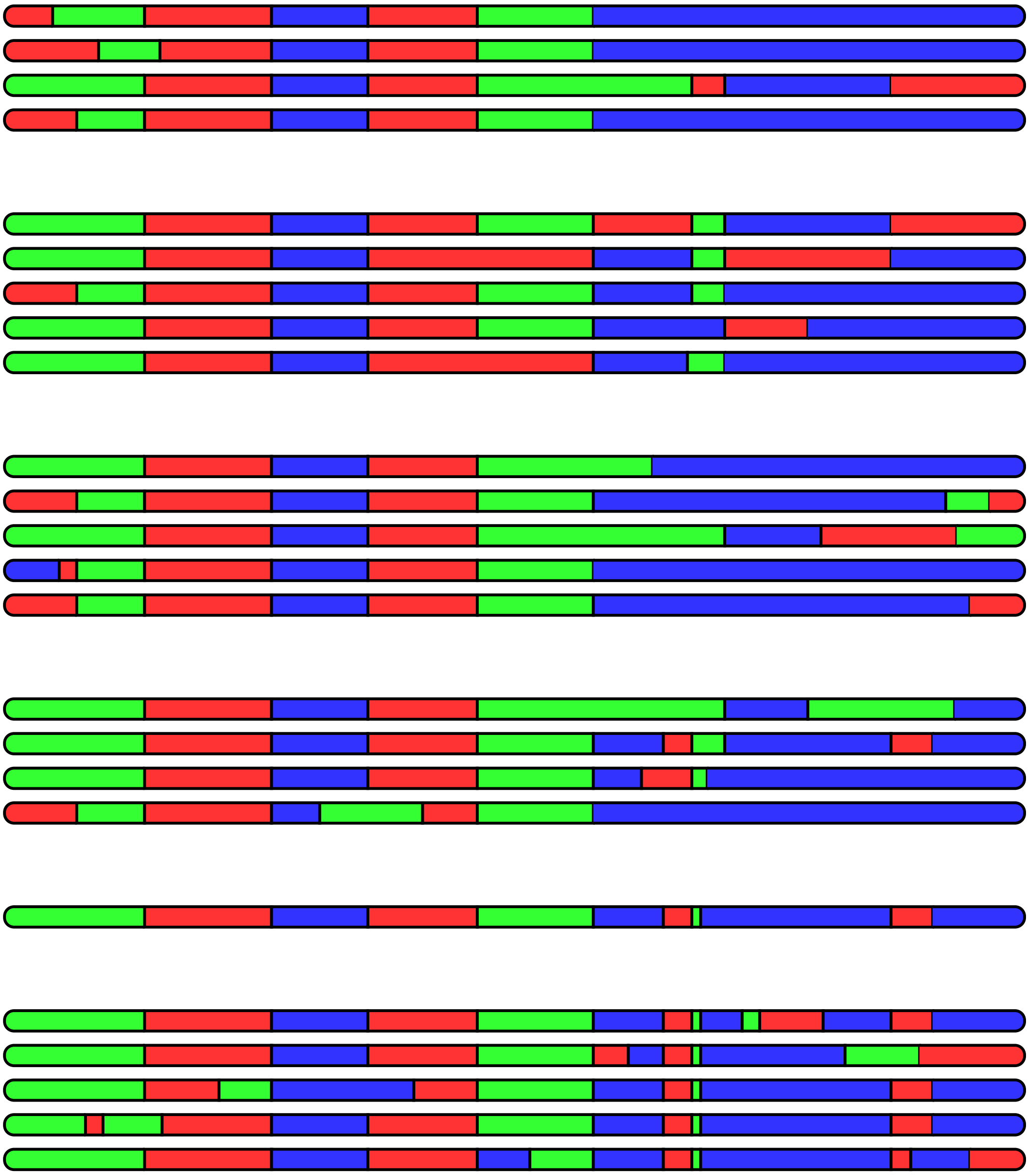
5.8 Supplementary Tables

```
 >C60, T50=58.9
MKETSPIPQPKTFGPLGNLPLIDKDKPTLSLIKLAEEQGPIFQIHTPAGTTIVVSGHELVKEACDESRFDKNL
SQALKFVRDFAGDGLATSWTHEKNWKKAHNILLPSFSQQAMKGYHAMMVDIAVQLVQKWERLNADEHIEVPED
MTRLTLDTIGLCGFNYRFNSFYRDQPHPFITSMVRALDEAMNKLQRANPDDPAYDENKRQFQEDIKVMNDLVD
KIIADRKASGEQSDDLLTHMLNGKDPETGEPLDDENIRYQIITFLIAGHETTSGLLSFALYFLVKNPHVLQKA
AEEAARVLVDPVPSYKQVKQLKYVGMVLNEALRLWPTAPAFSLYAKEDTVLGGEYPLEKGDELMVLIPQLHRD
KTIWGDDVEEFRPERFENPSAIPQHAFKPFGNGQRACIGQQFALHEATLVLGMMLKHFDFEDHTNYELDIKET
LTLKPEGFVVKAKSKKIPLGGIPSPST
>C142, T50=50.3
MKETSPIPQPKTFGPLGNLPLIDKDKPTLSLIKLAEEQGPIFQIHTPAGTTIVVSGHELVKEVCDEERFDKSI
EGALEKVRAFSGDGLATSWTHEPNWRKAHNILMPTFSQRAMKDYHEKMVDIAVQLIQKWARLNPNEAVDVPED
MTRLTLDTIGLCGFNYRFNSFYRDQPHPFITSMVRALDEAMNKLQRANPDDPAYDENKRQFQEDIKVMNDLVD
KIIADRKASGEQSDDLLTHMLNGKDPETGEPLDDENIRYQIITFLIAGHETTSGLLSFALYFLVKNPHVLQKA
AEEAARVLVDPVPSYKQVKQLKYVGMVLNEALRLWPTAPAFSLYAKEDTVLGGEYPLEKGDELMVLIPQLHRD
KTIWGDDVEEFRPERFENPSAIPQHAFKPFGNGQRACIGQQFALHEATLVLGMMLKHFDFEDHTNYELDIKET
LTLKPEGFVVKAKSKKIPLGGIPSPST
>C60_354, T50=46.7
MKETSPIPQPKTFGPLGNLPLIDKDKPTLSLIKLAEEQGPIFQIHTPAGTTIVVSGHELVKEACDESRFDKNL
SQALKFVRDFAGDGLATSWTHEKNWKKAHNILLPSFSQQAMKGYHAMMVDIAVQLVQKWERLNADEHIEVPED
MTRLTLDTIGLCGFNYRFNSFYRDQPHPFITSMVRALDEAMNKLQRANPDDPAYDENKRQFQEDIKVMNDLVD
KIIADRKASGEQSDDLLTHMLNGKDPETGEPLDDENIRYQIITFLIAGHETTSGLLSFALYFLVKNPHVLQKA
AEEAARVLVDPVPSYKQVKQLKYVGMVLNEALRLWPTAPAFSLYAKEDTVLGGEYPLEKGDRISVLIPQLHRD
RDAWGKDAEEFRPERFEHQDQVPHHAYKPFGNGQRACIGMQFALHEATLVLGMILKYFTLIDHENYELDIKQT
LTLKPGDFHISVQSRHQEAIHADVQAAE
>C142_354, T50=48.1
MKETSPIPQPKTFGPLGNLPLIDKDKPTLSLIKLAEEQGPIFQIHTPAGTTIVVSGHELVKEVCDEERFDKSI
EGALEKVRAFSGDGLATSWTHEPNWRKAHNILMPTFSQRAMKDYHEKMVDIAVQLIQKWARLNPNEAVDVPED
MTRLTLDTIGLCGFNYRFNSFYRDQPHPFITSMVRALDEAMNKLQRANPDDPAYDENKRQFQEDIKVMNDLVD
KIIADRKASGEQSDDLLTHMLNGKDPETGEPLDDENIRYQIITFLIAGHETTSGLLSFALYFLVKNPHVLQKA
AEEAARVLVDPVPSYKQVKQLKYVGMVLNEALRLWPTAPAFSLYAKEDTVLGGEYPLEKGDRISVLIPQLHRD
RDAWGKDAEEFRPERFEHQDQVPHHAYKPFGNGQRACIGMQFALHEATLVLGMILKYFTLIDHENYELDIKQT
LTLKPGDFHISVQSRHQEAIHADVQAAE
>C200, T50=47.4
MKETSPIPQPKTFGPLGNLPLIDKDKPTLSLIKLAEEQGPIFQIHTPAGTTIVVSGHELVKEVCDEERFDKSI
EGALEKVRAFSGDGLATSWTHEPNWRKAHNILMPTFSQRAMKDYHEKMVDIAVQLIQKWARLNPNEAVDVPGD
MTRLTLDTIGLCGFNYRFNSYYRETPHPFINSMVRALDEAMHQMQRLDVQDKLMDENKRQFQEDIKVMNDLVD
KIIADRKASGEQSDDLLTHMLNGKDPETGEPLDDENIRYQIITFLIAGHETTSGLLSFALYFLVKNPHVLQKA
AEEAARVLVDPVPSYKQVKQLKYVGMVLNEALRLWPTAPAFSLYAKEDTVLGGEYPLEKGDELMVLIPQLHRD
KTIWGDDVEEFRPERFENPSAIPQHAFKPFGNGQRACIGQQFALHEATLVLGMMLKHFDFEDHTNYELDIKET
LTLKPEGFVVKAKSKKIPLGGIPSPST
```

Supplementary Table 5.1: Thermostabilities (T_{50}) and sequences of unpublished single- and double-crossover chimeric P450s between CYP102A1 and CYP102A2. Thermostabilities were measured following the procedure presented in the Methods section.



Supplementary Figure 5.10: Gaussian process model predictions on an independent enzymatic activity data set [85]. Note the data set contained normalized activity values, whose logarithm should be monotonic with the Gaussian process predictions, but not necessarily linear. (*A*) Predictions for enzymatic activity on 2-phenoxyethanol ($r = 0.72$). (*B*) Predictions for enzymatic activity on ethoxybenzene ($r = 0.61$). (*C*) Predictions for enzymatic activity on ethyl phenoxyacetate ($r = 0.45$). (*D*) Predictions for enzymatic activity on propranolol ($r = 0.13$). (*E*) Predictions for enzymatic activity on chlorzoxazone ($r = 0.46$).



sequence name	sequence fragments	Soret peak	2PE activity	EOB activity	EPOA activity	PROP activity	CHLOR activity	11POD activity	DOP K_d (mM)	5HT Kd (mM)
PAR1	11111111	417	13.03	2.60	1.52	2.57	0.61	5.56	0.55	0.52
PAR2	22222222	417	2.71	1.01	1.26	1.46	0.53	1.18	3.51	1.05
PAR3	33333333	417	2.35	3.47	1.47	0.74	0.89	2.25	5.30	1.80
ED1	13222212	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
ED2	22132113	419	5.69	4.62	2.51	1.45	0.98	1.87	5.38	2.38
ED3	21232332	417	1.69	0.94	1.08	1.01	0.85	0.68	19.47	1.51
ED4	31233233	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
ED5	12112333	417	32.04	3.31	2.62	1.73	0.63	3.13	0.50	1.18
ED6	33332131	417	1.99	2.61	1.17	0.88	0.91	3.51	7.64	1.73
ED7	32211323	397	4.40	1.28	1.42	2.05	0.64	2.16	8.72	18.88
ED8	21313313	417	6.98	2.20	1.85	2.79	0.62	0.83	0.53	0.42
ED9	23211132	425	1.98	2.74	1.33	0.42	0.28	0.48	500.00	500.00
ED10	11323313	415	27.25	10.02	4.06	7.60	0.71	26.11	1.01	0.40
ED11	23133331	419	2.74	1.10	1.13	1.25	0.86	2.65	11.03	1.84
ED12	32213132	397	8.14	3.10	2.64	0.97	1.00	0.94	10.83	11.65
ED13	22222121	419	3.06	1.15	1.30	1.46	0.66	0.78	3.99	1.63
ED14	12212211	419	2.42	1.11	1.25	1.40	0.57	0.74	0.75	1.28
ED15	21132222	417	1.69	0.83	1.12	1.27	0.78	1.41	4.90	0.64
ED16	12331123	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
ED17	22331332	417	2.89	1.13	1.25	1.41	0.90	1.14	12.77	2.99
ED18	21312231	417	2.02	0.98	1.18	1.67	0.51	0.64	1.87	0.69
ED19	22323313	419	4.88	3.83	2.29	1.70	0.98	0.47	0.98	0.86
ED20	11333213	417	3.65	1.50	1.15	3.37	0.78	5.33	4.85	0.61
ED21	22213223	417	3.71	1.23	1.58	1.21	0.61	0.64	1.03	0.91
ED22	11331333	417	2.91	2.28	1.23	2.05	0.78	2.95	3.48	1.00
ED23	11123313	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
ED24	22113211	419	10.09	1.38	1.73	1.26	0.61	0.55	2.90	2.55
ED25	11121111	417	11.92	2.84	1.98	3.70	0.77	12.57	0.52	0.92
ED26	12322333	417	12.98	2.00	1.90	2.13	0.62	6.37	1.96	1.23
ED27	11313233	415	11.01	2.72	1.48	1.74	0.53	6.79	3.30	0.62
ED28	22312322	399	12.78	1.95	1.96	2.12	0.81	3.41	15.79	14.44
ED29	23333123	417	2.31	2.62	1.20	0.44	0.43	1.44	8.09	2.24
ED30	11322333	417	6.66	1.92	1.28	2.80	0.60	11.01	2.41	0.49

Supplementary Table 5.2: Experimental data collected on chimeric P450s. The sequences with "NaN"s as entries did not produce folded P450s. The reported enzymatic activity values are 100X the measured value. The substrate names are abbreviated as 2PE: 2-phenoxyethanol, EOB: ethoxybenzene, EPOA: ethyl phenoxyacetate, PROP: propranolol, CHLOR: chlorzoxazone, 11POD: 11-phenoxyundecanoic acid, DOP: dopamine, 5HT: serotonin.

	EOB activity	EPOA activity	PROP activity	CHLOR activity	11POD activity	DOP affinity	5HT affinity
2PE activity	0.5730	0.7943	0.6170	0.0595	0.5659	-0.4883	-0.1834
EOB activity		0.7642	0.2337	0.1111	0.4691	-0.1555	0.0065
EPOA activity			0.4324	0.2300	0.2922	-0.3520	-0.0291
PROP activity				0.2503	0.6434	-0.6178	-0.5486
CHLOR activity					0.1469	-0.1800	-0.3022
11POD activity						-0.2853	-0.3477
DOP affinity							0.8042

Supplementary Table 5.3: Pairwise correlations between the measured enzymatic activities and binding affinities. The substrate names are abbreviated as 2PE: 2-phenoxyethanol, EOB: ethoxybenzene, EPOA: ethyl phenoxyacetate, PROP: propranolol, CHLOR: chlorzoxazone, 11POD: 11-phenoxyundecanoic acid, DOP: dopamine, 5HT: serotonin. Some properties show strong correlations, such as 2-phenoxyethanol activity and ethyl phenoxyacetate activity or dopamine affinity and serotonin affinity. However, many of the pairwise correlations are less than the predictive ability of the model, suggesting the model is able to capture independent sequence properties.

Bibliography

[1] P. A. Romero and F. H. Arnold. Exploring protein fitness landscapes by directed evolution. *Nature Reviews Molecular Cell Biology*, 10(12):866–876, 2009.

[2] K. Chen and F. H. Arnold. Tuning the activity of an enzyme for unusual environments: sequential random mutagenesis of subtilisin E for catalysis in dimethylformamide. *Proceedings of the National Academy of Sciences of the United States of America*, 90(12):5618–5622, 1993.

[3] M. T. Reetz. Combinatorial and evolution-based methods in the creation of enantioselective catalysts. *Angewandte Chemie International Edition*, 40(2):284–310, 2001.

[4] E. T. Boder, K. S. Midelfort, and K. D. Wittrup. Directed evolution of antibody fragments with monovalent femtomolar antigen-binding affinity. *Proceedings of the National Academy of Sciences of the United States of America*, 97(20):10701–10705, 2000.

[5] R. E. Campbell, O. Tour, A. E. Palmer, P. A. Steinbach, G. S. Baird, D. A. Zacharias, and R. Y. Tsien. A monomeric red fluorescent protein. *Proceedings of the National Academy of Sciences of the United States of America*, 99(12):7877–7882, 2002.

[6] L. Jiang, E. A. Althoff, F. R. Clemente, L. Doyle, D. Röthlisberger, A. Zanghellini, J. L. Gallaher, J. L. Betker, F. Tanaka, C. F. Barbas, D. Hilvert, K. N. Houk, B. L. Stoddard, and D. Baker. De novo computational design of retro-aldol enzymes. *Science*, 319(5868):1387–1391, 2008.

[7] D. N. Bolon and S. L. Mayo. Enzyme-like proteins by computational design. *Proceedings of the National Academy of Sciences of the United States of America*, 98(25):14274–14279, 2001.

[8] D. Röthlisberger, O. Khersonsky, A. M. Wollacott, L. Jiang, J. DeChancie, J. Betker, J. L. Gallaher, E. A. Althoff, A. Zanghellini, O. Dym, S. Albeck, K. N. Houk, D. S. Tawfik, and D. Baker. Kemp elimination catalysts by computational enzyme design. *Nature*, 453(7192):190–195, 2008.

[9] N. Tokuriki and D. S. Tawfik. Protein dynamism and evolvability. *Science*, 324(5924):203–207, 2009.

[10] A. Shimotohno, S. Oue, T. Yano, S. Kuramitsu, and H. Kagamiyama. Demonstration of the importance and usefulness of manipulating non-active-site residues in protein design. *Journal of Biochemistry*, 129(6):943–948, 2001.

[11] B. Spiller, A. Gershenson, F. H. Arnold, and R. C. Stevens. A structural view of evolutionary divergence. *Proceedings of the National Academy of Sciences of the United States of America*, 96(22):12305–12310, 1999.

[12] A. Aharoni, L. Gaidukov, O. Khersonsky, S. McQ Gould, C. Roodveldt, and D. S. Tawfik. The 'evolvability' of promiscuous protein functions. *Nature Genetics*, 37(1):73–76, 2005.

[13] I. Sarkar, I. Hauber, J. Hauber, and F. Buchholz. HIV-1 proviral DNA excision using an evolved recombinase. *Science*, 316(5833):1912–1915, 2007.

[14] R. Fasan, M. M. Chen, N. C. Crook, and F. H. Arnold. Engineered alkane-hydroxylating cytochrome P450(BM3) exhibiting nativelike catalytic properties. *Angewandte Chemie International Edition*, 46(44):8414–8418, 2007.

[15] M. T. Reetz, J. D. Carballeira, and A. Vogel. Iterative saturation mutagenesis on the basis of B factors as a strategy for increasing protein thermostability. *Angewandte Chemie International Edition*, 45(46):7745–7751, 2006.

[16] T. H. Yoo, A. J. Link, and D. A. Tirrell. Evolution of a fluorinated green fluorescent protein. *Proceedings of the National Academy of Sciences of the United States of America*, 104(35):13887–13890, 2007.

[17] R. Y. Tsien. Constructing and exploiting the fluorescent protein paintbox (Nobel Lecture). *Angewandte Chemie International Edition*, 48(31):5612–26, 2009.

[18] N. C. Shaner, R. E. Campbell, P. A. Steinbach, B. N. G. Giepmans, A. E. Palmer, and R. Y. Tsien. Improved monomeric red, orange and yellow fluorescent proteins derived from Discosoma sp. red fluorescent protein. *Nature Biotechnology*, 22(12):1567–1572, 2004.

[19] Y. Yokobayashi, R. Weiss, and F. H. Arnold. Directed evolution of a genetic circuit. *Proceedings of the National Academy of Sciences of the United States of America*, 99(26):16587–16591, 2002.

[20] A. A. Beaudry and G. F. Joyce. Directed evolution of an RNA enzyme. *Science*, 257(5070):635–641, 1992.

[21] H. Alper, C. Fischer, E. Nevoigt, and G. Stephanopoulos. Tuning genetic control through promoter engineering. *Proceedings of the National Academy of Sciences of the United States of America*, 102(36):12678–12683, 2005.

[22] F. H. Arnold, P. L. Wintrode, K. Miyazaki, and A. Gershenson. How enzymes adapt: lessons from directed evolution. *Trends in Biochemical Sciences*, 26(2):100–106, 2001.

[23] P. L. Wintrode and F. H. Arnold. Temperature adaptation of enzymes: lessons from laboratory evolution. *Advances in Protein Chemistry*, 55:161–225, 2000.

[24] J. M. Smith. Natural selection and the concept of a protein space. *Nature*, 225(5232):563–564, 1970.

[25] W. Mandecki. The game of chess and searches in protein sequence space. *Trends in Biotechnology*, 16(5):200–202, 1998.

[26] S. Wright. The roles of mutation, inbreeding, crossbreeding and selection in evolution. *Proceedings of the 6th International Congress of Genetics*, 1:356–366, 1932.

[27] D. H. Wolpert and W. G. Macready. No free lunch theorems for optimization. *IEEE Transactions on Evolutionary Computation*, 1(1):67–82, 1997.

[28] S. A. Kauffman and E. D. Weinberger. The NK model of rugged fitness landscapes and its application to maturation of the immune response. *Journal of Theoretical Biology*, 141(2):211–245, 1989.

[29] A. Wagner. Robustness and evolvability: a paradox resolved. *Proceedings of the Royal Society B Biological Sciences*, 275(1630):91–100, 2008.

[30] J. D. Bloom, S. T. Labthavikul, C. R. Otey, and F. H. Arnold. Protein stability promotes evolvability. *Proceedings of the National Academy of Sciences of the United States of America*, 103(15):5869–5874, 2006.

[31] A. D. Keefe and J. W. Szostak. Functional proteins from a random-sequence library. *Nature*, 410(April):715–718, 2001.

[32] D. D. Axe. Estimating the prevalence of protein sequences adopting functional enzyme folds. *Journal of Molecular Biology*, 341(5):1295–1315, 2004.

[33] D. M. Taverna and R. A. Goldstein. Why are proteins marginally stable? *Proteins*, 46(1):105–109, 2002.

[34] S. Govindarajan and R. A. Goldstein. Evolution of model proteins on a foldability landscape. *Proteins*, 29(4):461–466, 1997.

[35] Y. Xia and M. Levitt. Simulating protein evolution in sequence and structure space. *Current Opinion in Structural Biology*, 14(2):202–207, 2004.

[36] J. D. Bloom, J. J. Silberg, C. O. Wilke, D. A. Drummond, C. Adami, and F. H. Arnold. Thermodynamic prediction of protein neutrality. *Proceedings of the National Academy of Sciences of the United States of America*, 102(3):606–611, 2005.

[37] H. H. Guo, J. Choe, and L. A. Loeb. Protein tolerance to random amino acid change. *Proceedings of the National Academy of Sciences of the United States of America*, 101(25):9205–9210, 2004.

[38] F. H. Arnold. Directed evolution: Creating biocatalysts for the future. *Chemical Engineering Science*, 51(23):5091–5102, 1996.

[39] J. L. England and E. I. Shakhnovich. Structural Determinant of Protein Designability. *Physical Review Letters*, 90(21):218101, 2003.

[40] T. L. O'Loughlin, W. M. Patrick, and I. Matsumura. Natural history as a predictor of protein evolvability. *Protein Engineering Design Selection*, 19(10):439–442, 2006.

[41] D. Umeno, A. V. Tobias, and F. H. Arnold. Diversifying carotenoid biosynthetic pathways by directed evolution. *Microbiology and Molecular Biology Reviews*, 69(1):51–78, 2005.

[42] M. E. Glasner, J. A. Gerlt, and P. C. Babbitt. Evolution of enzyme superfamilies. *Current Opinion in Chemical Biology*, 10(5):492–497, 2006.

[43] S. Bershtein, M. Segal, R. Bekerman, N. Tokuriki, and D. S. Tawfik. Robustness-epistasis link shapes the fitness landscape of a randomly drifting protein. *Nature*, 444(7121):929–932, 2006.

[44] J. Claren, C. Malisi, B. Höcker, and R. Sterner. Establishing wild-type levels of catalytic activity on natural and artificial (beta alpha)8-barrel protein scaffolds. *Proceedings of the National Academy of Sciences of the United States of America*, 106(10):3704–3709, 2009.

[45] D. A. Drummond, B. L. Iverson, G. Georgiou, and F. H. Arnold. Why high-error-rate random mutagenesis libraries are enriched in functional and improved proteins. *Journal of Molecular Biology*, 350(4):806–816, 2005.

[46] M. T. Reetz, M. Bocola, J. D. Carballeira, D. Zha, and A. Vogel. Expanding the range of substrate acceptance of enzymes: combinatorial active-site saturation test. *Angewandte Chemie International Edition*, 44(27):4192–4196, 2005.

[47] T. P. Treynor, C. L. Vizcarra, D. Nedelcu, and S. L. Mayo. Computationally designed libraries of fluorescent proteins evaluated by preservation and diversity of function. *Proceedings of the National Academy of Sciences of the United States of America*, 104(1):48–53, 2007.

[48] Y. Yoshikuni, T. E. Ferrin, and J. D. Keasling. Designed divergent evolution of enzyme function. *Nature*, 440(7087):1078–1082, 2006.

[49] L. You and F. H. Arnold. Directed evolution of subtilisin E in Bacillus subtilis to enhance total activity in aqueous dimethylformamide. *Protein Engineering*, 9(1):77–83, 1996.

[50] R. Fujii, M. Kitaoka, and K. Hayashi. RAISE: a simple and novel method of generating random insertion and deletion mutations. *Nucleic Acids Research*, 34(4):e30, 2006.

[51] Z. Qian and S. Lutz. Improving the catalytic activity of Candida antarctica lipase B by circular permutation. *Journal of the American Chemical Society*, 127(39):13466–13467, 2005.

[52] C. Neylon. Chemical and biochemical strategies for the randomization of protein encoding DNA sequences: library construction methods for directed evolution. *Nucleic Acids Research*, 32(4):1448–1459, 2004.

[53] D. Rennell, S. E. Bouvier, L. W. Hardy, and A. R. Poteete. Systematic mutation of bacteriophage T4 lysozyme. *Journal of Molecular Biology*, 222(1):67–88, 1991.

[54] D. D. Axe, N. W. Foster, and A. R. Fersht. A search for single substitutions that eliminate enzymatic function in a bacterial ribonuclease. *Biochemistry*, 37(20):7157–7166, 1998.

[55] S. Shafikhani, R. A. Siegel, E. Ferrari, and V. Schellenberger. Generation of large libraries of random mutants in Bacillus subtilis by PCR-based plasmid multimerization. *Biotechniques*, 23(2):304–310, 1997.

[56] D. A. Drummond, J. J. Silberg, M. M. Meyer, C. O. Wilke, and F. H. Arnold. On the conservative nature of intragenic recombination. *Proceedings of the National Academy of Sciences of the United States of America*, 102(15):5380–5385, 2005.

[57] J. C. Moore, H. M. Jin, O. Kuchner, and F. H. Arnold. Strategies for the in vitro evolution of protein function: enzyme evolution by random recombination of improved sequences. *Journal of Molecular Biology*, 272(3):336–347, 1997.

[58] W. P. C. Stemmer. Rapid evolution of a protein in vitro by DNA shuffling. *Nature*, 370(6488):389–391, 1994.

[59] F. J. Poelwijk, D. J. Kiviet, D. M. Weinreich, and S. J. Tans. Empirical fitness landscapes reveal accessible evolutionary paths. *Nature*, 445(7126):383–386, 2007.

[60] T. Aita, N. Hamamatsu, Y. Nomiya, H. Uchiyama, Y. Shibanaka, and Y. Husimi. Surveying a local fitness landscape of a protein with epistatic sites for the study of directed evolution. *Biopolymers*, 64(2):95–105, 2002.

[61] Y. Hayashi, T. Aita, H. Toyota, Y. Husimi, I. Urabe, and T. Yomo. Experimental Rugged Fitness Landscape in Protein Sequence Space. *PLoS ONE*, 1(1):8, 2006.

[62] J. D. Bloom and F. H. Arnold. In the light of directed evolution: pathways of adaptive protein evolution. *Proceedings of the National Academy of Sciences of the United States of America*, 106(Supplement 1):9995–10000, 2009.

[63] D. M. Weinreich, N. F. Delaney, M. A. Depristo, and D. L. Hartl. Darwinian evolution can follow only very few mutational paths to fitter proteins. *Science*, 312(5770):111–114, 2006.

[64] M. T. Reetz and J. Sanchis. Constructing and analyzing the fitness landscape of an experimental evolutionary process. *Chembiochem*, 9(14):2260–2267, 2008.

[65] K. Bernath, S. Magdassi, and D. S. Tawfik. Directed evolution of protein inhibitors of DNA-nucleases by in vitro compartmentalization (IVC) and nano-droplet delivery. *Journal of Molecular Biology*, 345(5):1015–1026, 2005.

[66] L. Liu, Y. Li, D. Liotta, and S. Lutz. Directed evolution of an orthogonal nucleoside analog kinase via fluorescence-activated cell sorting. *Nucleic Acids Research*, 37(13):4472–4481, 2009.

[67] M. A. Fischbach, J. R. Lai, E. D. Roche, C. T. Walsh, and D. R. Liu. Directed evolution can rapidly improve the activity of chimeric assembly-line enzymes. *Proceedings of the National Academy of Sciences of the United States of America*, 104(29):11951–11956, 2007.

[68] N. Halabi, O. Rivoire, S. Leibler, and R. Ranganathan. Protein sectors: evolutionary units of three-dimensional structure. *Cell*, 138(4):774–786, 2009.

[69] I. Matsumura and A. D. Ellington. In vitro evolution of beta-glucuronidase into a beta-galactosidase proceeds through non-specific intermediates. *Journal of Molecular Biology*, 305(2):331–339, 2001.

[70] S. Park, K. L. Morley, G. P. Horsman, M. Holmquist, K. Hult, and R. J. Kazlauskas. Focusing mutations into the P. fluorescens esterase binding site increases enantioselectivity more effectively than distant mutations. *Chemistry & Biology*, 12(1):45–54, 2005.

[71] J. Paramesvaran, E. G. Hibbert, A. J. Russell, and P. A. Dalby. Distributions of enzyme residues yielding mutants with improved substrate specificities from two different directed evolution strategies. *Protein Engineering Design Selection*, 22(7):401–411, 2009.

[72] R. Fasan, Y. T. Meharenna, C. D. Snow, T. L. Poulos, and F. H. Arnold. Evolutionary history of a specialized p450 propane monooxygenase. *Journal of Molecular Biology*, 383(5):1069–1080, 2008.

[73] C. A. Voigt, C. Martinez, Z.-G. Wang, S. L. Mayo, and F. H. Arnold. Protein building blocks preserved by recombination. *Nature Structural Biology*, 9(7):553–558, 2002.

[74] L. O. Hansson, R. Bolton-Grob, T. Massoud, and B. Mannervik. Evolution of differential substrate specificities in Mu class glutathione transferases probed by DNA shuffling. *Journal of Molecular Biology*, 287(2):265–276, 1999.

[75] A. Crameri, S.-A. Raillard, E. Bermudez, and W. P. C. Stemmer. DNA shuffling of a family of genes from diverse species accelerates directed evolution. *Nature*, 391(6664):288–291, 1998.

[76] M. Ostermeier, J. H. Shim, and S. J. Benkovic. A combinatorial approach to hybrid enzymes independent of DNA homology. *Nature Biotechnology*, 17(12):1205–1209, 1999.

[77] S. Lutz, M. Ostermeier, G. L. Moore, C. D. Maranas, and S. J. Benkovic. Creating multiple-crossover DNA libraries independent of sequence identity. *Proceedings of the National Academy of Sciences of the United States of America*, 98(20):11248–11253, 2001.

[78] K. Hiraga and F. H. Arnold. General method for sequence-independent site-directed chimeragenesis. *Journal of Molecular Biology*, 330(2):287–296, 2003.

[79] P. Heinzelman, C. D. Snow, I. Wu, C. Nguyen, A. Villalobos, S. Govindarajan, J. Minshull, and F. H. Arnold. A family of thermostable fungal cellulases created by structure-guided recombination. *Proceedings of the National Academy of Sciences of the United States of America*, 106(14):5610–5615, 2009.

[80] C. R. Otey, M. Landwehr, J. B. Endelman, K. Hiraga, J. D. Bloom, and F. H. Arnold. Structure-Guided Recombination Creates an Artificial Family of Cytochromes P450. *PLoS Biology*, 4(5):e112, 2006.

[81] R. K. Campbell, E. R. Bergert, Y. H. Wang, J. C. Morris, and W. R. Moyle. Chimeric proteins can exceed the sum of their parts: Implications for evolution and protein design. *Nature Biotechnology*, 15(5), 1997.

[82] J. D. Bloom, Z. Lu, D. Chen, A. Raval, O. S. Venturelli, and F. H. Arnold. Evolution favors protein mutational robustness in sufficiently large populations. *BMC Biology*, 5(1):29, 2007.

[83] G. Amitai, R. D. Gupta, and D. S. Tawfik. Latent evolutionary potentials under the neutral mutational drift of an enzyme. *HFSP Journal*, 1(1):67–78, 2007.

[84] S. Bershtein, K. Goldin, and D. S. Tawfik. Intense neutral drifts yield robust and evolvable consensus proteins. *Journal of Molecular Biology*, 379(5):1029–1044, 2008.

[85] M. Landwehr, M. Carbone, C. R. Otey, Y. Li, and F. H. Arnold. Diversification of catalytic function in a synthetic family of chimeric cytochrome p450s. *Chemistry & Biology*, 14(3):269–278, 2007.

[86] Y. Li, D. A. Drummond, A. M. Sawayama, C. D. Snow, J. D. Bloom, and F. H. Arnold. A diverse family of thermostable cytochrome P450s created by recombination of stabilizing fragments. *Nature Biotechnology*, 25(9):1051–1056, 2007.

[87] R. Couñago, S. Chen, and Y. Shamoo. In vivo molecular evolution reveals biophysical origins of organismal fitness. *Molecular Cell*, 22(4):441–449, 2006.

[88] X. Wang, G. Minasov, and B. K. Shoichet. Evolution of an Antibiotic Resistance Enzyme Constrained by Stability and Activity Trade-offs. *Journal of Molecular Biology*, 320(1):85–95, 2002.

[89] R. D. Gupta and D. S. Tawfik. Directed enzyme evolution via small and effective neutral drift libraries. *Nature Methods*, 5(11):939–942, 2008.

[90] G. N. Somero. Proteins and temperature. *Annual Review of Physiology*, 57(1):43–68, 1995.

[91] P. A. Fields. Protein function at thermal extremes: balancing stability and flexibility. *Comparative Biochemistry and Physiology Part A*, 129(2):417–431, 2001.

[92] L. Giver, A. Gershenson, P.-O. Freskgard, and F. H. Arnold. Directed evolution of a thermostable esterase. *Proceedings of the National Academy of Sciences of the United States of America*, 95(22):12809–12813, 1998.

[93] N. Tokuriki, F. Stricher, L. Serrano, and D. S. Tawfik. How Protein Stability and New Functions Trade Off. *PLoS Computational Biology*, 4(2):7, 2008.

[94] S. G. Peisajovich and D. S. Tawfik. Protein engineers turned evolutionists. *Nature Methods*, 4(12):991–994, 2007.

[95] A. M. Dean and J. W. Thornton. Mechanistic approaches to the study of evolution: the functional synthesis. *Nature Reviews Genetics*, 8(9):675–688, 2007.

[96] S. P. Miller, M. Lunzer, and A. M. Dean. Direct demonstration of an adaptive constraint. *Science*, 314(5798):458–461, 2006.

[97] S. Gavrilets. Evolution and speciation on holey adaptive landscapes. *Trends in Ecology & Evolution*, 12(8):307–312, 1997.

[98] A. Glieder, E. T. Farinas, and F. H. Arnold. Laboratory evolution of a soluble, self-sufficient, highly active alkane hydroxylase. *Nature Biotechnology*, 20(11):1135–1139, 2002.

[99] M. W. Peters, P. Meinhold, A. Glieder, and F. H. Arnold. Regio- and enantioselective alkane hydroxylation with engineered cytochromes P450 BM-3. *Journal of the American Chemical Society*, 125(44):13442–13450, 2003.

[100] M. G. Shapiro, G. G. Westmeyer, P. A. Romero, J. O. Szablowski, B. Küster, A. Shah, C. R. Otey, R. Langer, F. H. Arnold, and A. Jasanoff. Directed evolution of a magnetic resonance imaging contrast agent for noninvasive imaging of dopamine. *Nature Biotechnology*, 28(3):264–270, 2010.

[101] R. B. Buxton. *Introduction to functional magnetic resonance imaging: principles and techniques*. Cambridge University Press, 2002.

[102] S. Ogawa, T. M. Lee, A. R. Kay, and D. W. Tank. Brain magnetic resonance imaging with contrast dependent on blood oxygenation. *Proceedings of the National Academy of Sciences of the United States of America*, 87(24):9868–9872, 1990.

[103] N. K. Logothetis. What we can do and what we cannot do with fMRI. *Nature*, 453(7197):869–878, 2008.

[104] A. Jasanoff. MRI contrast agents for functional molecular imaging of brain activity. *Current Opinion in Neurobiology*, 17(5):593–600, 2007.

[105] Q. S. Li, U. Schwaneberg, P. Fischer, and R. D. Schmid. Directed evolution of the fatty-acid hydroxylase P450 BM-3 into an indole-hydroxylating catalyst. *Chemistry*, 6(9):1531–1536, 2000.

[106] P. Meinhold, M. W. Peters, M. M. Y. Chen, K. Takahashi, and F. H. Arnold. Direct conversion of ethane to ethanol by engineered cytochrome P450 BM3. *Chembiochem*, 6(10):1765–1768, 2005.

[107] B. Knutson and S. E. B. Gibbs. Linking nucleus accumbens dopamine and blood oxygenation. *Psychopharmacology*, 191(3):813–822, 2007.

[108] W. Schultz. Multiple dopamine functions at different time courses. *Annual Review of Neuroscience*, 30(1):259–288, 2007.

[109] S. E. Hyman, R. C. Malenka, and E. J. Nestler. Neural Mechanisms of Addiction: The Role of Reward-Related Learning and Memory. *Annual Review of Neuroscience*, 29(1):565–598, 2006.

[110] P. Damier, E. C. Hirsch, Y. Agid, and A. M. Graybiel. The substantia nigra of the human brain. *Brain*, 122(8):1437–1448, 1999.

[111] A. M. J. Young, M. Joseph, and J. A. Gray. Increased dopamine release in vivo in nucleus accumbens and caudate nucleus of the rat during drinking: A microdialysis study. *Neuroscience*, 48(4):871–876, 1992.

[112] P. A. Garris and R. M. Wightman. Different kinetics govern dopaminergic transmission in the amygdala, prefrontal cortex, and striatum: an in vivo voltammetric study. *Journal of Neuroscience*, 14(1):442–450, 1994.

[113] N. G. Gubernator, H. Zhang, R. G. W. Staal, E. V. Mosharov, D. B. Pereira, M. Yue, V. Balsanek, P. A. Vadola, B. Mukherjee, R. H. Edwards, D. Sulzer, and D. Sames. Fluorescent False Neurotransmitters Visualize Dopamine Release from Individual Presynaptic Terminals. *Science*, 324(June):1441–1444, 2009.

[114] K. P. Lindsey and S. J. Gatley. Applications of clinical dopamine imaging. *Neuroimaging Clinics Of North America*, 16(4):553–573, 2006.

[115] A. G. Ewing, J. C. Bigelow, and R. M. Wightman. Direct in vivo monitoring of dopamine released from two striatal compartments in the rat. *Science*, 221(4606):169–171, 1983.

[116] A. C. Michael, M. Ikeda, and J. B. Justice. Dynamics of the recovery of releasable dopamine following electrical stimulation of the medial forebrain bundle. *Neuroscience Letters*, 76(1):81–86, 1987.

[117] T. Q. Duong, D. S. Kim, K. Uurbil, and S. G. Kim. Spatiotemporal dynamics of the BOLD fMRI signals: toward mapping submillimeter cortical columns using the early negative response. *Magnetic Resonance in Medicine*, 44(2):231–242, 2000.

[118] A. W. Munro, D. G. Leys, K. J. McLean, K. R. Marshall, T. W. B. Ost, S. Daff, C. S. Miles, S. K. Chapman, D. A. Lysek, C. C. Moser, C. C. Page, and P. L. Dutton. P450 BM3: the very model of a modern flavocytochrome. *Trends in Biochemical Sciences*, 27(5):250–257, 2002.

[119] I. D. Macdonald, A. W. Munro, and W. E. Smith. Fatty acid-induced alteration of the porphyrin macrocycle of cytochrome P450 BM3. *Biophysical Journal*, 74(6):3241–3249, 1998.

[120] K. G. Ravichandran, S. S. Boddupalli, C. A. Hasermann, J. A. Peterson, and J. Deisenhofer. Crystal structure of hemoprotein domain of P450BM-3, a prototype for microsomal P450's. *Science*, 261(5122):731–736, 1993.

[121] S. Modi, W. U. Primrose, J. M. Boyle, C. F. Gibson, L. Y. Lian, and G. C. Roberts. NMR studies of substrate binding to cytochrome P450 BM3: comparisons to cytochrome P450 cam. *Biochemistry*, 34(28):8982–8988, 1995.

[122] D. F. V. Lewis. *Guide to Cytochrome P450 Structure and Function.* CRC Press, 2001.

[123] K. Ohnuma, Y. Hayashi, M. Furue, K. Kaneko, and M. Asashima. Serum-free culture conditions for serial subculture of undifferentiated PC12 cells. *Journal of Neuroscience Methods*, 151(2):250–261, 2006.

[124] A. G. Ewing, R. M. Wightman, and M. A. Dayton. In vivo voltammetry with electrodes that discriminate between dopamine and ascorbate. *Brain Research*, 249(2):361–370, 1982.

[125] G. A. Gerhardt, G. M. Rose, and B. J. Hoffer. Release of monoamines from striatum of rat and mouse evoked by local application of potassium: evaluation of a new in vivo electrochemical technique. *Journal of Neurochemistry*, 46(3):842–850, 1986.

[126] Z.-J. Chen, G. T. Gillies, W. C. Broaddus, S. S. Prabhu, H. Fillmore, R. M. Mitchell, F. D. Corwin, and P. P. Fatouros. A realistic brain tissue phantom for intraparenchymal infusion studies. *Journal Of Neurosurgery*, 101(2):314–322, 2004.

[127] N. Vykhodtseva, N. McDannold, and K. Hynynen. Progress and problems in the application of focused ultrasound for blood-brain barrier disruption. *Ultrasonics*, 48(4):279–296, 2008.

[128] H. J. Barnes, M. P. Arlotto, and M. R. Waterman. Expression and enzymatic activity of recombinant cytochrome P450 17 alpha-hydroxylase in Escherichia coli. *Proceedings of the National Academy of Sciences of the United States of America*, 88(13):5597–5601, 1991.

[129] T. Omura and R. Sato. The Carbon Monoxide-binding Pigment of Liver Microsomes. *The Journal of Biological Chemistry*, 239(7):2370–2378, 1964.

[130] R. J. Peroutka, N. Elshourbagy, T. Piech, and T. R. Butt. Enhanced protein expression in mammalian cells using engineered SUMO fusions: secreted phospholipase A2. *Protein Science*, 17(9):1586–1595, 2008.

[131] D. E. Brough, C. Hsu, V. A. Kulesa, G. M. Lee, L. J. Cantolupo, A. Lizonova, and I. Kovesdi. Activation of transgene expression by early region 4 is responsible for a high level of persistent transgene expression from adenovirus vectors in vivo. *Journal of Virology*, 71(12):9206–9213, 1997.

[132] E. E. Nilsson, S. D. Westfall, C. McDonald, T. Lison, I. Sadler-Riggleman, and M. K. Skinner. An in vivo mouse reporter gene (human secreted alkaline phosphatase) model to monitor ovarian tumor growth and response to therapeutics. *Cancer Chemotherapy and Pharmacology*, 49(2):93–100, 2002.

[133] V. E. Shashoua. Extracellular fluid proteins of goldfish brain: studies of concentration and labeling patterns. *Neurochemical Research*, 6(10):1129–1147, 1981.

[134] H. Li and T. L. Poulos. The structure of the cytochrome p450BM-3 haem domain complexed with the fatty acid substrate, palmitoleic acid. *Nature Structural Biology*, 4(2):140–146, 1997.

[135] G. Paxinos and C. Watson. *The rat brain in stereotaxic coordinates*, volume Second. Academic Press, 6th edition, 2007.

[136] V. López, R. Alarcón, M. S. Orellana, P. Enríquez, E. Uribe, J. Martínez, and N. Carvajal. Insights into the interaction of human arginase II with substrate and manganese ions by site-directed mutagenesis and kinetic studies. Alteration of substrate specificity by replacement of Asn149 with Asp. *The FEBS Journal*, 272(17):4540–4548, 2005.

[137] E. Cama, D. M. Colleluori, F. A. Emig, H. Shin, S. W. Kim, N. N. Kim, A. M. Traish, D. E. Ash, and D. W. Christianson. Human arginase II: crystal structure and physiological role in male and female sexual arousal. *Biochemistry*, 42(28):8445–8451, 2003.

[138] D. P. Dowling, L. Di Costanzo, H. A. Gennadios, and D. W. Christianson. Evolution of the arginase fold and functional diversity. *Cellular and molecular life sciences*, 65(13):2039–2055, 2008.

[139] C. M. Ensor, F. W. Holtsberg, J. S. Bomalaski, and M. A. Clark. Pegylated arginine deiminase (ADI-SS PEG20,000 mw) inhibits human melanomas and hepatocellular carcinomas in vitro and in vivo. *Cancer Research*, 62(19):5443–5450, 2002.

[140] L. G. Feun, A. Marini, H. Landy, A. Markoe, D. Heros, C. Robles, C. Herrera, and N. Savaraj. Clinical trial of CPT-11 and VM-26/VP-16 for patients with recurrent malignant brain tumors. *Journal of Neuro-Oncology*, 82(2):177–181, 2007.

[141] C.-Y. Yoon, Y.-J. Shim, E.-H. Kim, J.-H. Lee, N.-H. Won, J.-H. Kim, I.-S. Park, D.-K. Yoon, and B.-H. Min. Renal cell carcinoma does not express argininosuccinate synthetase and

is highly sensitive to arginine deprivation via arginine deiminase. *International Journal of Cancer*, 120(4):897–905, 2007.

[142] M.-Y. Shen and A. Sali. Statistical potential for assessment and prediction of protein structures. *Protein Science*, 15(11):2507–2524, 2006.

[143] L. Scott, J. Lamb, S. Smith, and D. N. Wheatley. Single amino acid (arginine) deprivation: rapid and selective death of cultured transformed and malignant cells. *British Journal of Cancer*, 83(6):800–810, 2000.

[144] P. A. Ascierto, S. Scala, G. Castello, A. Daponte, E. Simeone, A. Ottaiano, G. Beneduce, V. De Rosa, F. Izzo, M. T. Melucci, C. M. Ensor, A. W. Prestayko, F. W. Holtsberg, J. S. Bomalaski, M. A. Clark, N. Savaraj, L. G. Feun, and T. F. Logan. Pegylated arginine deiminase treatment of patients with metastatic melanoma: results from phase I and II studies. *Journal of Clinical Oncology*, 23(30):7660–7668, 2005.

[145] S. Jain-Ghaia, S. C. Sreenath Nagamanic, S. Blasera, K. Siriwardenaa, and A. Feigenbaum. Arginase I deficiency: Severe infantile presentation with hyperammonemia: More common than reported? *Molecular Genetics and Metabolism*, 104(1):107–111, 2011.

[146] Y. Segawa, M. Matsufuji, N. Itokazu, H. Utsunomiya, Y. Watanabe, M. Yoshino, and S. Takashima. A long-term survival case of arginase deficiency with severe multicystic white matter and compound mutations. *Brain & Development*, 33(1):45–48, 2011.

[147] T. Sakiyama, H. Nakabayashi, H. Shimizu, W. Kondo, S. Kodama, and T. Kitagawa. A Successful Trial of Enzyme Replacement Therapy in a Case of Argininemia. *The Tohoku Journal of Experimental Medicine*, 142(3):239–248, 1984.

[148] N. Mizutani, C. Hatakawa, M. Maehara, and K. Watanbe. Enzyme Replacement Therapy in a Patient with Hyperargininemia. *The Tohoku Journal of Experimental Medicine*, 151(3):301–307, 1987.

[149] E. M. Stone, E. S. Glazer, L. Chantranupong, P. Cherukuri, R. M. Breece, D. L. Tierney, S. A. Curley, B. L. Iverson, and G. Georgiou. Replacing Mn(2+) with Co(2+) in human arginase I enhances cytotoxicity toward l-arginine auxotrophic cancer cell lines. *ACS Chemical Biology*, 5(3):333–342, 2010.

[150] E. S. Glazer, E. M. Stone, C. Zhu, K. L. Massey, A. N. Hamir, and S. A. Curley. Bioengineered human arginase I with enhanced activity and stability controls hepatocellular and pancreatic carcinoma xenografts. *Translational Oncology*, 4(3):138–146, 2011.

[151] P. Heinzelman, C. D. Snow, M. A. Smith, X. Yu, A. Kannan, K. Boulware, A. Villalobos, S. Govindarajan, J. Minshull, and F. H. Arnold. SCHEMA recombination of a fungal cellulase uncovers a single mutation that contributes markedly to stability. *The Journal of Biological Chemistry*, 284(39):26229–26233, 2009.

[152] A. Krause and C. Guestrin. Near-optimal Observation Selection using Submodular Functions. *National Conference on Artificial Intelligence, Nectar Track*, 22(2):1650–1654, 2007.

[153] G. L. Nemhauser, L. A. Wolsey, and M. L. Fisher. An analysis of approximations for maximizing submodular set functions. *Mathematical Programming*, 14(1):265–294, 1978.

[154] A. Krause. SFO: A Toolbox for Submodular Function Optimization. *Journal of Machine Learning Research*, 11:1141–1144, 2010.

[155] J. C. Cox, J. Lape, M. A. Sayed, and H. W. Hellinga. Protein fabrication automation. *Protein Science*, 16(3):379–390, 2007.

[156] C. M. Bishop. *Pattern Recognition and Machine Learning.* Springer, New York, 1st edition, 2006.

[157] J. B. Endelman, J. J. Silberg, Z.-G. Wang, and F. H. Arnold. Site-directed protein recombination as a shortest-path problem. *Protein Engineering Design Selection*, 17(7):589–594, 2004.

[158] E. Y. Shishova, L. Di Costanzo, F. A. Emig, D. E. Ash, and D. W. Christianson. Probing the specificity determinants of amino acid recognition by arginase. *Biochemistry*, 48(1):121–131, 2009.

[159] P. Heinzelman, R. Komor, A. Kannan, P. A. Romero, X. Yu, S. Mohler, C. D. Snow, and F. H. Arnold. Efficient screening of fungal cellobiohydrolase class I enzymes for thermostabilizing sequence blocks by SCHEMA structure-guided recombination. *Protein Engineering Design Selection*, 23(11):871–880, 2010.

[160] D. W. Christianson and J. D. Cox. Catalysis by metal-activated hydroxide in zinc and manganese metalloenzymes. *Annual Review of Biochemistry*, 68(1):33–57, 1999.

[161] D. L. Nelson and M. M. Cox. *Lehninger Principles of Biochemistry*, volume 1. W. H. Freeman, 4th edition, 2005.

[162] S. Subramaniam. The biology workbench - A seamless database and analysis environment for the biologist. *Proteins*, 32(1):1–2, 1998.

[163] S. Kumar, C.-J. Tsai, and R. Nussinov. Thermodynamic differences among homologous thermophilic and mesophilic proteins. *Biochemistry*, 40(47):14152–14165, 2001.

[164] R. Nakon, P. R. Rechani, and R. J. Angelici. Copper(II) complex catalysis of amino acid ester hydrolysis. A correlation with complex stability. *Journal of the American Chemical Society*, 96(7):2117–2120, 1974.

[165] E. Alexov. Numerical calculations of the pH of maximal protein stability. *European Journal of Biochemistry*, 271(1):173–185, 2003.

[166] K. L. Shaw, G. R. Grimsley, G. I. Yakovlev, A. A. Makarov, and C. N. Pace. The effect of net charge on the solubility, activity, and stability of ribonuclease Sa. *Protein Science*, 10(6):1206–1215, 2001.

[167] J. P. Schmittschmitt and J. M. Scholtz. The role of protein stability, solubility, and net charge in amyloid fibril formation. *Protein Science*, 12(10):2374–2378, 2003.

[168] M. S. Lawrence, K. J. Phillips, and D. R. Liu. Supercharging proteins can impart unusual resilience. *Journal of the American Chemical Society*, 129(33):10110–10112, 2007.

[169] E. A. Crombez and S. D. Cederbaum. Hyperargininemia due to liver arginase deficiency. *Molecular Genetics and Metabolism*, 84(3):243–251, 2005.

[170] P. C. Rodriguez and A. C. Ochoa. Arginine regulation by myeloid derived suppressor cells and tolerance in cancer: mechanisms and therapeutic perspectives. *Immunological Reviews*, 222(1):180–191, 2008.

[171] P. N.-M. Cheng, T.-L. Lam, W.-M. Lam, S.-M. Tsui, A. W.-M. Cheng, W.-H. Lo, and Y.-C. Leung. Pegylated recombinant human arginase (rhArg-peg5,000mw) inhibits the in vitro and in vivo proliferation of human hepatocellular carcinoma through arginine depletion. *Cancer Research*, 67(1):309–317, 2007.

[172] N. H. Barton and B. Charlesworth. Why sex and recombination? *Science*, 281(5385):1986–1990, 1998.

[173] S. P. Otto and T. Lenormand. Resolving the paradox of sex and recombination. *Nature Reviews Genetics*, 3(4):252–261, 2002.

[174] W. B. Watt. Intragenic recombination as a source of population genetic variability. *The American Naturalist*, 106(952):737–753, 1972.

[175] C. Strobeck and K. Morgan. The Effect of Intragenic Recombination on the Number of Alleles in a Finite Population. *Genetics*, 88(4):829–844, 1978.

[176] M. Freeling. Allelic Variation at the Level of Intragenic Recombination. *Genetics*, 89(1):211–224, 1978.

[177] E. De Silva, L. A. Kelley, and M. P. H. Stumpf. The extent and importance of intragenic recombination. *Human Genomics*, 1(6):410–420, 2004.

[178] M. Carbone and F. H. Arnold. Engineering by homologous recombination: exploring sequence and function within a conserved fold. *Current Opinion in Structural Biology*, 17(4):454–459, 2007.

[179] D. M. LeMaster and G. Hernández. Additivity in both thermodynamic stability and thermal transition temperature for rubredoxin chimeras via hybrid native partitioning. *Structure*, 13(8):1153–1163, 2005.

[180] H. V. Thulasiram, H. K. Erickson, and C. D. Poulter. Chimeras of two isoprenoid synthases catalyze all four coupling reactions in isoprenoid biosynthesis. *Science*, 316(5821):73–76, 2007.

[181] S. G. Peisajovich, J. E. Garbarino, P. Wei, and W. A. Lim. Rapid diversification of cell signaling phenotypes by modular domain recombination. *Science*, 328(5976):368–372, 2010.

[182] M. F. Farrow and F. H. Arnold. Combinatorial recombination of gene fragments to construct a library of chimeras. *Current Protocols in Protein Science*, 26:2.1–2.20, 2010.

[183] R. J. Adler. *The Geometry of Random Fields.* Wiley & Sons, Chichester, 1st edition, 1981.

[184] S. Z. Li. *Markov Random Field Modeling in Image Analysis.* Springer, London, 3rd edition, 2009.

[185] M. L. Stein. *Interpolation of Spatial Data: Some Theory for Kriging.* Springer, New York, 1st edition, 1999.

[186] P. F. Stadler and R. Happel. Random Field Models For Fitness Landscapes. *Journal of Mathematical Biology*, 38(5):435–478, 1999.

[187] S. A. Kauffman and S. A. Levin. Towards a general theory of adaptive walks on rugged landscapes. *Journal of Theoretical Biology*, 128(1):11–45, 1987.

[188] S. Miyazawa and R. L. Jernigan. Estimation of effective interresidue contact energies from protein crystal structures: quasi-chemical approximation. *Macromolecules*, 18(3):534–552, 1985.

[189] C. Clementi, M. Vendruscolo, A. Maritan, and E. Domany. Folding Lennard-Jones proteins by a contact potential. *Proteins*, 37(4):544–553, 1999.

[190] M. Vendruscolo, R. Najmanovich, and E. Domany. Can a pairwise contact potential stabilize native protein folds against decoys obtained by threading? *Proteins*, 38(2):134–148, 2000.

[191] M. M. Meyer, J. J. Silberg, C. A. Voigt, J. B. Endelman, S. L. Mayo, Z.-G. Wang, and F. H. Arnold. Library analysis of SCHEMA-guided protein recombination. *Protein Science*, 12(8):1686–1693, 2003.

[192] F. Mingardon, D. L. Trudeau, M. A. Smith, C. D. Snow, and F. H. Arnold. Structure-guided SCHEMA recombination of GH9-CBM3c bacterial cellulases. *In preparation*, 2012.

[193] J. B. Endelman. *Design and analysis of combinatorial protein libraries created by site-directed recombination.* Dissertation, California Institute of Technology, 2005.

[194] R. A. Fisher. *The Genetical Theory of Natural Selection.* Clarendon, Oxford, 1930.

[195] B. Charlesworth and D. Charlesworth. *Elements of evolutionary genetics.* Roberts and Co. Publishers, 2010.

[196] M. Carneiro and D. L. Hartl. Adaptive landscapes and protein evolution. *Proceedings of the National Academy of Sciences of the United States of America*, 107(suppl. 1):1747–1751, 2010.

[197] E. Bornberg-Bauer. How are model protein structures distributed in sequence space? *Biophysical Journal*, 73(5):2393–2403, 1997.

[198] M. M. Meyer, L. Hochrein, and F. H. Arnold. Structure-guided SCHEMA recombination of distantly related beta-lactamases. *Protein Engineering Design Selection*, 19(12):563–570, 2006.

[199] M. F. Farrow, T. M. Lee, C. D. Snow, P. A. Romero, and F. H. Arnold. Recombination of distantly related Cel5a enzymes creates novel active sites. *Submitted*, 2012.

[200] M. A. Smith, A. Rentmeister, C. D. Snow, T. Wu, M. F. Farrow, F. Mingardon, and F. H. Arnold. Synthetic diversity created for family 48 bacterial cellulases. *In preparation*, 2012.

[201] P. A. Romero, E. Stone, C. Lamb, L. Chantranupong, A. Krause, A. Miklos, R. A. Hughes, B. Fechtel, A. D. Ellington, F. H. Arnold, and G. Georgiou. SCHEMA-Designed Variants of Human Arginase I and II Reveal Sequence Elements Important to Stability and Catalysis. *ACS Synthetic Biology*, 2012.

[202] C. J. Clopper and E. S. Pearson. The use of confidence or fiducial limits illustrated in the case of the binomial. *Biometrika*, 26(4):404–413, 1934.

[203] L. Tierney and J. B. Kadane. Accurate approximations for posterior moments and marginal densities. *Journal of the American Statistical Association*, 81(393):82–86, 1986.

[204] S. Gavrilets. *Fitness Landscapes and the Origin of Species*, volume series of *Monographs in Population Biology*. Princeton University Press, Princeton, NJ, 2004.

[205] S. P. Yip, J. U. Lovegrove, N. A. Rana, D. A. Hopkinson, and D. B. Whitehouse. Mapping recombination hotspots in human phosphoglucomutase (PGM1). *Human Molecular Genetics*, 8(9):1699–1706, 1999.

[206] A. D. McBee, D. J. Wegner, C. S. Carlson, J. A. Wambach, P. Yang, H. B. Heins, O. D. Saugstad, M. A. Trusgnich, J. Watkins-Torry, L. M. Nogee, H. Henderson, F. S. Cole, and A. Hamvas. Recombination as a mechanism for sporadic mutation in the surfactant protein-C gene. *Pediatric Pulmonology*, 43(5):443–450, 2008.

[207] R. A. Watson, D. M. Weinreich, and J. Wakeley. Genome structure and the benefit of sex. *Evolution*, 65(2):523–536, February 2011.

[208] M. Dash and H. Liu. Feature selection for classification. *Intelligent Data Analysis*, 1:131–156, 1997.

[209] J. D. M. Rennie. Regularized Logistic Regression is Strictly Convex. *Unpublished manuscript*, pages 1–4, 2005.

[210] A. G. Murzin, S. E. Brenner, T. Hubbard, and C. Chothia. SCOP: a structural classification of proteins database for the investigation of sequences and structures. *Journal of Molecular Biology*, 247(4):536–40, 1995.

[211] J. H. Holland. *Adaptation in Natural and Artificial Systems.* University of Michigan Press, Ann Arbor, MI, 1st edition, 1975.

[212] N. A. Pierce and E. Winfree. Protein design is NP-hard. *Protein Engineering*, 15(10):779–782, 2002.

[213] H. A. Orr. The distribution of fitness effects among beneficial mutations in Fisher's geometric model of adaptation. *Journal of Theoretical Biology*, 238(2):279–285, 2006.

[214] B. I. Dahiyat and S. L. Mayo. De novo protein design: fully automated sequence selection. *Science*, 278(5335):82–87, 1997.

[215] J. K. Lassila, D. Baker, and D. Herschlag. Origins of catalysis by computationally designed retroaldolase enzymes. *Proceedings of the National Academy of Sciences of the United States of America*, 107(11):4937–4942, 2010.

[216] D. Baker. An exciting but challenging road ahead for computational enzyme design. *Protein Science*, 19(10):1817–1819, 2010.

[217] C. K. I. Williams and C. E. Rasmussen. Gaussian processes for regression. In D. S. Touretzky, M. C. Mozer, and M. E. Hasselmo, editors, *Advances in Neural Information Processing Systems*, volume 8, pages 514–520. MIT Press, 1996.

[218] C. E. Rasmussen and C. Williams. *Gaussian Processes for Machine Learning*, volume 14 of *Adaptive Computation and Machine Learning.* MIT Press, Cambridge, MA, 2006.

[219] C. K. I. Williams and D. Barber. Bayesian classification with Gaussian processes. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(12):1342–1351, 1997.

[220] C. R. Otey, J. J. Silberg, C. A. Voigt, J. B. Endelman, G. Bandara, and F. H. Arnold. Functional evolution and structural conservation in chimeric cytochromes p450: calibrating a structure-guided approach. *Chemistry & Biology*, 11(3):309–318, 2004.

[221] C. Guestrin, A. Krause, and A. P. Singh. Near-optimal sensor placements in Gaussian processes. *Proceedings of the 22nd International Conference on Machine Learning*, 1(June):265–272, 2005.

[222] A. Krause, H. B. Mcmahan, C. Guestrin, and A. Gupta. Robust Submodular Observation Selection. *Journal of Machine Learning Research*, 9(January):2761–2801, 2008.

[223] R. S. Sutton and A. G. Barto. *Reinforcement Learning.* MIT Press, Cambridge, MA, 1998.

[224] D. A. Berry and B. Fristedt. *Bandit Problems: Sequential Allocation of Experiments.* Monographs on statistics and applied probability. Chapman and Hall, 1985.

[225] D. Chakrabarti, F. Radlinski, R. Kumar, and E. Upfal. Mortal Multi-Armed Bandits. *Advances in Neural Information Processing Systems*, pages 273–280, 2008.

[226] W. H. Press. Bandit solutions provide unified ethical models for randomized clinical trials and comparative effectiveness research. *Proceedings of the National Academy of Sciences of the United States of America*, 106(52):22387–22392, 2009.

[227] V. Krishnamurthy, B. Wahlberg, and F. Lingelbach. A Value Iteration Algorithm for Partially Observed Markov Decision Process Multi-armed Bandits. *IEEE Transactions on Automatic Control*, pages 1–10, 2001.

[228] P. Auer. Using Confidence Bounds for Exploitation-Exploration Trade-offs. *Journal of Machine Learning Research*, 3(3):397–422, 2003.

[229] N. Srinivas, A. Krause, S. M. Kakade, and M. Seeger. Gaussian Process Optimization in the Bandit Setting: No Regret and Experimental Design. *Proceedings of the 27th International Conference on Machine learning*, pages 1015–1022, 2010.

[230] A. Krause. Batch mode bandit optimization. *In preparation*, 2012.

[231] A. K. Kelmans and B. N. Kimelfeld. Multiplicative submodularity of a matrix's principal minor as a function of the set of its rows and some combinatorial applications. *Discrete Mathematics*, 44(1):113–116, 1983.

[232] M. Minoux. Accelerated greedy algorithms for maximizing submodular set functions. *Optimization Techniques*, 7:234–243, 1978.

[233] C. R. Otey and J. M. Joern. High-throughput screen for aromatic hydroxylation. *Methods in Molecular Biology*, 230:141–148, 2003.

[234] C. R. Otey. High-throughput carbon monoxide binding assay for cytochromes p450. *Methods in Molecular Biology*, 230:137–139, 2003.

[235] D. Peña and I. Guttman. Comparing probabilistic methods for outlier detection in linear models. *Biometrika*, 80(3):603–610, 1993.

[236] W. Rocchia, S. Sridharan, A. Nicholls, E. Alexov, A. Chiabrera, and B. Honig. Rapid grid-based construction of the molecular surface for both molecules and geometric objects: applications to the finite difference Poisson-Boltzmann method. *Journal of Computational Chemistry*, 23:128–137, 2002.

[237] E. M. Brustad, V. S. Lelyveld, C. D. Snow, N. C. Crook, F. Bi, F. M. Martinez, T. J. Scholl, A. Jasanoff, and F. H. Arnold. Directed Evolution and Structural Characterization of Highly Selective P450-based Functional MRI Reporters for Non-invasive Dopamine and Serotonin Imaging. *In preparation*, 2012.

[238] I. V. Loksha, J. R. Maiolo, C. W. Hong, A. Ng, and C. D. Snow. SHARPEN-systematic hierarchical algorithms for rotamers and proteins on an extended network. *Journal of Computational Chemistry*, 30(6):999–1005, 2009.

[239] M. Valiev, E. J. Bylaska, N. Govind, K. Kowalski, T. P. Straatsma, H. J. J. Van Dam, D. Wang, J. Nieplocha, E. Apra, and T. L. Windus. NWChem: a comprehensive and scalable

open-source solution for large scale molecular simulations. *Computer Physics Communications*, 181(9):1477–1489, 2010.

[240] C. Zhang, S. Liu, H. Zhou, and Y. Zhou. An accurate, residue-level, pair potential of mean force for folding and binding based on the distance-scaled, ideal-gas reference state. *Protein Science*, 13(2):400–411, 2004.

[241] L. Kocsis and C. Szepesvári. Bandit based Monte-Carlo Planning. *Machine Learning ECML 2006*, 4212:282–293, 2006.

[242] E. Bonilla, K. M. Chai, and C. Williams. Multi-task Gaussian Process Prediction. *Advances in Neural Information Processing Systems*, 20(October):153–160, 2008.