OVERVIEW

The goal of this proposal is to determine how increasing the dimensionality of a genotype-phenotype map alters evolutionary trajectories. To achieve this goal, members of the Harms lab will experimentally characterize a genotype-phenotype map containing all possible combinations of 15 historical substitutions that occurred between an ancestral green GFP-like protein and a derived red protein. This map will contain 49,152 genotypes and 1.7×10^{14} potential forward trajectories—large enough for neutral networks and indirect paths to arise and profoundly alter possible evolutionary trajectories through the map. The Harms lab will also develop advanced software for studies of trajectories through this high-dimensional genotype-phenotype map. They will borrow approaches from chemical physics, including Transition Path Theory (which allows rapid calculation of evolutionary trajectories through vast genotype-phenotype maps) and Perron-Cluster Cluster Analysis (which identifies fitness peaks and allows coarse-graining of the space). Finally, using these tools, the Harms lab will investigate the interplay between dimensionality and population genetics in determining evolutionary trajectories through the map. Using both the GFP-like protein map from above and published genotype-phenotype maps, the Harms lab will study how the effects of neutral networks and high-order epistasis scale was the size of the map increases. These computational studies will be done in a variety of population genetics regimes, incorporating phenomena such as neutral drift and multi-step mutations.

INTELLECTUAL MERIT

Understanding what determines evolutionary trajectories is a fundamental question in protein evolution. Studies of small genotype-phenotype maps have shown that, as a result of epistasis, protein evolution is often constrained to a few, high probability evolutionary trajectories. While those studies were done using maps with 8 or fewer amino acid substitutions, natural protein evolution often involves 15 or more substitutions. As the dimensionality of genotype-phenotype maps increases, phenomena like neutral networks and high-order epistasis are expected to emerge, profoundly altering the accessibility and probabilities of evolutionary trajectories through the map. As a result, observations in small genotype-phenotype maps may not translate to these larger maps. The proposed work addresses this gap by investigating the effect of increasing dimensionality on evolutionary trajectories for an evolving protein.

This work will provide: 1) A publicly available, 15-dimensional genotype-phenotype map covering the evolution of a new protein function; 2) Powerful software for studies of evolutionary trajectories; and 3) Deep insight into the determinants of protein evolutionary trajectories in high-dimensional genotype-phenotype maps.

BROADER IMPACTS

Many issues facing society are fundamentally problems in protein evolution, from the evolution of antibiotic/pesticide resistance to attempts to engineer designer proteins using directed evolutionary methods. The proposed work will provide new insight into how universal features of proteins determine evolutionary outcomes. This will inform everything from strategies to combat evolution of drug resistance to attempts to engineer more efficient industrial enzymes.

In addition to expanding our understanding of protein evolution, this project leverages ongoing research in the Harms lab to improve high school evolutionary biology education. The Harms lab will host high school science teachers from local rural schools as summer researchers. Members of the Harms lab will then collaborate with these teachers and other on-campus groups to develop a new, inquiry based high school evolution curriculum that focuses on “tree thinking” and mechanisms of protein evolution. As part of developing this curriculum, the PI and graduate students in the lab will join these teacher-researchers in their high school classrooms. Finally, the developed materials will be disseminated internationally through organizations such as the Open Education Resource Commons.

TABLE OF CONTENTS

For font size and page formatting specifications, see PAPPG section II.B.2.

*Proposers may select any numbering mechanism for the proposal. The entire proposal however, must be paginated. Complete both columns only if the proposal is numbered consecutively.

1. BACKGROUND AND JUSTIFICATION

Understanding what determines evolutionary trajectories is a fundamental question in molecular evolution. Why is one trajectory taken relative to others? To what extent do evolutionary trajectories depend on the details of the genotype-phenotype map versus global features like map dimensionality? What are the roles of epistasis, contingency, and constraint in evolution?

These questions are particularly interesting and pressing for studies of protein evolution. Protein evolution is a key driver for the amazing structural and functional diversity in living things.[1–6] Further, some of the greatest problems facing humanity are fundamentally questions in protein evolution: protein-coding mutations underlie the evolution of many drug resistant pathogens[7–12] and pesticide/herbicide resistant pests.[13–16] The choice of management strategy depends strongly on understanding the evolutionary dynamics that underlie these evolutionary changes.[17–21] In other circumstances, we wish to promote protein evolution: state-of-the-art protein engineering often consists of a combination of rational design and optimization by directed evolution.[22–26] Understanding protein evolution at a fundamental level is key to improving these approaches.

One powerful way to study protein evolution is with a combinatorial genotype-phenotype map.[8,27–31] Such a map contains all combinations of mutations between two protein sequences. This approach can be illustrated with John Maynard Smith's word game,[5,8,32,33] which imagines protein sequences as English words and mutations as changes in letters (Fig 1). Some words are meaningful (functional) while others are gibberish (non-functional). The ability to traverse this map is determined by the connectivity of meaningful words. For example, Fig 1 shows the evolution of the word "FAST" into the word "LOVE." There are 16 (2^4) words in this map, but only eight viable English words. Further, because of the distribution of those words in the map, only 2 of the 24 (4!) possible forward trajectories are accessible.

Fig 1: A combinatorial genotype-phenotype map reveals accessible evolutionary trajectories5. Meaningless (inviable) words are faded out.

Studies of combinatorial protein genotype-phenotype maps have revealed that protein evolution is strongly constrained by the details of the map.[8,33–37] This arises because mutations exhibit strong epistasis,[3,8,27–29,31,34,36–48] making the effects of mutations contingent upon previous substitutions.[3,49–51] As a result, in many instances there are relatively few, high-probability paths through each map.[3,8,31,38,39]

One limitation of previous work is the small size of the maps. The largest published map has 8 sites (2^8 = 256 genotypes);[40] however, many historical evolutionary transitions are much larger. We took a random sample of published ancestral sequence reconstruction studies and looked at the number of substitutions that occurred on the branches of interest.[6,27,52–59] These branches had 35 ± 17 substitutions—corresponding to an average of 6×10^{10} genotypes.

Evolutionary transitions through small genotype-phenotype maps are almost certainly different than transitions through large maps. There are strong theoretical arguments that maps with many mutations ("high-dimensional" maps) will exhibit qualitatively different evolutionary properties than maps with few mutations. As a result, conclusions about epistasis and the nature of protein evolutionary trajectories should be revisited in a high-dimensional context.

One key feature of high-dimensional maps is the emergence of neutral networks.[60–66] Fig 2A shows how the number of genotypes and number of forward trajectories scale as the number of sites in a combinatorial map increases. The number of forward trajectories grows by $L!$—much more rapidly than the number of genotypes (2^L). These scaling relationships mean that, as the size of a map increases, the probability of viable trajectories between the ancestral and derived

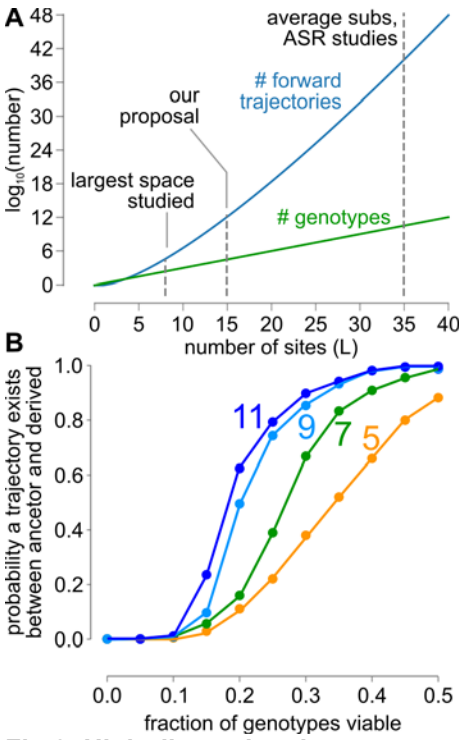

Fig 2: High-dimensional genotype-phenotype maps are qualitatively different than low-dimensional maps. A) Scaling relationships for the number of genotypes (green) and forward trajectories (blue) for binary genotype-phenotype maps with increasing numbers of sites. B) Probability at least one trajectory exists between an ancestral and derived genotype for maps with a random fraction of genotypes inviable.

states increases. This can be seen in Fig 2B, which shows the probability that there is a viable trajectory across a map ranging from 5 to 11 sites where each genotype is randomly set to either viable or inviable. If 25% of genotypes are viable in a 5-site map, there is only a 20% chance that a path exists between the ancestral and derived state. In an 11-site map, however, there is an 80% chance that a path exists.

A second important feature of high-dimensional genotype-phenotype maps is the presence of high-order epistasis.[31,37,43,48,67–69] High-order epistasis is an interaction between three or more mutations that cannot be explained by individual and pairwise mutational effects. High-order epistasis is a ubiquitous feature of genotype-phenotype maps[48] that can play a profound role in evolutionary trajectories.[37,43] We recently found that we could detect epistasis of the L^{th} order in an L-site map (meaning, for example, that we find six-way epistasis in a six-site map).[48] There is no indication that this epistasis decays to zero with increasing order.[43,48] Indeed, we have argued that the statistical thermodynamics of proteins may lead to epistasis at ultra high-orders.[70] Such high-order epistasis, if general, would mean that the effect of a mutation would change as substitutions accumulated. This would create profound unpredictability in protein evolution, as even knowing the individual and pairwise epistatic effects of all mutations would be insufficient to predict evolutionary trajectories.[37,47,71]

Because of the possible emergence of neutral networks and high-order epistasis, we expect that many historical transitions in protein function will be qualitatively different than the small transitions that have been studied previously. New experiments, on much larger genotype-phenotype maps, are required to address this gap.

We propose experimentally and computationally dissecting the evolution of new protein function through a 15-dimensional map. This map will be built from all combinations of 15 historical substitutions that converted an ancestral green GFP-like protein into a red protein.[58,72] This map is in a profoundly different evolutionary regime than all previously reported combinatorial maps. One way to see this is by considering the ratio of the number of possible forward trajectories versus the number of genotypes. For the largest previous map,[40] this ratio is 158; for our map, this ratio is 3×10^9. As a result, we expect a vast number of viable low flux trajectories, rather than a few high flux trajectories as observed in small maps.

In the following three aims, we will develop our experimental, computational, and conceptual approach. In the first aim, we describe how we will experimentally measure this 15-dimensional map. In the second aim, we describe computational tools that we will develop to allow studies of such large maps. This will represent a significant portion of our effort given the vast number of trajectories in the map. Finally, in the third aim, we describe the investigations we will perform on our genotype-phenotype map and other published maps to understand the interplay between increasing dimensionality, high-order epistasis, and population genetics in determining evolutionary trajectories.

Research Products:

- A 15-dimensional combinatorial genotype-phenotype map of the historical substitutions that converted an ancestral GFP-like protein from green to red.
- Novel, powerful, open-source software for analyses of large genotype-phenotype maps, drawing from existing conceptual frameworks in computational physical chemistry.
- Deep insight into the interplay between genotype-phenotype map dimensionality, high-order epistasis, and population genetics in determining evolutionary trajectories.

2. PLANNED RESEARCH ACTIVITIES

2.1: Experimentally measure phenotypes in a 15-dimensional "transition space" between ancestral green and derived red GFP-like proteins

Our goal is to construct a large, combinatorial genotype-phenotype map encompassing an evolutionary transition. To achieve this aim, we will combine high-throughput protein characterization[73–78] and Ancestral Sequence Reconstruction (ASR).[52–54,58,72,79–91] High-throughput protein characterization is often used for deep mutational scanning, which provides a high-resolution picture of the local genotype-phenotype map surrounding a genotype of interest (Fig 1A). In contrast, ASR reveals the end-states of an evolutionary transition. Manipulative experiments can reveal a sparse sample of the genotypes in the evolutionary transition, but the number of genotypes characterized in the transitions themselves is usually limited (Fig 3B).

We propose combining these approaches, taking two ancestral proteins known to have different properties, and then characterizing all possible combinations of mutations between those end states (Fig 3C, 3D). We will use technology developed for deep mutational scanning and apply it in an explicitly evolutionary context, revealing the complete combinatorial map between an ancestral and modern protein function (Fig 3D).

As a model system, we will characterize the fluorescent phenotype of 49,152 possible intermediates between an ancestral green GFP-like protein and a red variant containing 15 historical substitutions. The Matz group previously used ancestral sequence reconstruction to trace the evolution of red GFP-like proteins from *Favia* corals.[58,72] They found the ancestor of the clade was green (Fig 4). They then rank-ordered the effects of the mutations when introduced into the ancestral genotype. We will use the top 15 substitutions from this list to construct the map. This balances creating a large map against technical feasibility. We have made the "ancGFP+15" construct (see Preliminary Data below) and found that it is, like the derived protein, red.

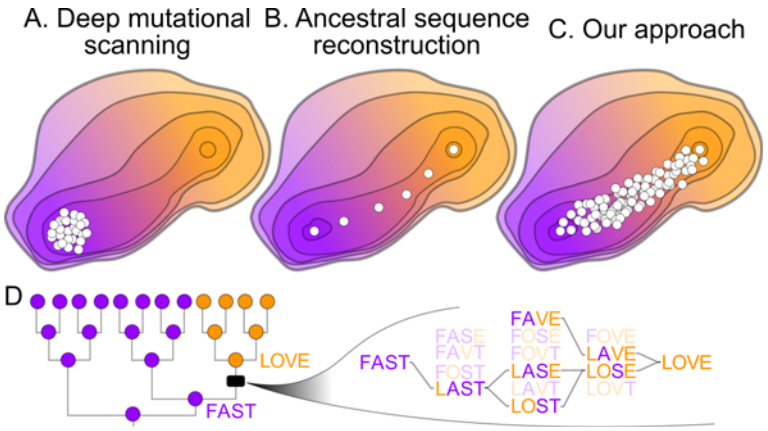

Fig 3: Combining ASR with high-throughput characterization allows construction of a large genotype-phenotype map spanning an evolutionary transition. Top panels show a schematic of a genotype-phenotype map that ranges continuously from from one phenotype (purple) to another (orange). The three sub-panels show different strategies to sample the map: deep mutational scanning (dense, local), ASR (sparse, long-distance), and our method (dense, long-distance). Our method is shown on the bottom row: unpack an ancient evolutionary transition into a combinatorial genotype-phenotype map by screening combinations of historical substitutions.

GFP-like proteins are a powerful model to study the general features of protein evolution. This transition exhibits characteristics shared by a large

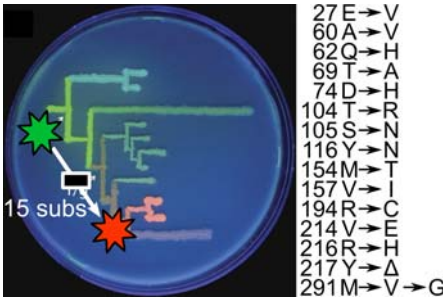

Fig 4: GFP-like proteins evolved from a green ancestor. Plate figure is adapted from Uglade et al.[72] Fifteen substitutions (listed at right) are sufficient to convert the green ancestor into a red GFP-like protein. The M291G substitution must pass through a V codon if only single nucleotide steps are allowed.

number of protein genotype-phenotype maps. 1) The historical substitutions exhibit extensive epistasis.[58,72,92] 2) The effect size distribution matches that noted in many other studies: a few large-effect mutations that are then attenuated by a large collection of epistatically interacting modifier mutations.[58] 3) Finally, like many maps, there is one amino acid substitution (M291G) that requires more than one base substitution at the codon level. This requires either an evolutionary intermediate Valine or a multi-step move that could represent a high-dimensional bypass.[30] Studies in this map will therefore allow us to investigate how the effects of these common features of protein genotype-phenotype maps scale in higher dimensions.

Another key feature of the model system is technical ease. GFP-like proteins are readily expressed in bacterial cells and can be characterized without purification. There are well-established protocols for sorting GFP-like proteins based on phenotype. Indeed, the protocol we describe is derived from a recently published, massive characterization of the local fitness landscape surrounding a modern GFP.[78]

2.1.1: Library construction

The transition consists of 1 two-step mutation and 14 single-step mutations (Fig 4). This gives a total of 3×2^{14} = 49,152 genotypes in the transition. We will order a combinatorially complete DNA library containing all of these mutations in the ancGFP background (Genewiz). We will order a trimer-controlled synthetic library, which allows precise control of the frequency of each codon at each position in the protein. At each transitional site, we will order codons encoding ancestral and derived amino acids at even frequencies: 33.3% each at the triple site and 50% each for the binary sites. We will use *E. coli* optimized codons at all sites. We will also include a completely random 15-bp fragment just 3' of the gene stop codon.

We will clone this library into the pQE30 vector,[78] which will express members of the ancGFP library linked to a blue fluorescent protein (eBFP2) (Fig 5A). The eBFP2 allows us control for protein expression in our FACS experiments by gating on blue fluorescence, which should be identical across samples. In this, we are following the approach taken by a recent deep mutation scan of GFP.[78]

In our FACS experiments, we will measure the frequency of each clone in each intensity bin by sequencing a 15 bp barcode unique to each library member (Fig 5B-C). This requires an up-front experiment to map barcodes to genotypes. We will do this by including a 15 bp random sequence just 3' of the gene stop codon during library synthesis. After cloning, we will take ~1,000,000 *E. coli* transformants as our production library. In 1,000,000 clones, we expect to see all of of our 49,152 sequences at least five times, but expect to see each barcode only once. We will then use a PacBio Sequel (UO HTS Facility) to sequence through the barcode and entire 700 bp gene of each clone. With an expected 20 million reads, we expect to see each genotype/barcode pair at least 5 times, allowing unambiguous assignment of each barcode to a genotype.

2.1.2: Phenotype measurement

We will measure the red and green fluorescence intensity of each genotype using FACS coupled to next-generation sequencing.[78] We will transform our library into *E. coli,* grow them to log phase, and then induce for a fixed period of time using IPTG. To mature the expressed fluorophores, we will expose these bacteria to broad-spectrum UV-B at 4 °C.[58,72] We will take this population of bacteria and then use FACS to sort independently on green and red

fluorescence intensity (Fig 5B). We will first employ a gate on blue fluorescence intensity (EX/EM: 405 nm/450 nm), taking only cells expressing eBFP2 and thus controlling for protein expression. We will follow this by sorting into bins of green and red intensity. We will use EX/EM of 428 nm/525 nm and 488 nm/600 nm for this purpose. We will use established best practices for selecting bin width and the size of the bacterial population sorted.[78,93]

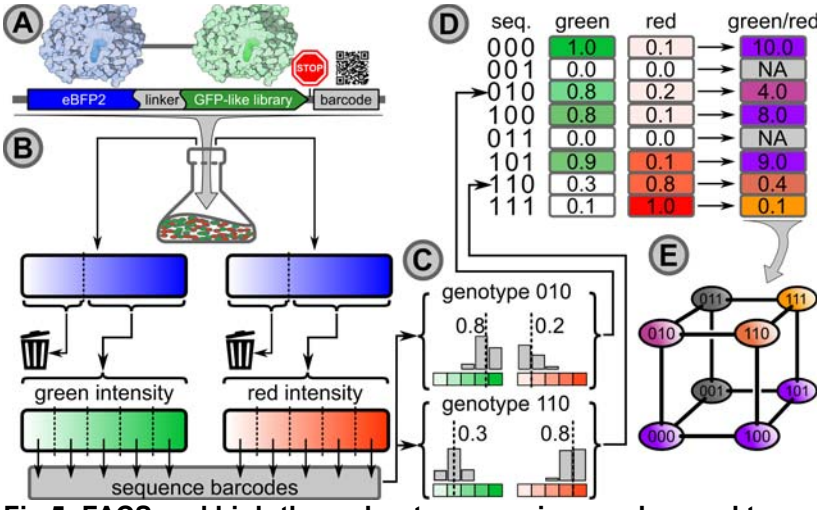

Fig 5: FACS and high-throughput sequencing can be used to construct a genotype-phenotype map. A) We will clone the GFP-like library as a fusion with a blue fluorescent protein. Each library member will have a unique barcode. B) After expressing the library in bacteria, we will independently sort the bacteria. We will first sort on blue fluorescence as an expression control. We will then independently sort the library on green and red intensity. C) By sequencing each intensity pool, we can quantify the distribution of each clone across bins and quantify its green and red fluorescence. D) We can then take the ratio of red and green fluorescence to E) construct a complete genotype-phenotype map.

After sorting, we will measure the frequency of each genotype in each intensity bin. We will also sequence the pool that failed our expression gate to identify genotypes that do not express well. We will do a standard plasmid prep on each pool, followed by PCR amplification of the 15 bp barcode from each plasmid and attachment of standard Illumina flow cell adapters. We will then sequence the contents of each bin using Illumina HiSeq 4000. We will use standard software pipelines to estimate the frequency of each genotype. Because our library complexity is small (49,152 genotypes), we expect a total sequencing depth of ~10,000x per lane, and thus extremely high precision estimates of frequency. We will do the experiment in biological triplicate.

Once we have the frequency of each genotype in each intensity bin, we can estimate the underlying green and red fluorescence intensity of that genotype. We will do so by taking the weighted average of each populated bin (Fig 5C). Finally, we will take the ratio of red/green intensity, thus quantifying how far along the evolutionary transition the genotype in question is (Fig 5D). Once this is done for all genotypes, we will use this to construct a complete, 15-dimensional genotype-phenotype map for the transition (Fig 5E).

2.1.3: Validation

We will validate our FACS results by randomly sampling genotypes for direct spectroscopic measurement. This measurement can be done in 96-well plate format, allowing us to easily characterize on the order of ~100 clones. By comparing this output to the output of the FACS runs, we will be able to directly measure the relationship between our FACS results and the spectroscopic phenotype .

2.1.4: Preliminary Data

We have established that we can work with these proteins in the lab. We have cloned, expressed and purified three GFP-like proteins: ancGFP (the deepest, green ancestor), ancGFP+10 (an intermediate genotype with ten substitutions), and ancGFP+15 (our “derived” state). All three proteins exhibit the expected fluorescence spectra when studied in bacteria or when purified (Fig 6A). Further, we have showed that we can express these proteins as fusions

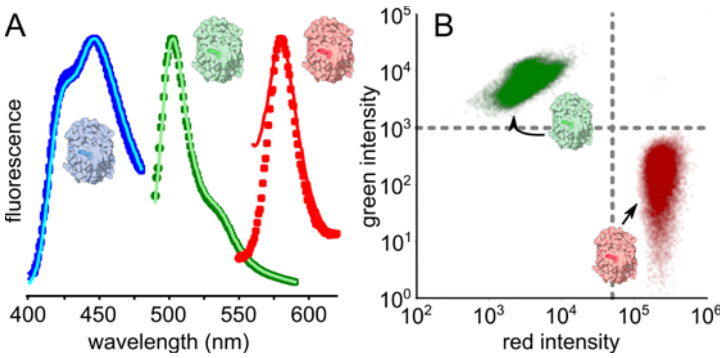

Fig 6: GFP-like proteins are amenable to FACS sorting. A) Emission spectra of eBFP2 (blue), ancGFP (green), and ancGFP+15 (red) expressed in bacteria. Solid lines are spectra measured for individual proteins. Points are spectra measured for eBFP2-ancGFP and eBFP2-ancGFP+15 fusion proteins. B) ancGFP and ancGFP+15 are readily separable by FACS. Experiment sorted with an 1:1 mixture of bacteria expressing ancGFP and ancGFP+15. Bacteria appear as discrete populations after FACS sorting.

with eBFP2. The fluorescence spectra of both the ancGFP constructs and eBFP2 are identical when measured individually or when in fusion, indicating minimal FRET or direct interaction between the proteins (Fig 6A). We have also validated our ability to sort these bacteria by creating a mixture of bacteria containing ancGFP and ancGFP+15 and then performing FACS (Fig 6B)

2.1.5: Analysis

We will discuss our analysis in more detail in Aim 3. Briefly, however, we plan three immediate studies using this dataset. 1) We will look for the emergence of large, connected networks with increased dimensionality. We predict that a 15-dimensional map will have a vast number of allowable trajectories and, as a result, be robust to large numbers of inviable intermediates. 2) We will analyze high-order epistasis in the ratio of red/green fluorescence. We previously analyzed a large number of published binary genotype-phenotype maps and found extensive high-order epistasis that does not decay with increasing map size. We hypothesize that high-order epistasis continues to arbitrarily high-order. This high-resolution GFP map will allow us to test this hypothesis. 3) We will study the robustness of our conclusions about these trajectories to more realistic and relaxed evolutionary models that include neutral drift and multi-step mutations.

2.1.6: Anticipated Problems and Solutions

Given the ease of expressing and characterizing these proteins in bacteria, we anticipate few problems on that front. We have experience in high-throughput screens and are also unconcerned about the molecular biological aspects of the project. The most likely issue we will face is statistical power. Mathematically, we should be able to characterize all 49,152 clones, but issues such as library bias could lead to less-than-perfect coverage. In the event we cannot measure all genotypes, we can simplify the problem. Twelve mutations are sufficient to create the red phenotype,[58] so we can productively study maps anywhere from 2^{12} to 2^{15} genotypes while still gaining insight into th system. We have also identified local experts that can help us with unforeseen technical problem, including our staffed HTS and FACS cores (see Facilities) and UO colleagues who regularly use FACS in their research.

2.1.7: Scientific Product

Achieving this aim will result in a rich dataset capturing the transition map between two discrete protein functions. We will release the entire dataset in a standardized format upon publication. This is particularly powerful for the evolutionary genetics and evolutionary biochemistry communities, as many meta-analyses are performed on existing combinatorial datasets.[43,48,67,94] Aside from evolutionary insights, this work will also provide a detailed map of the sequence-function relationship for a class of technologically important fluorescent proteins.

2.2: Develop computational tools to study high-dimensional genotype-phenotype maps

We propose studying a map with ~2 billion-fold more trajectories than any yet studied. Current computational tools are not up to this task. To solve this problem, we will borrow a page from chemical physicists, who have developed powerful mathematical machinery for treating

Markovian processes that explore many degrees of freedom. Because molecules have many degrees of freedom (translation, bond rotation, vibrational modes, etc.) a typical chemical reaction can populate a vast number of possibilities between the reactant state and the product state. This is directly analogous to an evolutionary process that can explore many different genotypes between an ancestral and derived state. We can therefore take computational tools designed for studies of chemical reactions, and use them for studies of evolutionary trajectories.

We start with the common assumption that evolution behaves as a Markov process[95–101]—meaning that the probability of a substitution depends only on the current genotype and current evolutionary scenario, not specific past events. This allows us to write a transition matrix *T* that describes the probability of fixing some genotype *j* given that the population starts with genotype *i*. This local substitution probability can be calculated using any appropriate population genetics model.[95,97,99] Once *T* is known, the dynamics of trajectories through the map can then be studied.[8,37,46,47] Fig 7 shows this calculation for the strong-selection/weak-mutation model.[97,99] Under strong selection for orange and single-step moves, most evolutionary moves are not allowed. Only a few transitions have some probability (000 → 010, for example). If one iteratively applies this transition matrix to a population, one obtains the trajectories seen in Fig 7D.

This approach is extremely powerful, as it provides a simple and transparent way to link arbitrary population genetics models to trajectories and dynamics through the overall genotype-phenotype map. Population genetics determines local transition probabilities; the matrix then allows determination of trajectories and other dynamical properties. In 2.3.3, we describe how we can use this formulation to sweep through different evolutionary scenarios to probe which features of trajectories depend strongly on a specific population genetic scenario.

In this Aim, we will describe three pieces of software we will develop to allow characterization of trajectories in large genotype-phenotype maps. These tools will be useful for many studies enabled by high-throughput protein characterization. We will adapt existing approaches and software libraries from other disciplines, rather than reinventing the wheel. All software will be written in Python extended with the Python scientific computing stack.

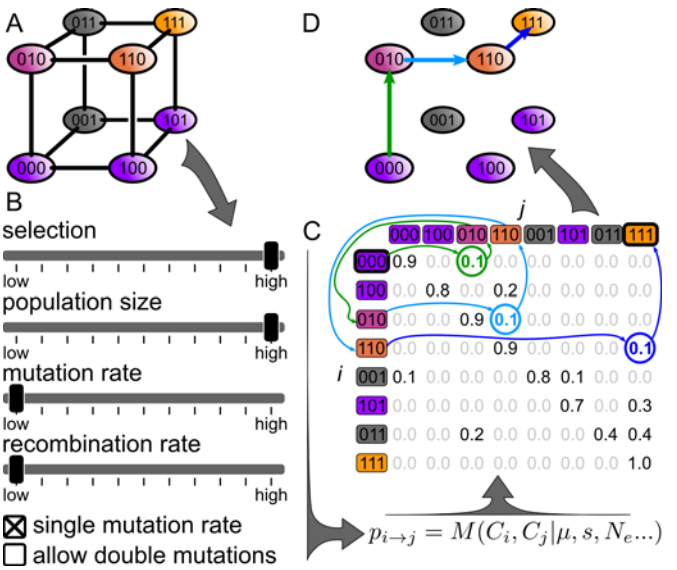

Fig 7: Use transition matrices to encode evolutionary models. A) A measured genotype-phenotype map (purple to orange). B) Using a population genetics model, one can calculate the probability of reaching *j* given *i* ($p_{i\to j}$). C) This can be placed in a transition matrix. By applying the matrix (colored arrows within matrix), one can find evolutionary trajectories (shown in D). The matrix shown is for the strong-selection/weak-mutation model. By altering the model used for *T*, one can test how population genetics shapes trajectories.

2.2.1: Efficiently extract evolutionary trajectories and their probabilities using Transition Path Theory (TPT)

One fundamental question is the relative probabilities of evolutionary trajectories given a genotype-phenotype map and population genetics scenario.[8,46,102,103] The number of trajectories in a map grows rapidly as the number of sites increases (Fig 2A). This explosion of trajectories is an important, general problem facing researchers performing high-throughput studies of genotype-phenotype maps. For example, the 15-site map we plan to characterize will have 1.7×10^{14} forward trajectories—far

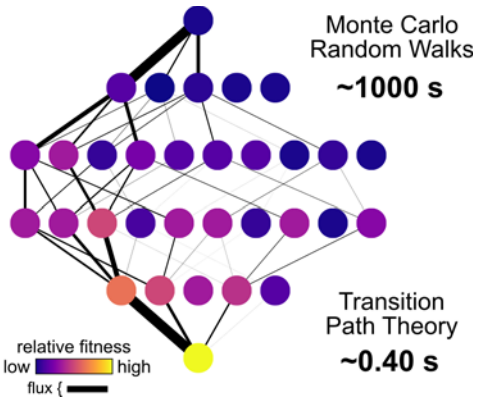

Fig 8: Transition Path Theory allows rapid identification of evolutionary trajectories. Paths through a 32 genotype-fitness map calculated using Monte Carlo walks and TPT are identical. Color denotes fitness; edge width denotes calculated flux. Map is from Weinreich et al.[8]

too large to characterize effectively by random sampling.

We will solve this evolutionary genetics problem using technology from chemical physics: Transition Path Theory (TPT.)[104–106] TPT was developed for tracing the computed trajectories of chemical compounds across "potential surfaces." As an input, it takes a transition matrix describing the probability of each genotype fixing given each each starting genotype. One then identifies an ancestor and derived sequence of interest—a "source" and "sink" in chemical parlance—and constructs possible paths between them as a series of individual transitions. TPT uses a series of linear algebraic transformations to efficiently prune inaccessible trajectories, allowing determination of the probabilities of every populated trajectory.

The output of a TPT calculation is the probability of each pathway relative to all others. This same information can be achieved using Monte Carlo simulation of possible evolutionary trajectories, albeit in a computationally inefficient manner.[37] Fig 8 shows the probabilities of evolutionary trajectories through the an experimental genotype-fitness map under a strong-selection/weak-mutation model.[97,99] We obtain identical results using Monte Carlo sampling versus TPT, but the timing is dramatically different. On a laptop, the Monte Carlo run took ~1,000 seconds to converge, while TPT took 0.4 s. This 2,500-fold speed up means that much larger maps are accessible for trajectories. We simulated a 16-site map—larger than our experimentally proposed map from Aim I—and were able to run a TPT calculation in 95 minutes. In contrast, the projected time for a Monte Carlo simulation would be 5.5 *months.*

Our preliminary work shows the utility of the method, but further development is required to make this into a useful, stable tool. This includes creating a general pipeline for calculating transition matrices given a genotype-phenotype map and generic populations genetics model, careful treatment of numerical error, and reliable error checking and validation.

This novel tool will be useful for many studies outside our own. The number of high-throughput studies of protein function increases daily, but methods for studying trajectories through maps remain ad hoc. Our package will fill this important gap.

2.2.2: Efficiently identify local fitness peaks using Perron-Cluster Cluster Analysis

Another key goal is the identification of local fitness peaks in a genotype-phenotype map. These have been identified in a variety of ways, from visual inspection to identifying highly populated genotypes by sampling evolutionary paths.[40,94] These approaches do not scale to large maps, where the number of peaks are unknown a priori and the number of genotypes may be very large.

We will solve this problem using another strategy used by chemical physicists: identification of meta-stable states through Perron-Cluster Cluster Analysis (PCCA+).[106,107] Like TPT, PCCA+ takes a transition matrix as input. It then identifies clusters of genotypes that exchange more rapidly between one another than with other genotypes. These correspond to local fitness peaks: local regions of high fitness that require a slow step (drop in fitness) to escape. In mathematical terms: if you solve the appropriate eigenvalues of transition matrix, any eigenvector with an eigenvalue of one corresponds to the equilibrium distribution of states at infinite time. PCCA+ looks for eigenvalues that are close to one, thereby identifying quasi-steady state populations of genotypes.

Fig 9 shows an application of PCCA+ to a eight site genotype-fitness map. Out of the

"hairball" of trajectories and genotypes (Fig 9A), it identifies four fitness peaks in which genotypes exchange rapidly between one another, while slowly between peaks (Fig 9B). As can be seen from Fig 9, this method is also powerful for coarse graining evolutionary trajectories. Rather than studying all trajectories, one can first cluster the map into peaks, then study trajectories between local peaks. This is directly analogous to Markov State Modeling as employed in studies of long molecular dynamics simulations.[108]

As with TPT, there is significant software development yet to be done before this is a useful tool; however, our preliminary work shows the method is feasible and should be quite powerful.

2.2.3: Predicting phenotypes of unmeasured genotypes using global phenotypic scale

The final tool we will develop does not rely on transition matrices, but instead focuses on the practical problem of "filling in" large genotype-phenotype maps given incomplete measurement of phenotypes. Although our experiment in Aim 1 is theoretically exhaustive, we may miss genotypes in our characterization. This is a common problem in high-throughput studies of protein function; therefore, there is strong interest in developing models to "fill in" these missing phenotypes given what has been measured.[109,110]

We recently found that we could explain a large amount of variation in phenotype maps using a power transform to find an empirical phenotype "scale".[48] Finding and applying such a scale should allow prediction of phenotypes using much sparser sampling of phenotypes than previously known, as it describes the variation in each map with a few, global parameters. By coupling this nonlinear analysis to a linear, pairwise epistasis model, we can predict any phenotype in the map. In preliminary analyses, addition of a nonlinear scale term led to dramatic improvements in predictions from sparse data.

We will turn this observation into useful software by reformulating the model as a Bayesian predictor. We will iteratively test our predictive model using simulated data, our own data sets (e.g. Aim 1), as well as using experimental data generated in collaboration with Rowena Martin's group (Australia National University). Dr. Martin's group previously measured 52 of 256 genotypes in an eight-site genotype-phenotype map of malarial chloroquine resistance.[111,112] As a first-pass, we performed a blind prediction of 10 additional phenotypes. Dr. Martin's group then measured these values. We successfully predicted 8 of the 10 phenotypes (R^2=0.92); the other two were predicted to be functional when they were not. Follow up analyses suggested that our model failed to correctly describe phentoypes that were below the detection limit of their assay. We will update our model with a logistic classifier, allowing us to describe each genotype with a mixture model capturing phenotypes below the detection limit.

2.2.4: Anticipated problems and solutions

We anticipate few problems with the software development described in this aim. Scientific

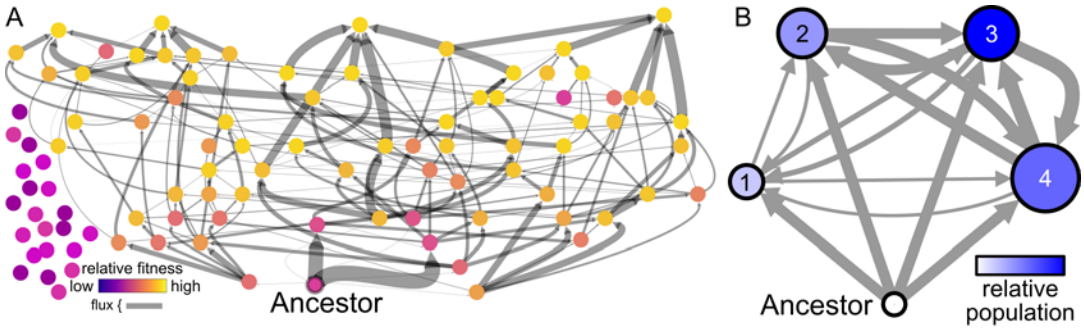

Fig 9: PCCA+ rapidly identifies local fitness peaks. A) Full experimental genotype-fitness map[40] with trajectories calculated using the strong-selection, weak mutation model. B) The same map, analyzed by PCCA+. Node size corresponds to number of genotypes; color to peak population at infinite time. Arrow width corresponds to the probability of a transition. This analysis identifies peak 3 as the global peak.

programming is in our wheelhouse (https://github.com/harmslab/). Further, the computational approaches we will use are mature and have been rigorously tested in other disciplines, so we anticipate few problems in applying them to problems in evolutionary genetics.

2.1.5: Scientific Product

This work will result in a collection of novel, powerful, open-source software packages for analyses of large genotype-phenotype maps. This includes tools to extract trajectories, identify local fitness peaks, and fill in missing phenotypes. All software will be made freely available and placed in the public domain using the UNLICENSE.[113]

2.3: Characterize the determinants of evolutionary trajectories in high-dimensional genotype-phenotype maps

With a rich experimental dataset (Aim 1) and powerful new computational tools (Aim 2), we will be positioned to probe how the dimensionality of the genotype-phenotype map shapes evolutionary trajectories. We will systematically investigate three aspects of this phenomenon: the relationship between map size and connectivity, the contributions of high-order epistasis to phenotype and trajectories, and whether the dimensionality of the map or population genetics plays a more important role in shaping evolutionary trajectories.

For these analyses, we use both the GFP-like genotype-phenotype map from Aim 1, as well as published binary genotype-phenotype maps ranging from four to eight sites.[48,94] We have amassed a curated collection of these maps and, as more results are published, will continue to update our database (https://github.com/harmslab/genotype-phenotype-maps).

2.3.1: Measure the relationship between genotype-phenotype map size and evolutionary trajectories

How does increasing map size alter the nature of evolutionary trajectories through a functional transition in a genotype-phenotype map? To ask this question, we will study the features of transitions across low-dimensional maps sampled from the main map. We will create 2^1-genotype through 2^{15}-genotype sub maps from the main map. All sub maps will include the ancestral genotype as the reference state.

We will look for neutral networks as a function of map dimensionality. As a starting point, we will define any protein that exhibits some fluorescence as functionally equivalent. If GFP-like proteins behave like other genotype-phenotype maps,[8,50,73,76,78] we expect a fraction of the intermediate phenotypes to fail to fold and/or fluoresce. We can detect variants that do not fold as those that pass the blue “expression” control (Fig 5A) but fail to exhibit either red or green fluorescence. If we score those that fluoresce as viable and those that do not as inviable, we have a dataset similar to that analyzed by Gavrilets in his original neutral network publications.[60,61] We will then identify the number and size of connected components in the map using Tarjan’s algorithm as implemented in MSMTools.[106,114]

We expect the emergence of a large neutral network as we increase the dimensionality of the map. Given the likely non-random distribution of phenotypes in the genotype map, we expect this will occur at a higher dimensionality for the protein map than a random map.

We will then probe how increasing dimensionality affects evolutionary trajectories. We will treat the ratio of red:green GFP-like fluorescence as a feature under selection. We will calculate a transition matrix for the each of the genotypes in the map using a strong-selection/weak-mutation model.[97,99] We will decompose trajectories in each sub-map through the application of PCCA+ (2.2.2) and TPT (2.2.1). We will study four characteristics of the trajectories: number of trajectories, distribution of flux through trajectories (e.g. few high-flux or many low-flux), average trajectory length, and fraction of population that ends up at the derived genotype.

For low dimensional maps (~10 sites or fewer), we predict that there will be fewer, higher-flux trajectories from the ancestral to the derived genotype for smaller maps. We also anticipate observing a few fitness peaks, corresponding to trajectories that get “stuck” before reaching the

global optimum. This reflects the apparent strong constraints on evolution observed in low-dimensional maps.

For high-dimensional maps, we anticipate a transition to a huge number of longer, lower flux trajectories corresponding to the emergence of large neutral networks. We also anticipate seeing a single, global fitness peak because the high dimensionality of the space will make it less likely that a trajectory will end up stuck at a local maximum.

We will then study the relationship between map dimensionality and antagonistic pleiotropy in shaping evolutionary trajectories. One key feature of high-dimensional maps is the emergence of connected trajectories, even when a large fraction of genotypes in the map are inviable (Fig 2B). To probe for this behavior in our measured genotype-phenotype map, we will assign random deleterious effects to individual mutations and regenerate pseudo genotype-phenotype maps. This models antagonistic pleiotropy,[115–117] where different genotypes in the space are compromised on some unmeasured fitness component. This will make the space of viable genotypes sparser, and thus allow us to study the robustness of trajectories to fractions of the space becoming inviable. By tuning the magnitude of the random deleterious effects, we can tune the sparseness of the space.

We predict that trajectories through the high-dimensional maps will be much more robust to antagonistic pleiotropy than the low-dimensional maps. This would indicate that the dimensionality of the map is a strong determinant of evolutionary trajectories.

2.3.2: Measure the contribution of high-order epistasis to phenotypic variation and evolutionary trajectories in large genotype-phenotype maps

We recently dissected high-order interactions in a collection of published genotype-phenotype maps ranging from 5-7 sites. We found extensive high-order epistasis that could not be reduced by accounting for global nonlinear scale.[43,48] This epistasis also potently shaped evolutionary trajectories.[37,43]

Intriguingly, the magnitude of the high-order epistasis does not decay with increasing order.[48] Put another way, the contribution of a five-way interaction was not (necessarily) smaller than the contribution of two-way interactions. This raises two questions: 1) how “high-order” does high-order epistasis go? and 2) is there an order at which it ceases to affect evolutionary trajectories? We have a manuscript, currently in review, where we make a statistical thermodynamic argument that high-order epistasis should not decay with order.[70] This would make protein evolution fundamentally unpredictable: knowing the phenotypic effects of mutations in the ancestral background—or even their pairwise and higher epistatic effects—would not allow prediction of future phenotypes.[37,47,71]

We will first measure the magnitude and contribution of high-order epistasis to phenotype. Measuring an L^{th}-order interaction requires measuring all 2^L combinations of mutations.[43,48,67] Our dataset from Aim 1 will nearly double L relative to previous publications (15 vs. 8) and will allow us to ask whether the contribution of epistasis decreases with increasing order. We will decompose the epistasis in this map using our established computational pipeline,[48] which identifies an empirical nonlinear scale for the map and then decomposes the epistasis using Walsh polynomials. We will propagate uncertainty in each phenotype to each epistatic coefficient, allowing us to determine whether each coefficient is statistically different from zero.

We will ask to what extent these high-order interactions shape on evolutionary trajectories. We will extract epistatic coefficients and then truncate the model to progressively lower orders.[37] Using these low-order models, we can then recalculate the phenotype of each genotype. Finally, we will determine the number and nature of evolutionary trajectories through each of these truncated maps. This will reveal the extent to which each order of epistasis shapes trajectories. Based on our previous analysis,we hypothesize that high-order epistatic interactions will potently shape evolutionary trajectories, even up to 15^{th} order.[37] This would point to deep unpredictability in evolution that grows as the number of mutations increases.[36,70,71]

2.3.3: Determine the robustness of our conclusions to more realistic evolutionary models

Our transition-matrix treatment of the evolutionary process allows us to readily explore the effects of different evolutionary assumptions on trajectories. To this point, we have proposed using a strong-selection/weak-mutation model for trajectories.[8,30,37,97,99,102] This assumes strong selection for the measured phenotype, an effectively infinite population, and that mutations occur one at a time. This does not treat important phenomena such as neutral drift or multi-step mutations. Allowing for neutral processes could—like increasing dimensionality—dramatically expand the number of trajectories and make the evolutionary process much less sensitive to the details of the genotype-phenotype map.[118,119] Likewise, allowing for multi-step mutations is expected to open up trajectories.[120–122]

We will therefore repeat the analyses described in 2.3.1 and 2.3.2 using more sophisticated population genetics models. We will calculate fixation probabilities using relaxed selection, decreased population size, and a non-zero probability for multi-step mutations. Because we need only calculate a transition matrix, we can rapidly iterate over different evolutionary models. This will allow us to determine which features we observe above are determined by the genotype-phenotype map—and are therefore robust to a variety of population genetics scenarios—and which are strongly dependent on the evolving population.

2.3.4: Product

Completion of this aim will result in a detailed understanding of the nature of the interplay between the dimensionality of a large genotype-phenotype map and evolutionary trajectories. By leveraging the experimental and computational tools developed in Aims 1 and 2, we will describe how dimensionality affects key properties of the map: neutral networks, evolutionary trajectories, and the robustness of those trajectories to external perturbation such as pleiotropy or neutral drift.

3. TIMELINE

We will use the first few years of the award period to develop the experimental and computational tools to make the desired measurements and analyses. We anticipate doing full data collection in year 2-3 and then following this up with downstream population genetics analyses over the next several years. We are firmly committed to maintaining our software; therefore, we expect to perform software maintenance throughout the award period.

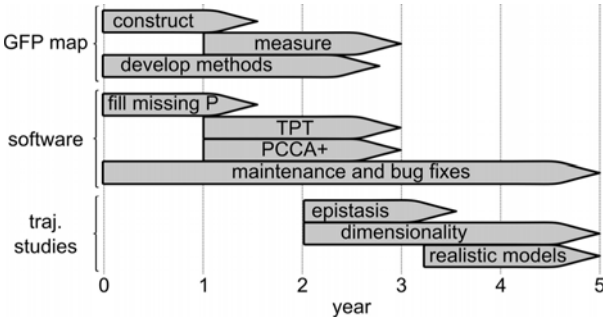

Fig 10: Project timeline. Bars show anticipated timing for each of the three aims of the grant, broken into sub-tasks.

4. BROADER IMPACTS

Completion of this work will provide deep insight into the interplay between genotype-phenotype maps, their dimensionality, and population genetics in determining evolutionary trajectories. It will also result a large, high-resolution dataset with much higher dimensionality than any yet published. This will prove useful for future meta-analyses and population genetics studies. Likewise, the novel, open source software we generate will be useful in a large number of future studies—particularly as high-throughput protein characterization continues to grow in power and prevalence.

Lastly, the proposed work will result in the rigorous training of 1.5 graduate students in the growing field of evolutionary biochemistry. These students will be well prepared to make important contributions to future evolutionary biological studies, other quantitative STEM disciplines, or industry.

5. EDUCATION PLAN

Evolutionary biology can be a challenging topic for students. This arises both from conceptual difficulties[123,124] and social factors[125]. There is an emerging consensus that evolutionary education should be restructured relative to its classic presentation.[123,126–129] Recent work has shown that focused lessons on "tree-thinking" and macromolecular processes—rather than the traditional focus on microevolutionary processes—leads to better learning outcomes.[123,127,129] Further, teaching about the mechanisms of evolution is a powerful alternative to teaching it within a purely observational and ecological context.[128]

We are in an excellent position to make measurable contributions to modern, inquiry-based[130,131] evolutionary education, with a specific emphasis on tree-thinking and mechanisms of evolution. Our research using phylogenetics to dissect the mechanisms by which proteins evolve directly overlaps these educational strategies. Further, GFP-like proteins are an excellent model to use to teach evolution. It has well-understood phylogenetics (Fig 3), known mechanisms of evolutionary change,[58,92,132] and a readily observable phenotype (color).

Our goal is to increase interest and understanding of evolution among high school-age students, with a particular emphasis on reaching students who identify as members of underrepresented groups. To achieve this goal, we propose a series of educational activities as part of the education plan. Briefly, the phases are as follows: (1) host 2-4 teachers for immersive summer research experiences; (2) collaborate with teachers to develop innovative and authentic lesson sequences; (3) support the delivery of lessons drawing on the NSF GK-12 model in which the PI and graduate students co-lead high school lessons; and (4) revise, document, and share research-connected lesson plans.

5.1 Offer Teacher Research Experience

We will recruit high school STEM teachers through the networks of UO's **STEM C**areers through **O**utreach, **R**esearch and **E**ducation (STEM CORE) and Lane STEM (see letters of collaboration and Facilities document). Recruitment efforts will target rural high school science teachers who serve high proportions of underrepresented students in STEM disciplines. These include women, potential first-generation college students, economically disadvantaged students, and minorities. 13 of the 16 public school districts in surrounding Lane County serve rural communities, most with fewer than one thousand students total. High schools in these districts have limited resources: none have committed to updating their science curriculum since Oregon's adoption of Next Generation Science Standards in 2014.[133] Further, teachers have limited opportunities for professional development. An opportunity to engage in cutting edge research and develop innovative, updated lessons would make a meaningful impact.

Participant teachers will be selected based on the following criteria: (1) educational background and preparedness to work in the Harms lab; (2) leadership experience; (3) commitment to ongoing involvement as outlined; (4) recommendation and commitment of the principal; (5) consideration for the diversity of underrepresented students served. The application process will involve an online application using a secure service (Qualtrics) followed by interviews for top candidates.

Similar to other such programs, the purpose of offering these teacher research experiences (TREs) is to allow teachers to conduct hands-on research in the laboratories of professional scientists for short periods of time (generally 1-2 summers), with the ultimate goal of improving teachers' preparedness to improve student learning outcomes. Case studies of teachers who participated in TREs have shown that science and math teachers show gains in indicators of teaching preparedness including understanding of the nature of science and confidence in teaching science and math.[134] In addition, students of teachers who participated in such programs showed significantly improved performance on standardized science exams.[135] The structure of the research experiences described here is informed by the literature that suggests that positive impacts of teacher research experiences are associated with (1) sufficient duration

of the experience; (2) clearly targeted outcomes with appropriately aligned support, and (3) deep involvement in the research process.[136]

In Year 1 and 2, teachers will engage in mentored research experiences in the Harms lab for 8-10 weeks. Ideally, we will host two teachers for two whole summers each (10 weeks x 8 hours x 2 teachers x 2 summers). In the event that we cannot recruit teachers for this length of time, we will set up shorter experiences—totaling the budgeted 1,600 total teacher research hours. Research projects will be selected in conversation with the teacher-researchers, allowing us to select a project in-line with their skills in interests. Projects could include experimental studies of GFP-like proteins of interest identified in our high-throughput analysis or direct assistance with the high-throughput screen. We will also offer computational projects investigating what determines evolutionary trajectories in these genotype-maps. We have extensive experience in getting beginning programmers up to speed—both students in the lab and through a formal scientific computing course (https://github.com/harmsm/pythonic-science).

Because we are collaborating with STEM CORE, teachers in the Harms lab will have extensive resources for translating their research into teaching practice. They will benefit from STEM CORE's participation in the Collaborative Around Research Experiences for Teachers (CARET), a partnership with twelve other institutions supporting TRE programs. CARET's mission is to promote cross-institutional collaboration involving research into and assessment of the impact of teacher-researcher programs on graduates and undergraduates. Benefits of participation in this collaborative include ready access to resources from successful TRE programs for supporting participants. Moreover, STEM CORE supports a growing teacher research community at UO in which teachers (supported by grants to the hosting faculty members) and undergraduate science majors aspiring to teaching careers (NSF #1660724) participate in summer TREs with science faculty and attend weekly education workshops.

In the last three years of the project, teachers will return to the Harms lab for three weeks each summer to continue to refine and develop their teaching materials. This sustained interaction will allow them to continue to engage in our research program, as well as refine and extend the new evolution curriculum.

5.2 Support Lesson Plan Development

In their first two summers of research experience, teachers' primary focus will be engaging in research. During weekly summer education workshops facilitated by STEM CORE, teachers will reflect their research experience and begin to generate ideas for lesson plans that draw on their experience. In consultation with me and other lab team members, teachers will produce two products by the end of their first summer in the lab: a research poster and a lesson plan. Already I have developed outlines for lessons that may inspire teachers. These include an inquiry-based study of evolutionary trajectories based on John Maynard Smith's word game[32] and a tree-thinking exercise based on tracing GFP-like colors across a phylogenetic tree.

STEM CORE faculty will guide teachers to ensure that lessons are designed for NGSS which emphasizes having students engage in science practices, or "doing real science." Teachers will pilot their lessons in the following school year with our lab team's support in the classroom. In the third summer and beyond, teachers will commit 2-3 weeks to the further development of lesson plans, expanding on the original lesson to create a full unit, in consultation with our lab team, STEM CORE, and each other.

5.3 Co-Delivery of Lessons

To support the implementation of lesson plans, the PI and graduate students from the Harms lab will join our teacher-researchers in their classrooms. We draw on the NSF GK-12 model for teacher professional development and classroom support for science involving interactions between graduate students, teachers, and K-12 students and involving many of the effective research-based characteristics of professional development.[137] STEM CORE staff has over 12 years of experience managing GK-12 programs (NSF #0231997 & 0742540) and

similarly inspired programs (Chasing Icebergs NSF award #1552232, Oregon Department of Education STEM Lab School Project). This approach has many demonstrated benefits, among them: teacher gains in confidence and content, stronger school-university connections, context for linking math and language arts to the sciences, and role models for students (MSI GK-12 program white paper, 2013). In northeastern Oregon evaluation studies showed that two or more 'doses' of GK-12 graduate students in classrooms raised state science test scores.

5.4 Document and Disseminate Lesson Plans

The goal for the third-fifth summers is to document a complete unit of lessons (2-3 weeks of plans) in an easily accessibly format that is useful to other teachers. STEM CORE staff has considerable experience facilitating such endeavors, and will provide a unit plan template and guide teachers through an iterative revision process to ensure lessons are complete, accessible, and designed for NGSS. Teachers will share their lessons publicly via Open Educational Resources (oercommons.org) website, a searchable and often-referenced source for lesson plans that allows users to update their shared plans at any time. As part of this activity, my lab team will work with teachers to ensure that the lessons are supported by necessary resources and materials assembled into ready-to-go kits available for check out. To promote the dissemination of lessons, we will co-present the project at a professional educator conference such as the Oregon Science Teachers Association annual meeting.

5.5 Education Activities Evaluation

We will collaborate with STEM CORE to evaluate our teaching outcomes using these metrics:

Activity	**Objectives**	**Metrics Gathered**
Host Teacher Research Experiences	Enhance teachers' preparedness to improve student learning outcomes especially with respect to awareness of and preparation for STEM careers	Teachers' awareness of STEM careers and use of 21st Century Skills[138] and teachers' understanding of NGSS science & engineering practices[139]
Lesson Plan Development	Teachers will develop lesson plans that introduce aspects of the research	Lesson plans analyzed for research connections and NGSS alignment using the EQuIP Rubric[140]
Co-lead high school lessons	Teachers gain confidence in their ability to engage students in authentic science activities; PI and graduate students enhance science communication skills	Semi-structured interviews with teachers, graduate students and PI, as well as open-ended responses about the research experience
Document and disseminate lesson plans	Lessons are accessible and widely adopted	Records gathered of downloads, conference presentation attendance, and supporting kit checkout

5.6 Conclusion

Through the work described in this proposal, we anticipate making important contributions to evolutionary biology education. This will occur both locally, in under-served nearby schools, and nationally through the publication of the curriculum we develop in collaboration with teachers.

REFERENCES CITED

1. Zuckerkandl, E. & Pauling, L. Evolutionary divergence and convergence in proteins. in *Evolving genes and proteins* (eds. Bryson, V. & Vogel, H. J.) 97–166 (Academic Press, New York, 1965).
2. Perutz, M. F. Species adaptation in a protein molecule. *Mol. Biol. Evol.* **1,** 1–28 (1983).
3. Ortlund, E. A., Bridgham, J. T., Redinbo, M. R. & Thornton, J. W. Crystal Structure of an Ancient Protein: Evolution by Conformational Epistasis. *Science* **317,** 1544–1548 (2007).
4. Blount, Z. D., Borland, C. Z. & Lenski, R. E. Historical contingency and the evolution of a key innovation in an experimental population of Escherichia coli. *Proc. Natl. Acad. Sci.* **105,** 7899–7906 (2008).
5. Harms, M. J. & Thornton, J. W. Evolutionary biochemistry: revealing the historical and physical causes of protein properties. *Nat. Rev. Genet.* **14,** 559–571 (2013).
6. Wilson, C. *et al.* Using ancient protein kinases to unravel a modern cancer drug's mechanism. *Science* **347,** 882–886 (2015).
7. Gabryszewski, S. J., Modchang, C., Musset, L., Chookajorn, T. & Fidock, D. A. Combinatorial Genetic Modeling of pfcrt-Mediated Drug Resistance Evolution in Plasmodium falciparum. *Mol. Biol. Evol.* **33,** 1554–1570 (2016).
8. Weinreich, D. M., Delaney, N. F., DePristo, M. A. & Hartl, D. L. Darwinian Evolution Can Follow Only Very Few Mutational Paths to Fitter Proteins. *Science* **312,** 111–114 (2006).
9. Robbins, N., Caplan, T. & Cowen, L. E. Molecular Evolution of Antifungal Drug Resistance. *Annu. Rev. Microbiol.* **71,** null (2017).
10. Anderson, J. B. Evolution of antifungal-drug resistance: mechanisms and pathogen fitness. *Nat. Rev. Microbiol.* **3,** 547–556 (2005).
11. Davies, J. & Davies, D. Origins and Evolution of Antibiotic Resistance. *Microbiol. Mol. Biol. Rev.* **74,** 417–433 (2010).
12. Wang, X., Minasov, G. & Shoichet, B. K. Evolution of an Antibiotic Resistance Enzyme Constrained by Stability and Activity Trade-offs. *J. Mol. Biol.* **320,** 85–95 (2002).
13. Hirschberg, J. & Mcintosh, L. Molecular Basis of Herbicide Resistance in *Amaranthus hybridus*. *Science* **222,** 1346–1349 (1983).
14. Karn, E. & Jasieniuk, M. Nucleotide Diversity at Site 106 of EPSPS in Lolium perenne L. ssp. multiflorum from California Indicates Multiple Evolutionary Origins of Herbicide Resistance. *Front. Plant Sci.* **8,** (2017).
15. Naqqash, M. N., Gökçe, A., Bakhsh, A. & Salim, M. Insecticide resistance and its molecular basis in urban insect pests. *Parasitol. Res.* **115,** 1363–1373 (2016).
16. Wang, X. *et al.* Mutations on M3 helix of Plutella xylostella glutamate-gated chloride channel confer unequal resistance to abamectin by two different mechanisms. *Insect Biochem. Mol. Biol.* **86,** 50–57 (2017).
17. Summers, R. L., Nash, M. N. & Martin, R. E. Know your enemy: understanding the role of PfCRT in drug resistance could lead to new antimalarial tactics. *Cell. Mol. Life Sci. CMLS* (2012). doi:10.1007/s00018-011-0906-0
18. Darmency, H., Colbach, N. & Le Corre, V. Relationship between weed dormancy and herbicide rotations: implications in resistance evolution. *Pest Manag. Sci.* n/a-n/a doi:10.1002/ps.4611
19. Evans, J. A. *et al.* Managing the evolution of herbicide resistance. *Pest Manag. Sci.* **72,** 74–80 (2016).
20. Takahashi, D., Yamanaka, T., Sudo, M. & Andow, D. A. Is a larger refuge always better? Dispersal and dose in pesticide resistance evolution. *Evolution* **71,** 1494–1503 (2017).
21. Liang, J., Tang, S. & Cheke, R. A. Beverton–Holt discrete pest management models with pulsed chemical control and evolution of pesticide resistance. *Commun. Nonlinear Sci.*

Numer. Simul. **36,** 327–341 (2016).
22. Packer, M. S. & Liu, D. R. Methods for the directed evolution of proteins. *Nat. Rev. Genet.* **16,** 379–394 (2015).
23. Bornscheuer, U. T. & Pohl, M. Improved biocatalysts by directed evolution and rational protein design. *Curr. Opin. Chem. Biol.* **5,** 137–143 (2001).
24. Huang, P.-S., Boyken, S. E. & Baker, D. The coming of age of de novo protein design. *Nature* **537,** 320–327 (2016).
25. Romero, P. A. & Arnold, F. H. Exploring protein fitness landscapes by directed evolution. *Nat. Rev. Mol. Cell. Biol.* **10,** 866–876 (2009).
26. Lutz, S. Beyond directed evolution—semi-rational protein engineering and design. *Curr. Opin. Biotechnol.* **21,** 734–743 (2010).
27. Bridgham, J. T., Carroll, S. M. & Thornton, J. W. Evolution of Hormone-Receptor Complexity by Molecular Exploitation. *Science* **312,** 97–101 (2006).
28. Bridgham, J. T., Ortlund, E. A. & Thornton, J. W. An epistatic ratchet constrains the direction of glucocorticoid receptor evolution. *Nature* **461,** 515–519 (2009).
29. O'Maille, P. E. *et al.* Quantitative exploration of the catalytic landscape separating divergent plant sesquiterpene synthases. *Nat. Chem. Biol.* **4,** 617–623 (2008).
30. Wu, N. C., Dai, L., Olson, C. A., Lloyd-Smith, J. O. & Sun, R. Adaptation in protein fitness landscapes is facilitated by indirect paths. *eLife* **5,** e16965 (2016).
31. Anderson, D. W., McKeown, A. N. & Thornton, J. W. Intermolecular epistasis shaped the function and evolution of an ancient transcription factor and its DNA binding sites. *eLife* e07864 (2015). doi:10.7554/eLife.07864
32. Maynard Smith, J. Natural Selection and the Concept of a Protein Space. *Nature* **225,** 563–564 (1970).
33. Gong, L. I., Suchard, M. A. & Bloom, J. D. Stability-mediated epistasis constrains the evolution of an influenza protein. *eLife* **2,** e00631 (2013).
34. Weinreich, D. M., Watson, R. A. & Chao, L. Perspective: Sign Epistasis and Genetic Costraint on Evolutionary Trajectories. *Evolution* **59,** 1165–1174 (2005).
35. da Silva, J., Coetzer, M., Nedellec, R., Pastore, C. & Mosier, D. E. Fitness epistasis and constraints on adaptation in a human immunodeficiency virus type 1 protein region. *Genetics* **185,** 293–303 (2010).
36. Lobkovsky, A. E., Wolf, Y. I. & Koonin, E. V. Predictability of Evolutionary Trajectories in Fitness Landscapes. *PLOS Comput. Biol.* **7,** e1002302 (2011).
37. Sailer, Z. R. & Harms, M. J. High-order epistasis shapes evolutionary trajectories. *PLOS Comput. Biol.* **13,** e1005541 (2017).
38. Brown, K. M. *et al.* Compensatory Mutations Restore Fitness during the Evolution of Dihydrofolate Reductase. *Mol. Biol. Evol.* **27,** 2682–2690 (2010).
39. Lozovsky, E. R. *et al.* Stepwise acquisition of pyrimethamine resistance in the malaria parasite. *Proc. Natl. Acad. Sci.* **106,** 12025–12030 (2009).
40. Palmer, A. C. *et al.* Delayed commitment to evolutionary fate in antibiotic resistance fitness landscapes. *Nat. Commun.* **6,** 7385 (2015).
41. Yokoyama, S. *et al.* Adaptive evolutionary paths from UV reception to sensing violet light by epistatic interactions. *Sci. Adv.* **1,** e1500162 (2015).
42. Yokoyama, S., Yang, H. & Starmer, W. T. Molecular Basis of Spectral Tuning in the Red- and Green-Sensitive (M/LWS) Pigments in Vertebrates. *Genetics* **179,** 2037–2043 (2008).
43. Weinreich, D. M., Lan, Y., Wylie, C. S. & Heckendorn, R. B. Should evolutionary geneticists worry about higher-order epistasis? *Curr. Opin. Genet. Dev.* **23,** 700–707 (2013).
44. Manhart, M. & Morozov, A. V. Statistical Physics of Evolutionary Trajectories on Fitness Landscapes. *ArXiv13051352 Cond-Mat Q-Bio* (2014). doi:10.1142/9789814590297_0017

45. Poelwijk, F. J., Kiviet, D. J., Weinreich, D. M. & Tans, S. J. Empirical fitness landscapes reveal accessible evolutionary paths. *Nature* **445,** 383–386 (2007).
46. Franke, J., Klözer, A., Visser, J. A. G. M. de & Krug, J. Evolutionary Accessibility of Mutational Pathways. *PLOS Comput. Biol.* **7,** e1002134 (2011).
47. de Visser, J. A. G. M. & Krug, J. Empirical fitness landscapes and the predictability of evolution. *Nat. Rev. Genet.* **15,** 480–490 (2014).
48. Sailer, Z. R. & Harms, M. J. Detecting High-Order Epistasis in Nonlinear Genotype-Phenotype Maps. *Genetics* **205,** 1079–1088 (2017).
49. Bloom, J. D., Gong, L. I. & Baltimore, D. Permissive Secondary Mutations Enable the Evolution of Influenza Oseltamivir Resistance. *Science* **328,** 1272–1275 (2010).
50. Harms, M. J. & Thornton, J. W. Historical contingency and its biophysical basis in glucocorticoid receptor evolution. *Nature* **512,** 203–207 (2014).
51. Shah, P., McCandlish, D. M. & Plotkin, J. B. Contingency and entrenchment in protein evolution under purifying selection. *Proc. Natl. Acad. Sci.* 201412933 (2015). doi:10.1073/pnas.1412933112
52. Gaucher, E. A., Govindarajan, S. & Ganesh, O. K. Palaeotemperature trend for Precambrian life inferred from resurrected proteins. *Nature* **451,** 704–707 (2008).
53. Voordeckers, K. *et al.* Reconstruction of Ancestral Metabolic Enzymes Reveals Molecular Mechanisms Underlying Evolutionary Innovation through Gene Duplication. *PLOS Biol.* **10,** e1001446 (2012).
54. Hart, K. M. *et al.* Thermodynamic System Drift in Protein Evolution. *PLOS Biol.* **12,** e1001994 (2014).
55. McKeown, A. N. *et al.* Evolution of DNA Specificity in a Transcription Factor Family Produced a New Gene Regulatory Module. *Cell* **159,** 58–68 (2014).
56. Bickelmann, C. *et al.* The molecular origin and evolution of dim-light vision in mammals. *Evolution* **69,** 2995–3003 (2015).
57. Kratzer, J. T. *et al.* Evolutionary history and metabolic insights of ancient mammalian uricases. *Proc. Natl. Acad. Sci.* **111,** 3763–3768 (2014).
58. Field, S. F. & Matz, M. V. Retracing Evolution of Red Fluorescence in GFP-Like Proteins from Faviina Corals. *Mol. Biol. Evol.* **27,** 225–233 (2010).
59. Howard, C. J. *et al.* Ancestral resurrection reveals evolutionary mechanisms of kinase plasticity. *eLife* **3,** (2014).
60. Gavrilets, S. Evolution and speciation on holey adaptive landscapes. *Trends Ecol. Evol.* **12,** 307–312 (1997).
61. Gavrilets, S. & Gravner, J. Percolation on the Fitness Hypercube and the Evolution of Reproductive Isolation. *J. Theor. Biol.* **184,** 51–64 (1997).
62. Aguirre, J., Buldú, J. M., Stich, M. & Manrubia, S. C. Topological Structure of the Space of Phenotypes: The Case of RNA Neutral Networks. *PLOS ONE* **6,** e26324 (2011).
63. Babajide, A., Hofacker, I. L., Sippl, M. J. & Stadler, P. F. Neutral networks in protein space: a computational study based on knowledge-based potentials of mean force. *Fold. Des.* **2,** 261–269 (1997).
64. Bornberg-Bauer, E. Randomness, Structural Uniqueness, Modularity and Neutral Evolution in Sequence Space of Model Proteins. *Z. Für Phys. Chem. Int. J. Res. Phys. Chem. Chem. Phys.* **216,** 139 (2009).
65. Reidys, C. M. & Stadler, P. F. Neutrality in fitness landscapes. *Appl. Math. Comput.* **117,** 321–350 (2001).
66. Wagner, A. Neutralism and selectionism: a network-based reconciliation. *Nat. Rev. Genet.* **9,** 965–974 (2008).
67. Poelwijk, F. J., Krishna, V. & Ranganathan, R. The Context-Dependence of Mutations: A Linkage of Formalisms. *PLOS Comput. Biol.* **12,** e1004771 (2016).

68. Ritchie, M. D. *et al.* Multifactor-Dimensionality Reduction Reveals High-Order Interactions among Estrogen-Metabolism Genes in Sporadic Breast Cancer. *Am. J. Hum. Genet.* **69,** 138–147 (2001).
69. Sun, J. *et al.* Hidden risk genes with high-order intragenic epistasis in Alzheimer's disease. *J. Alzheimers Dis. JAD* **41,** 1039–1056 (2014).
70. Sailer, Z. R. & Harms, M. J. Molecular ensembles make evolution unpredictable. *Submitted* (2017).
71. Miton, C. M. & Tokuriki, N. How mutational epistasis impairs predictability in protein evolution and design. *Protein Sci.* **25,** 1260–1272 (2016).
72. Ugalde, J. A., Chang, B. S. W. & Matz, M. V. Evolution of Coral Pigments Recreated. *Science* **305,** (2004).
73. Hietpas, R. T., Jensen, J. D. & Bolon, D. N. A. Experimental illumination of a fitness landscape. *Proc. Natl. Acad. Sci.* **108,** 7896–7901 (2011).
74. Hietpas, R., Roscoe, B., Jiang, L. & Bolon, D. N. A. Fitness analyses of all possible point mutations for regions of genes in yeast. *Nat. Protoc.* **7,** 1382–1396 (2012).
75. Roscoe, B. P., Thayer, K. M., Zeldovich, K. B., Fushman, D. & Bolon, D. N. A. Analyses of the Effects of All Ubiquitin Point Mutants on Yeast Growth Rate. *J. Mol. Biol.* **425,** 1363–1377 (2013).
76. Fowler, D. M. *et al.* High-resolution mapping of protein sequence-function relationships. *Nat. Methods* **7,** 741–746 (2010).
77. Araya, C. L. & Fowler, D. M. Deep mutational scanning: assessing protein function on a massive scale. *Trends Biotechnol.* **29,** 435–442 (2011).
78. Sarkisyan, K. S. *et al.* Local fitness landscape of the green fluorescent protein. *Nature* **533,** 397–401 (2016).
79. Pauling, L. & Zuckerkandl, E. Chemical Paleogenetics: Molecular 'Restoration Studies' of Extinct Forms of Life. *Acta Chem. Scand.* **17,** (1963).
80. Malcolm, B. A., Wilson, K. P., Matthews, B. W., Kirsch, J. F. & Wilson, A. C. Ancestral lysozymes reconstructed, neutrality tested, and thermostability linked to hydrocarbon packing. *Nature* **345,** 86–89 (1990).
81. Chang, B. S. W., Jonsson, K., Kazmi, M. A., Donoghue, M. J. & Sakmar, T. P. Recreating a Functional Ancestral Archosaur Visual Pigment. *Mol Biol Evol* **19,** 1483–1489 (2002).
82. Thornton, J. W. Resurrecting ancient genes: experimental analysis of extinct molecules. *Nat. Rev. Genet.* **5,** 366–375 (2004).
83. Hult, E. F., Weadick, C. J., Chang, B. S. W. & Tobe, S. S. Reconstruction of ancestral FGLamide-type insect allatostatins: A novel approach to the study of allatostatin function and evolution. *J. Insect Physiol.* **54,** 959–968 (2008).
84. Harms, M. J. & Thornton, J. W. Analyzing protein structure and function using ancestral gene reconstruction. *Curr. Opin. Struct. Biol.* **20,** 360–366 (2010).
85. Thomson, J. M. *et al.* Resurrecting ancestral alcohol dehydrogenases from yeast. *Nat Genet* **37,** 630–635 (2005).
86. Shih, P., Malcolm, B., Rosenberg, S., Kirsch, J. & Wilson, A. Reconstruction and testing of ancestral proteins. *Methods Enzymol.* **224,** 576–590 (1993).
87. Martinez, C. *et al.* Ancestral Resurrection of the Drosophila S2E Enhancer Reveals Accessible Evolutionary Paths through Compensatory Change. *Mol. Biol. Evol.* **31,** 903–916 (2014).
88. Kuang, D. *et al.* Ancestral reconstruction of the ligand-binding pocket of Family C G protein-coupled receptors. *Proc. Natl. Acad. Sci.* **103,** 14050–14055 (2006).
89. Liberles, D. A. *Ancestral sequence reconstruction*. (Oxford University Press, USA, 2007).
90. Gaucher, E. A., Thomson, J. M., Burgan, M. F. & Benner, S. A. Inferring the palaeoenvironment of ancient bacteria on the basis of resurrected proteins. *Nature* **425,**

285–288 (2003).
91. Risso, V. A. *et al.* Mutational Studies on Resurrected Ancestral Proteins Reveal Conservation of Site-Specific Amino Acid Preferences throughout Evolutionary History. *Mol. Biol. Evol.* msu312 (2014). doi:10.1093/molbev/msu312
92. Kim, H. *et al.* A Hinge Migration Mechanism Unlocks the Evolution of Green-to-Red Photoconversion in GFP-like Proteins. *Structure* **23,** 34–43 (2015).
93. Metzger, B. P. H., Yuan, D. C., Gruber, J. D., Duveau, F. & Wittkopp, P. J. Selection on noise constrains variation in a eukaryotic promoter. *Nature* **521,** 344–347 (2015).
94. Szendro, I. G., Schenk, M. F., Franke, J., Krug, J. & Visser, J. A. G. M. de. Quantitative analyses of empirical fitness landscapes. *J. Stat. Mech. Theory Exp.* **2013,** P01005 (2013).
95. Moran, P. a. P. Random processes in genetics. *Math. Proc. Camb. Philos. Soc.* **54,** 60–71 (1958).
96. Kimura, M. Diffusion models in population genetics. *J. Appl. Probab.* **1,** 177–232 (1964).
97. Gillespie, J. H. Molecular Evolution Over the Mutational Landscape. *Evolution* **38,** 1116–1129 (1984).
98. Hudson, R. R. Generating samples under a Wright–Fisher neutral model of genetic variation. *Bioinformatics* **18,** 337–338 (2002).
99. Orr, H. A. The genetic theory of adaptation: a brief history. *Nat. Rev. Genet.* **6,** 119–127 (2005).
100. Sella, G. & Hirsh, A. E. The application of statistical physics to evolutionary biology. *Proc. Natl. Acad. Sci. U. S. A.* **102,** 9541–9546 (2005).
101. Carinci, G., Giardinà, C., Giberti, C. & Redig, F. Dualities in population genetics: A fresh look with new dualities. *Stoch. Process. Their Appl.* **125,** 941–969 (2015).
102. Flynn, K. M., Cooper, T. F., Moore, F. B.-G. & Cooper, V. S. The Environment Affects Epistatic Interactions to Alter the Topology of an Empirical Fitness Landscape. *PLOS Genet.* **9,** e1003426 (2013).
103. Chou, H.-H., Chiu, H.-C., Delaney, N. F., Segrè, D. & Marx, C. J. Diminishing Returns Epistasis Among Beneficial Mutations Decelerates Adaptation. *Science* **332,** 1190–1192 (2011).
104. E, W. & Vanden-Eijnden, E. Transition-Path Theory and Path-Finding Algorithms for the Study of Rare Events. *Annu. Rev. Phys. Chem.* **61,** 391–420 (2010).
105. Vanden-Eijnden, E. Transition Path Theory. in *An Introduction to Markov State Models and Their Application to Long Timescale Molecular Simulation* 91–100 (Springer, Dordrecht, 2014).
106. Noe, F. *MSMTools*. (2017).
107. Deuflhard, P. & Weber, M. Robust Perron cluster analysis in conformation dynamics. *Linear Algebra Its Appl.* **398,** 161–184 (2005).
108. Pande, V. S., Beauchamp, K. & Bowman, G. R. Everything you wanted to know about Markov State Models but were afraid to ask. *Methods* **52,** 99–105 (2010).
109. Otwinowski, J. & Plotkin, J. B. Inferring fitness landscapes by regression produces biased estimates of epistasis. *Proc. Natl. Acad. Sci.* **111,** E2301–E2309 (2014).
110. Hopf, T. A. *et al.* Mutation effects predicted from sequence co-variation. *Nat. Biotechnol.* **35,** 128–135 (2017).
111. Martin, R. E. *et al.* Chloroquine Transport Via the Malaria Parasite's Chloroquine Resistance Transporter. *Science* **325,** 1680–1682 (2009).
112. Summers, R. L. *et al.* Diverse mutational pathways converge on saturable chloroquine transport via the malaria parasite's chloroquine resistance transporter. *Proc. Natl. Acad. Sci.* **111,** E1759–E1767 (2014).
113. Unlicense.org » Unlicense Yourself: Set Your Code Free. Available at:

https://unlicense.org/. (Accessed: 19th July 2017)
114. Tarjan, R. Depth-First Search and Linear Graph Algorithms. *SIAM J. Comput.* **1,** 146–160 (1972).
115. Fraebel, D. T. *et al.* Environment determines evolutionary trajectory in a constrained phenotypic space. *eLife* **6,** e24669 (2017).
116. Østman, B., Hintze, A. & Adami, C. Impact of epistasis and pleiotropy on evolutionary adaptation. *Proc. R. Soc. Lond. B Biol. Sci.* rspb20110870 (2011). doi:10.1098/rspb.2011.0870
117. Pal, C., Papp, B. & Lercher, M. J. An integrated view of protein evolution. *Nat Rev Genet* **7,** 337–348 (2006).
118. Kimura, M. *The Neutral Theory of Molecular Evolution*. (Cambridge University Press, 1983).
119. Ohta, T. Slightly deleterious mutant substitutions in evolution. *Nature* **246,** 96–98 (1973).
120. Krasovec, M., Eyre-Walker, A., Sanchez-Ferandin, S. & Piganeau, G. Spontaneous Mutation Rate in the Smallest Photosynthetic Eukaryotes. *Mol. Biol. Evol.* **34,** 1770–1779 (2017).
121. Besenbacher, S. *et al.* Multi-nucleotide de novo Mutations in Humans. *PLOS Genet.* **12,** e1006315 (2016).
122. Park, S.-C. & Krug, J. Clonal interference in large populations. *Proc. Natl. Acad. Sci.* **104,** 18135–18140 (2007).
123. Baum, D. A., Smith, S. D. & Donovan, S. S. S. The Tree-Thinking Challenge. *Science* **310,** 979–980 (2005).
124. Short, S. D. & Hawley, P. H. The Effects of Evolution Education: Examining Attitudes toward and Knowledge of Evolution in College Courses. *Evol. Psychol.* **13,** 147470491501300100 (2015).
125. Funk, C. & Rainie, L. Chapter 3: Attitudes and Beliefs on Science and Technology Topics. *Pew Research Center: Internet, Science & Tech* (2015).
126. Gould, S. J. *Bully for Brontosaurus: Reflections in Natural History*. (W. W. Norton & Company, 1992).
127. Catley, K. M. Darwin's missing link—a novel paradigm for evolution education. *Sci. Educ.* **90,** 767–783 (2006).
128. White, P. J. T., Heidemann, M. K. & Smith, J. J. A New Integrative Approach to Evolution Education. *BioScience* **63,** 586–594 (2013).
129. Novick, L. R., Schreiber, E. G. & Catley, K. M. Deconstructing evolution education: The relationship between micro- and macroevolution. *J. Res. Sci. Teach.* **51,** 759–788 (2014).
130. Science, A. A. for the A. of. *Benchmarks for Science Literacy*. (Oxford University Press, 1994).
131. Council, N. R. *National Science Education Standards*. (1996).
132. Pakhomov, A. A. & Martynov, V. I. GFP Family: Structural Insights into Spectral Tuning. *Chem. Biol.* **15,** 755–764 (2008).
133. Council, N. R. *Next Generation Science Standards: For States, By States*. (2013).
134. Rebar, B. M., Keller, J. & Conoley, C. Exploring the teacher-researcher model for impacts on pre-service teachers' preparation for science and math teaching. in (2012).
135. Silverstein, S. C., Dubner, J., Miller, J., Glied, S. & Loike, J. D. Teachers' Participation in Research Programs Improves Their Students' Achievement in Science. *Science* **326,** 440–442 (2009).
136. Sadler, T. D., Burgin, S., McKinney, L. & Ponjuan, L. Learning science through research apprenticeships: A critical review of the literature. *J. Res. Sci. Teach.* **47,** 235–256 (2010).
137. Cormas, Peter C. & Barufaldi, James P. The Effective Research-Based Characteristics of Professional Development of the National Science Foundation's GK-12 Program: Journal

of Science Teacher Education: Vol 22, No 3. **22,** 255–272 (2011).
138. Friday Institute for Education Innovation. Teacher efficacy and attitudes toward STEM survey. (2012).
139. Hayes, K. N., Lee, C. S., DiStefano, R., O'Connor, D. & Seitz, J. C. Measuring Science Instructional Practice: A Survey Tool for the Age of NGSS. *J. Sci. Teach. Educ.* **27,** 137–164 (2016).
140. EQuIP. *Achieve* (2012). Available at: https://www.achieve.org/EQuIP. (Accessed: 17th July 2017)