EXPLORATION OF A HOLEY VERSION OF AN NK FITNESS LANDSCAPE

A Senior Honors Thesis Presented to

the Faculty of the Department of Biology and Biochemistry

University of Houston

In Partial Fulfillment

of the Requirements for the Degree

Bachelor of Science

By

Robert Aurelius Sereno Laroche

December 2018

EXPLORATION OF A HOLEY VERSION OF AN NK FITNESS LANDSCAPE

Robert Aurelius Sereno Laroche

APPROVED:

Dr. Ricardo Azevedo
Department of Biology and Biochemistry

Dr. Lisa Farmer
Department of Biology and Biochemistry

Dr. W. Anthony Frankino
Department of Biology and Biochemistry

Dr. Dan Wells, Dean
College of Natural Sciences and Mathematics

ACKNOWLEDGEMENTS

Throughout the duration of this thesis, I have been lucky enough to find support all around me in the form of academic mentors, teachers, peers, family, and friends. Of these people, I will name a few here who were critical in assisting in the success of this endeavor. First and foremost, I would like to thank Dr. Ricardo Azevedo, my primary thesis advisor. When I first approached Dr. Azevedo about the possibility of doing research in his lab I was hesitant. I had little to no coding experience and knew that if I wanted to be productive, there would be a lot to learn. Despite this, Dr. Azevedo encouraged me to explore my interest and welcomed me into an environment that would help me do just that. At times, he believed in my potential to absorb new information, obtain new skills, and conduct meaningful research more than I did. Without Dr. Azevedo's constant guidance and support, I would never have been able to accomplish what I did in this project. I would also like to thank the graduate student members of Dr. Azevedo's lab, Logan Chipkin, Hao Zhang, Kedar Karkare, and Elias Grimaldo. Over the course of this thesis, they each offered insights into both the work I was doing and the details of my future academic trajectory.

Seeing their efforts and learning about their research was a large part of the reason that I decided to embark on this journey in the first place. Each of them was equally as welcoming as Dr. Azevedo, and they made his lab space one where I was comfortable to work, ask questions, and share my ideas. I also owe great thanks to the other faculty who guided me through this process. Dr. Jennifer Asmussen as well as my two thesis readers, Dr. Lisa Farmer and Dr. Tony Frankino, were instrumental in my thesis process. Without their encouragement and willingness to support me throughout the year over which this project has taken place, I would not have been able to complete my thesis and would likely have missed one or more important thesis deadlines. The next person I would like thank is Catalina Pinto. Her kindness, patience, and willingness to listen to my sometimes lengthy thesis-related struggles helped me to move past even the biggest of obstacles that, at the time, seemed insurmountable. Talking to her about the challenge I had been facing that week, day, or hour never failed to relieve some of my stress and frustration. Finally, I want to thank my parents, Anne and Sylvain. The support they provided me in the past year and throughout all the years before then is incomparable. Without their guidance, wise words and warnings, I never would have been confident or capable enough to undertake this research or to pursue my Ph.D., which I plan to begin at the beginning of the

next academic year. Beyond all else, this experience has been humbling and has shown me just how much there is left to learn, from future research and from those that surround me.

EXPLORATION OF A HOLEY VERSION OF AN NK FITNESS LANDSCAPE

An Abstract of a Senior Honors Thesis

Presented to

the Faculty of the Department of Biology and Biochemistry

University of Houston

In Partial Fulfillment

of the Requirements for the Degree

Bachelor of Science

By

Robert Aurelius Sereno Laroche

December 2018

ABSTRACT

Evolution is an incredibly complex process that has been the subject of scientific study for well over a century. The complexity of evolution has made discovery by empirical studies alone challenging, as they often offer only a glimpse into much larger patterns. This has increased the importance of theoretical models to research in the field, which allow predictions to be made for empirical studies and enable these studies to be analyzed in a broader theoretical context. One set of models that have been especially important are those of fitness landscapes. These models describe the relationship between each genotype in a set and its fitness value and have been useful in understanding the mechanisms of divergence and speciation. Computationally, fitness landscapes can be produced to better represent the multidimensional nature of true biological systems. One insight of multidimensional landscapes is that they contain networks of genotypes of the same fitness, through which evolution and speciation could occur by neutral mutations alone. With the Nk model, the ruggedness, size and dimensionality of created networks can be adjusted. By imposing a fitness threshold on a quantitative trait specified by the Nk model, I am able to

investigate how epistasis in the quantitative trait impacts the development and characteristics of these neutral networks of genotypes. In this thesis, I explore the properties of this novel fitness landscape model and examine how landscape features influence the structure and composition of neutral networks. I show that the neutral networks that exist in landscapes of varying levels of epistasic interaction respond very differently to increasing fitness thresholds.

CONTENTS

INTRODUCTION

1.1 Mathematical Models and the Study of Evolution

Evolution is the process through which all species, past and present, have come into existence. The theory of evolution provides a basis for understanding how life developed from the simplest of organisms into the incredible biodiversity seen on Earth today. Since Darwin's famed *On the Origin of Species* in 1859, scientists and laymen alike have been captivated both by evolution and by the mechanisms that drive and guide it. Despite this fact, the field of evolutionary biology remains relevant and multitudes of researchers strive constantly to better elucidate the complexities of evolutionary change. For this reason, I have decided to conduct my senior honors thesis research in this field and contribute to the greater understanding of the forces responsible for life as we know it.

Historically, both the sheer complexity of the evolutionary process and the time scale on which evolution occurs have introduced challenges to experimental studies of the subject. In light of this, the use of mathematical models to represent biological phenomena and complement empirical work

by serving as a proof of concept is of incredible importance to the field (Servedio et al. 2014). Specifically, the study of speciation is a component of evolutionary biology that has benefitted immensely from mathematical modeling. For the first time in the 1960's and 1970's, different mathematical models were used to study speciation to varying extents (Maynard Smith 1962, Bazykin 1969, and Dickinson and Antonovics 1973). One concept, which has been the basis for many such mathematical models, is a fitness landscape.

1.2 Wright's Fitness Landscape

The idea of a fitness landscape was first proposed by Sewall Wright in 1932. Fitness landscapes imagine the relationship between genotypes and fitness as a three-dimensional surface where height represents fitness, and any given point on the surface represents a different genotype. Points closer together represent genotypes more similar in composition. Wright depicted this in his 1932 paper with a topographic map (**Figure 1**). In this metaphor, the surface represents genotype space, the set of every possible genotype. Genotype space is characterized by its dimensionality or, the number of

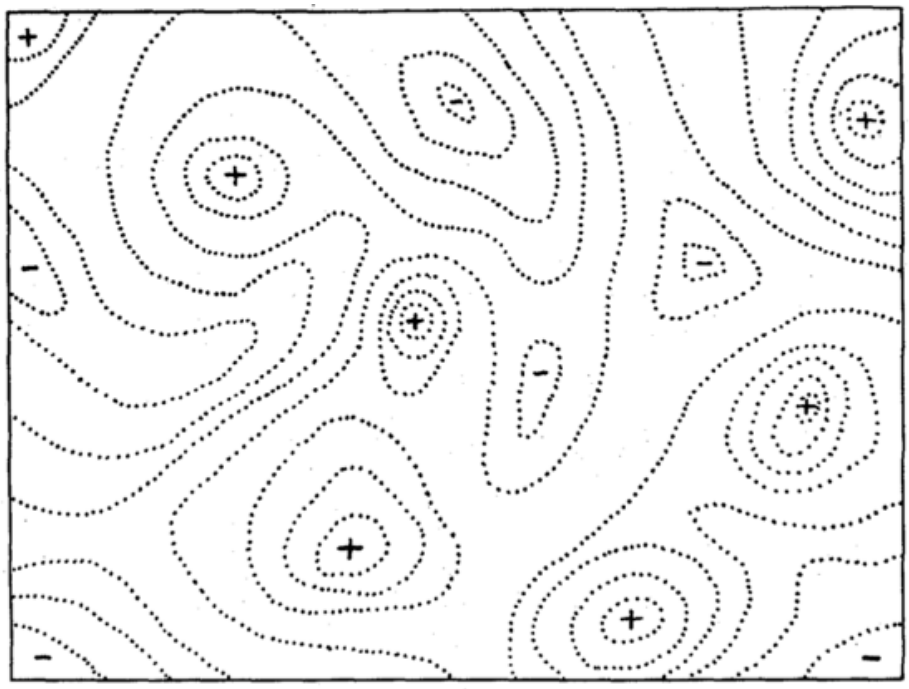

Figure 1. Wright's Two-Dimensional Fitness Landscape Diagram. Here, + signs represent areas of higher fitness and – signs represent areas of lower fitness. If the areas of high fitness are imagines as peaks, and the areas of low fitness as valleys, it is easy to see the ruggedness that Wright suggested defined this fitness landscape. Positions on the x-y plane in this diagram that are closer together represent genotypes more similar in composition.

different genotypes that can be produced from a given genotype by changing just one gene. To a certain extent, this dimensionality can and has been visualized on paper through lines connecting each genotype to every other genotype within one gene of difference from it (**Figure 2**) (Wright 1932). In Wright's fitness landscape, the more similar the composition of two genotypes, the closer they are in proximity, so the dimensionality can also be

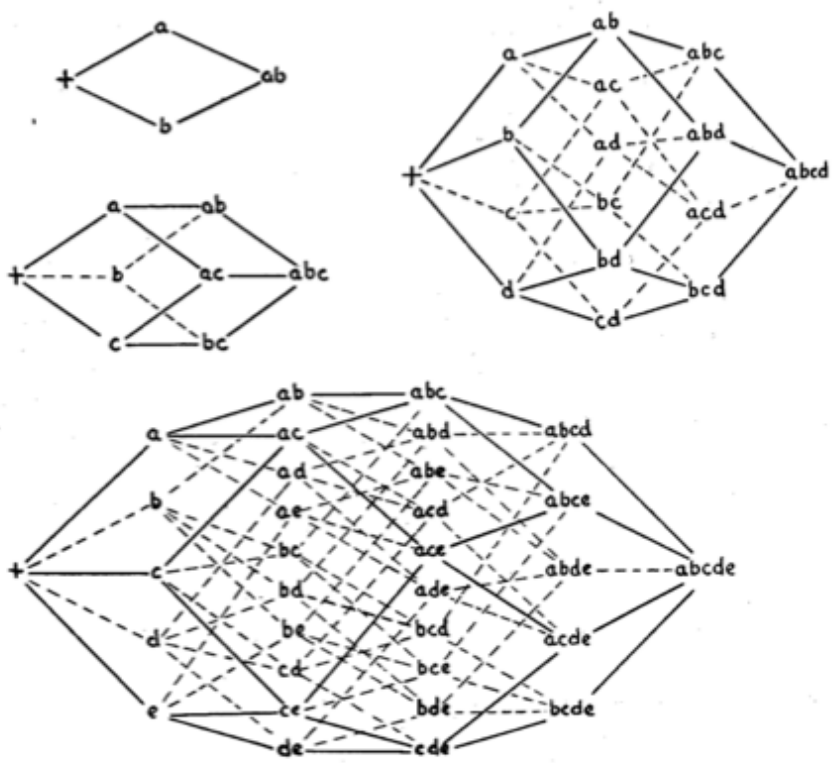

Figure 2. Wright's Multidimensional Networks. Sets of genotype spaces, increasing in dimensionality from 2 to 5, produced in the same paper in which Wright proposed his famous three dimensional fitness landscape. Every genotype is represented by a unique combination of letters and is connected to each of its one step neighbors by a line. As dimensionality increases, it becomes increasingly difficult to depict the relationship between position and genotype similarity in a two-dimensional drawing, meaning that genotypes similar in composition are not always closer to each other than to less similar genotypes.

described as the number of neighbors each genotype has. Naturally, these landscapes consist of peaks, where every surrounding genotype is of lower fitness, and valleys, where every surrounding genotype is of higher fitness. Fitness landscapes allow for the visualization of evolutionary change. As populations accumulate mutations or other genetic differences, their ability

to survive and reproduce changes and they can be imagined as moving across the landscape

Wright thought of fitness landscapes as being rugged, or having many different local peaks and valleys. In his metaphor, isolated peaks could represent distinct species, peaks clustered together could represent closely related species, and groups of peaks separated by expansive valleys might represent different higher level taxonomic groups. In this case, for evolution to occur on one such landscape, populations would need to cross the low fitness valleys between two peaks. In light of this, it might seem natural to conclude that stochastic changes in genotype frequencies could take populations into these valleys and then natural selection could drive populations up different peaks. However, this is where Wright's metaphor falls short.

In his 2004 book, *Fitness Landscapes and the Origin of Species*, Sergey Gavrilets explains why the assumption made in Wright's rugged fitness landscape that populations cross fitness valleys during evolutionary divergence is flawed. Possibly the most straightforward way for populations to move into valleys of lower fitness would be through genetic drift. In this scenario, deleterious mutations could accumulate and fix in a population, driving it away from its current peak towards another. Despite this, Gavrilets

claims that it is very unlikely that stochastic changes would ever be enough to move a population through the bottom of any fitness valley. In his book, he considers multiple models in which this might occur; a one locus, two allele model of underdominant selection (heterozygous disadvantage), a two peak model of disruptive selection on an additive quantitative trait, and a two locus haploid model of compensatory mutations in asexual populations or sexual populations with tight linkage (no recombination). In the end, he concludes that in all of these models, only under very stringent conditions (low population size and shallow valleys between peaks) can a stochastic shift from one peak to another actually occur. Further, for speciation to occur, there must be adequate reproductive isolation, a condition that is in contrast with shallow valleys between peaks. This proves then that under the conditions of Wright's rugged fitness landscape, it is implausible for stochastic changes to be the mechanism by which species diverge in natural populations.

The hypothesis that Gavrilets puts forward is that this seemingly irreconcilable problem does not lie in the mechanism by which populations diverge, but rather in the rugged landscape metaphor that Wright described. Wright's rugged fitness landscape is imagined as a three dimensional geographic terrain, something that made it an accessible concept, easily

understood and used to conceptualize evolution. However, evolution is a much more complex, multidimensional process. In a three dimensional landscape, thin ridges or very narrow, shallow valleys of high or low fitness relative to their surroundings seem unlikely to exist. In a genotype space with many more dimensions though, these ridges connecting high fitness genotypes are very common. In consideration of this, the strength of Wright's metaphor is also its weakness. The simplicity of his fitness landscape model is both what made it so broadly understood and what disguises a key structural component of biological landscapes. For this reason, Wright's rugged fitness landscape is not a good model on which to explore or imagine the properties of evolutionary change. Since the proposal of his model, there have been a number of attempts to analyze fitness landscapes produced empirically which are reviewed in Weinreich et al, 2013. These have met varying levels of success for a number of different organisms including *Escherichia coli* and *Drosophila melanogaster*, but are generally limited in terms of size and complexity (Weinreich et al, 2013).

1.3 Multidimensional Models of Fitness Landscapes

There are a multitude of different multidimensional fitness landscape models that have been proposed. The first of these that I will discuss is the multidimensional House of Cards model (Szendro et al. 2013). To understand this model, imagine genotypes with a large number of loci, N, each capable of taking on one of two allelic identities. This defines a genotype space in which each genotype has N possible one step neighbors, genotypes that differ at only one locus. Then, each locus offers a different random fitness contribution for every genotypes, producing genotypes of equally random fitness. The landscape produced by this model is therefore very rough and uncorrelated, meaning that the distance between genotypes in genotype space is completely unrelated to each genotype's individual fitness value.

The second multidimensional fitness landscape model to be considered is the multiplicative fitness model. The largest difference between this model and the House of Cards model is that the multiplicative fitness model is of a correlated fitness landscape. Genotypes closer together in genotype space have more similar fitness values. In this model, each genotype again consists of N loci, each which can take on two values. Of

these values, one is considered advantageous and the other deleterious. The fitness of each genotype as a whole is determined by the multiplication of the fitness contributions of each of its individual loci. This model is characterized by a single fitness peak and a spread of lower fitness genotypes that surround it. Other models of this nature, in which genotype space is correlated, exist as well. One example of this is the additive fitness model, where the overall fitness is determined by the addition of the fitness values of each locus rather than their multiplication.

1.4 The Nk Model

The two models discussed in the previous section can be viewed as existing on two opposite extremes of the same spectrum. In the House of Cards model, the fitness landscape is extremely rugged and fitness values of genotypes are entirely uncorrelated with those genotypes' positions in genotype space. On the other end of the spectrum is the multiplicative fitness model, in which genotypes of similar composition share very similar fitness values and the landscape is therefore very smooth.

In 1989, Stuart Kauffman and Edward Weinberger bridged the gap between these two extremes by creating what is known as the Nk model

(Kauffman and Weinberger 1989). In this model, which Kauffman and Weinberger described as “tunably rugged,”, the extent to which similar genotypes share similar fitness values is dependent on a variable, k, while the number of loci that each genotype is comprised of is defined by N. In the Nk model, k is a measure of the order of epistasis or, the level of interaction between genes in a genotype. The fitness contribution of each locus in any given genotype is determined by the fitness values of k other loci in that genotype. These other loci can be selected randomly, or in some more specified manner. Each genotype’s fitness is then determined by the average fitness contribution of each of its loci.

The value of k can range from 0 to N – 1. When k is equal to 0, there is no epistasis so each locus has a fitness value uninfluenced by any other locus in the genotype. In this case, the associated fitness landscape is identical to that of the additive fitness model and the similarity of two genotypes’ loci is highly correlated with the similarity of those genotypes’ fitness values. Conversely, when k is equal to N – 1, the fitness contribution of each locus is dependent on every other locus in the genotype. The fitness landscape associated with these parameters is very rugged and completely uncorrelated. Under these parameters, the Nk model is the same as the House of Cards model. With k equal to N – 1, the fitness of each genotype is

effectively stochastic. Until recently, many of the properties of Nk model fitness landscapes had been difficult to analyze because they are much more complex than the models at either of the extreme ends of the correlation spectrum.

1.5 Holey Fitness Landscapes

Under certain restrictions, a fitness landscape can be considered holey. In a holey fitness landscape, some fitness threshold is determined for a given network and all genotypes that exist below this threshold are considered to be dead, while all genotypes that exist above it are considered to be alive. One of the more extreme examples of this is the Russian roulette model.

In the Russian roulette model, each genotype is assigned a fitness value of either 0, reflecting an inviable organism, or 1, reflecting a viable one. The probability, P, of any given genotype to be viable can be varied. Consequently, at a P of 1, all of the genotypes in genotype space are viable, and at a P of 0, none of them are. As the name suggests, the model can be thought of in terms of playing a game of Russian roulette with each genotype with a probability of 1 – P that there is a bullet in the gun's

chamber when fired. The Russian roulette model is effectively a House of Cards model landscape with a fitness threshold imposed. In a House of Cards landscape each fitness value is random, so imposing a threshold at some level would be the equivalent of selecting a certain probability that any given genotype would be viable. Higher fitness thresholds would be equivalent to lower probabilities of viability and lower fitness thresholds would be equivalent to higher probabilities of viability. Like the House of Cards model landscape, the landscape produced by the Russian roulette model is very uncorrelated, meaning that the closeness of genotypes in genotype space is unrelated to the fitness values of those genotypes.

In the Russian roulette model, a viable genotype and any other viable genotypes that can be arrived at by undergoing changes at singular loci are considered connected. These connections can be thought of as pathways on which evolutionary change can occur. Seeing as this model has only two fitness values, 0 or 1 (alive or dead), movement through the network of connected genotypes is equivalent to a series of neutral mutations. As the value of P increases from 0 to 1, the network of genotypes becomes more connected overall until most viable genotypes are part of a single network spread throughout genotype space. The value of P at which the network transforms from a series of disconnected genotype clusters into one large

connected component is referred to as the percolation threshold (Gavrilets 2004). For this model, the percolation threshold is inversely related to N, the number of loci per genotype. This is simply because as N increases, there is a greater probability that for any given genotype, at least one neighbor will have survived the game of Russian roulette. When thresholds are imposed on correlated landscapes like the ones produced by the multiplicative fitness model, the relationship between N and the percolation threshold is not so easily defined.

Thus far I have described fitness landscapes in a mainly theoretical context, imagining loci as alleles with only two potential values. However, the next point is best illustrated by taking into account the incredible size and variability of even the simplest organisms' genomes. In this case, the number of loci, if representative of the number of genes in an organism (each with many more than two alleles) would be, by a conservative estimate, in the order of thousands and would define a genotype space with similarly high dimensionality. Taken a step further, if each site is representative of one of four base pairs in the DNA sequences that make up an organism's genome, the dimensionality would be in the order of millions. Fitness values, on the other hand, have a range that can be condensed to values between 0, where organisms have no chance of survival or

reproduction and 1, where organisms are best equipped to so do. As a result, in a biological context, there likely exists a multitude of genotypes for each possible value of fitness. Therefore, each fitness increment in genotype space is expected to contain many different neutral genotypes connected across the network. A specific biological example of this can be seen in RNA-based fitness landscapes (Lee et al. 1997).

In RNA fitness landscapes, the composition and structure of RNA molecules is used to represent fitness and genotype (Gavrilets 2004). The primary structure of RNA, the sequence of nucleotides in each molecule, is interpreted as the genotype. The secondary structure of the molecule is defined as the pattern of base pairing which can either be the classic G-C and A-U base pairs, or the weaker G-U pairs. Secondary structure is used to determine a fitness value for each RNA molecule and therefore, molecules with similar structures are imagined to have similar fitness values. With four potential bases, the possible number of genotypes in RNA landscapes is 4^N, where N is the number of bases that make up each sequence. In comparison, the maximum number of different secondary sequence structures is estimated close to $1.4848 \times N^{(-3/2)} \times 1.8488^N$, which is much smaller than 4^N (Schuster et al. 1994). This suggests again that there are many different genotypes that exist at each fitness value and together form neutral

networks. In simplified holey fitness landscapes, fitness values are reduced to a binary of either being alive or dead so that all genotypes above the fitness threshold are neutral, with no difference in fitness between them.

Networks of neutral genotypes that exist at high fitness levels demonstrate a potential solution to the problems associated with Wright's three dimensional rugged fitness landscape. Even if the genotypes associated with a small range of high fitness values are considered, populations would be able to move easily across these nearly neutral networks through stochastic changes alone. Speciation could occur as populations shifted from one high fitness genotype to another without ever having to cross any significant fitness valley. While holey fitness landscapes are a simplification of traditional fitness landscape models, they serve as useful models for the study of certain specific evolutionary concepts. An example of one such concept, is Dobzhansky-Muller incompatibilities (Gravner et al. 2007).

1.6 Dobzhansky-Muller Incompatibilities

In the early 1900's William Bateson, Theodosius Dobzhansky, and Herman Muller each contributed to the proposition of a model of the evolution of genetic incompatibility (Dobzhansky 1936, Orr 1996). This

model is best illustrated by two diverging populations which independently accumulate mutations overtime. After some period of time, each of the two populations has fixed unique advantageous or neutral genetic mutations. The model is realized when the two populations hybridize and the independently viable accumulated mutations combine to produce inviable offspring. The sets of genetic differences that can exist alone, but produce drastic decreases in fitness when observed together are known as Dobzhansky-Muller incompatibilities. Since being defined theoretically, such incompatibilities have been identified in real species pairs (reviewed in Presgraves 2010, Maheshwari and Barbash 2011). An example of one of these pairs is *D. simulans* and *D. melanogaster.* Crosses between the two species with differences in the *Lhr* locus, which encodes a protein associated with heterochromatin, caused hybrid F1 male lethality (Maheshwari and Barbash 2011). One of the most important insights of this model is that it provides a method for reproductive isolation that does not require populations to cross low fitness valleys as suggested by Wright's rugged fitness landscape.

In the metaphor of a holey fitness landscape, speciation can be thought of as the point at which two populations of genotypes in a network are separated by an area of inviability, or, more simply put, a hole. In the context of Dobzhansky-Muller Incompatibilities, these holes represent areas

where, when high fitness genotypes on either side are brought together, a DMi is produced and the resulting offspring are inviable. For this reason, a holey fitness landscape provides a useful tool for visualizing and exploring the development of DMis in diverging populations.

1.7 Thesis Objectives

In this thesis, I propose a novel model for a holey fitness landscape that confers varying orders of epistasis. In the following sections, I will explore this new Nk-model and elucidate its properties. Specifically, I investigate the shift from a disconnected network to one containing a single large component across all different levels of epistasis and suggest how this unique model may be of use in future studies.

METHODS

2.1 The Model

Networks created for this thesis were defined by three main parameters. The first two of these parameters are the ones classically used to describe Nk model networks: N, the number of loci in each genotype, and k, the number of other loci that influence the contribution of a given locus. The third parameter added in this research is the fitness threshold, above which genotypes were considered viable. Since, in this model, every genotype is either viable or inviable, the value produced for each genotype can be thought of as a measurement of some quantitative trait that exists on a range from 0 to 1 rather than a measure of fitness directly. With a strict fitness threshold, the model imposes truncation selection on this trait.

The networks produced had N values of 10. Due to the fact that each locus could take a binary value of 0 or 1, this defined networks with a maximum of 1024 genotypes (2^{10}). An N value of this size reflects the computational limits associated with generating and analyzing these networks. The k values analyzed ranged from 0 to 9, 9 being the maximum

number of genes in a 10 loci genotype that can interact epistatically with each locus (N–1). For each genotype, the trait value contribution of each locus was determined by the next k loci in the genotype sequence. If the end of the sequence was reached, the next locus considered would be selected from the start of the genotype. To expedite network creation and analysis, a lookup table of uniformly distributed random numbers between 0 and 1/N was generated under some value of k and N which calculated the trait values produced by each possible combination of loci. Trait values were calculated additively; the contribution of each locus was summed to determine the trait value produced by a given genotype, resulting in a value between 0 and 1. Each genotype was then iterated through, looked up in the table, and, if the trait value was above the fitness threshold, added to a network. Once all viable genotypes were in the network, edges were added connecting genotypes with only one locus that differed.

A range of fitness threshold values were considered from 0.2 to 0.95 in 0.05 increments. This range is justified by the fact that networks at both of the model's extremes, k = 0 and k = 9, are relatively static in terms of size, connectivity, and number of components outside of this range. With fitness thresholds less than 0.2, almost every genotype is alive, and, because the next fitness threshold increment above 0.95 is 1, every genotype above this

value is, by definition, inviable. Because the primary interest of my thesis is investigating how networks produced by this novel model change in size, composition and organization across varying fitness threshold values, these extremes of the fitness threshold range will be excluded from most future figures. Additionally, in the interest of brevity, most following figures will depict data from only the networks of k = 0, k = 3, k = 6, and k = 9. This will provide a sample of data from networks across the full range of k values analyzed in this thesis. Tables with data for each variable measured in networks of every k value are included in the appendix.

2.2 Programming Language, Packages, and Associated Applications

For this research, all coding was done in the Python programming language version 2.7.13. Python is well known for the ease with which it can be read and understood. Seeing as the work I did for this project was one of my first formal coding experiences, this was an important consideration in choosing this language for my thesis. Python's versatility stems from its simplicity and its usage of a large number of software libraries used for different specific purposes. In my thesis project, I used a number of these

libraries to create, store, analyze and represent data from the novel holey fitness Nk model I present.

NumPy is a Python library that allows for the creation and manipulation of large, multidimensional arrays. This was primarily used to create the table on which the fitness values of each genotype in the landscape can be determined. Another library used was igraph, which allows for the creation and manipulation of complex networks. This library was used to make model networks of viable genotypes and more easily measure their properties. Subsequently, the pandas Python library was used to structure, store, and analyze the data produced by the exploration of Nk model landscapes. Using this library, all data was saved as .csv files. Finally, the last library used throughout this thesis was Matplotlib. This Python library offers tools to produce plots and figures. It was used to produce many of the visual data representations pictured in this thesis.

Often, in order to better write, organize, and edit code, applications that provide better user interfaces are utilized. For my thesis project, I used one such application, Jupyter Notebook. Jupyter Notebook produces browser-based notebook documents in which I can both write and run code and view outputs in the form of tables, figures, and graphs. This operation was essential in preliminary analyses of different networks and in ensuring

that functions were performing as expected. All programming was conducted using Jupyter Notebook.

RESULTS

3.1 Describing the Network

As shown in **Figure 2**, networks are best visualized as a number of points connected to each other by lines. Therefore, the holey fitness Nk-model networks produced in this study can be separated into two basic components. The first of these components is the vertices or viable genotypes, two terms that I use interchangeably throughout this thesis. The second is the edges that connect viable genotypes that differ at only a single locus, which I will refer to as connections. Together, the quantities of these components give an idea of the size and connectedness of a landscape and begin to describe its structure. For every value of k (0–9), 10 networks were generated and analyzed across the fitness threshold range (0.2–0.9).

I first measured the average number of viable genotypes across this range (**Figure 3**). As expected, at the lower end of the fitness threshold range, the number of viable genotypes is close to 1024, the total number of genotype combinations. As the fitness threshold increases, the number of viable genotypes remains relatively constant until a threshold of

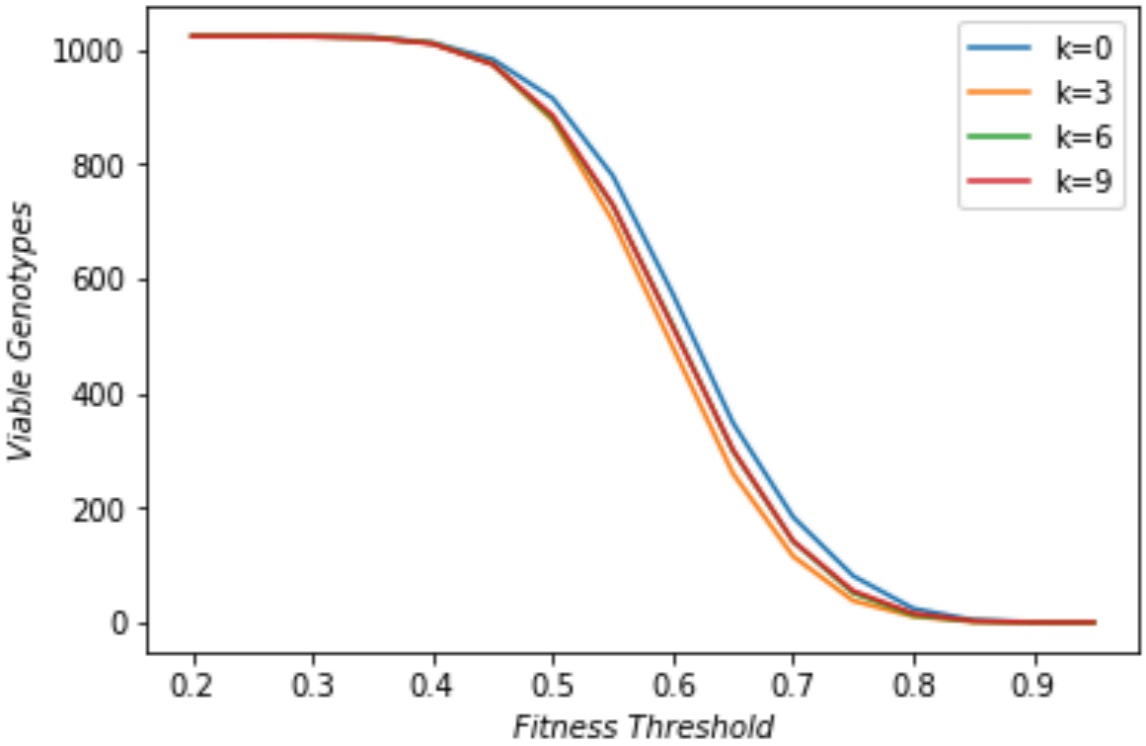

Figure 3. Network Viable Genotypes. Number of living genotypes in a sample of 10 networks as the fitness threshold increases in landscapes of k = 0, k = 3, k = 6 and k = 9 are shown above.

approximately 0.4. At this point, the number of genotypes becoming inviable at every new threshold increment accelerates until approximately 0.5, when the rate of genotypes becoming inviable becomes more constant. The viable genotypes continue to decrease at this near-constant rate until approximately 0.65, when the rate of decline slows as the total number of viable genotypes remaining in the network approaches 0. By 0.85, almost every network has no remaining viable genotypes. These trends are seen in every network, regardless of the value of k.

Secondly, I measured the number of connections in the same networks over the same range of fitness thresholds (**Figure 4**). At the low end of the fitness threshold range, the number of connections remained constant, around 5120. This number is the maximum number of connections that exists in a network made up of genotypes with 10 diallelic loci ($N \times V_{max}/2$). The connections begin to decrease at an accelerating rate from fitness thresholds of 0.4 to 0.5 before decreasing at a relatively constant rate until 0.65. Then, the rate of decrease slows until there are no connections left in the network, something that occurred in almost every network by a fitness threshold value of 0.85. This is almost identical to the trend followed by the number of viable genotypes shown previously. Again similar to what was observed for measurements of viable genotypes, the trends that the number of connections followed as the fitness threshold increased were the same across all values of k.

Another important metric mentioned previously that is used to describe fitness landscapes is dimensionality or, the number of new genotypes that can be arrived at by changing one locus of a target genotype.

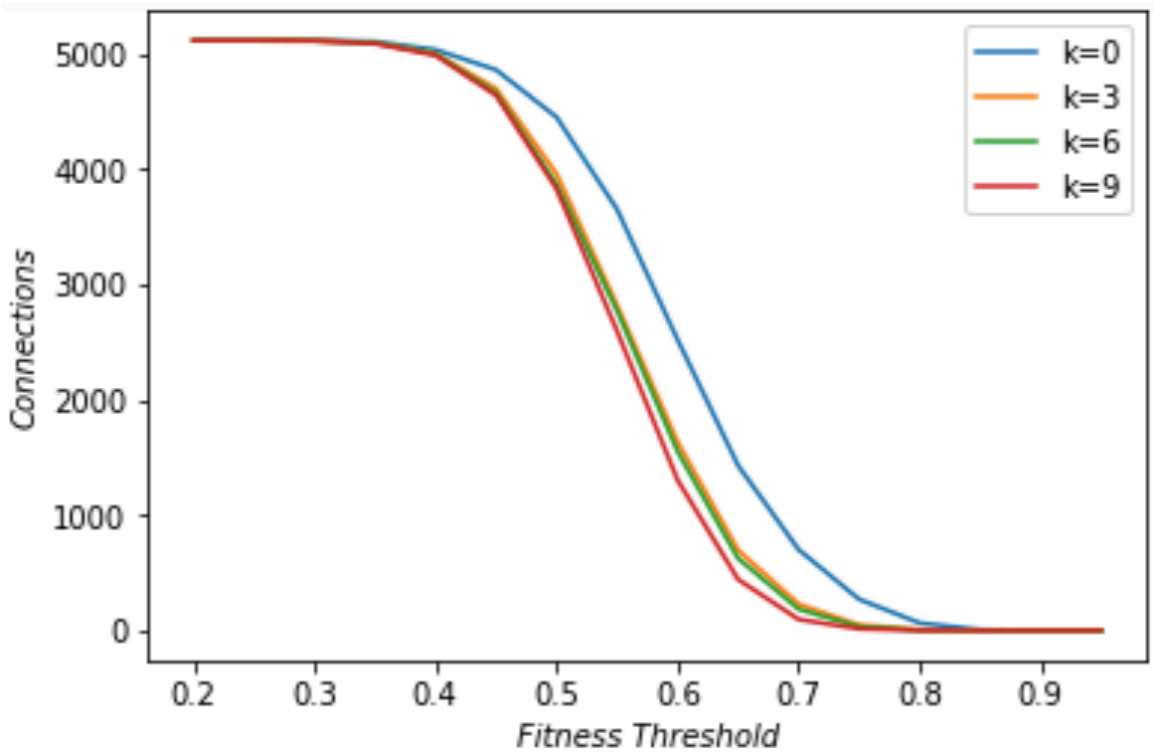

Figure 4. Network Connections. Number of connections between genotypes in a sample of 10 networks as the fitness threshold increases in landscapes of k = 0, k = 3, k = 6 and k = 9 are shown above.

In terms of connections and viable genotypes, dimensionality is the number of connections made to each viable genotype. The average dimensionality can therefore be calculated simply by dividing the total number of connections at any fitness threshold by the total number of viable genotypes at that same threshold. Then, because each connection makes a link between two separate genotypes, this number must be multiplied by two. Notably, as this duplication occurred after the connections in each network were enumerated, the values in the appendix represent half the true dimensionality

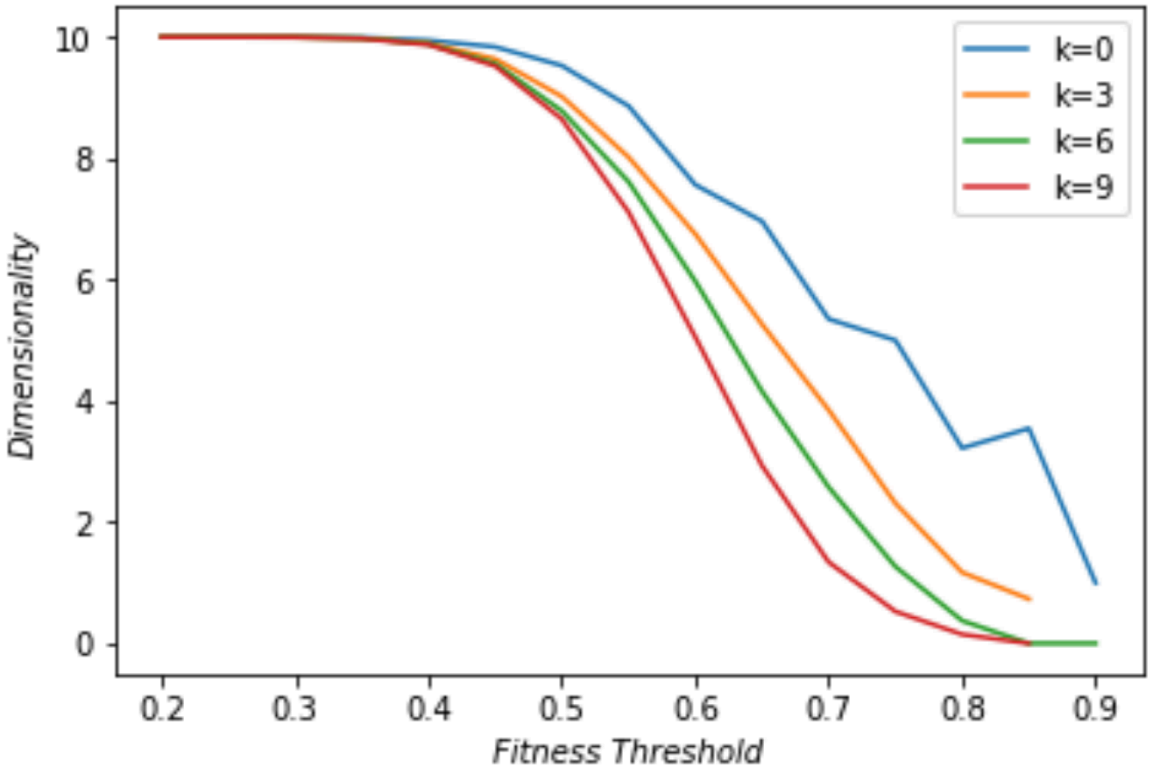

Figure 5. Network Dimensionality. Average number of connections to each genotype (dimensionality) in a sample of 10 networks as the fitness threshold increases in landscapes of k = 0, k = 3, k = 6 and k = 9 are shown above.

of networks analyzed. Naturally, because the number of connections and the number of viable genotypes respond similarly to incremental increases in the fitness threshold, the dimensionality of the network follows a similar trend (**Figure 5**). The dimensionality remains constant for fitness threshold increments early in the range before entering an accelerating decline and then decreasing at a relatively constant rate. Towards the high end of the fitness threshold range, the rate at which dimensionality is decreasing levels

out until there are no more viable genotypes or connections left in the network. In networks of low k, the slope of dimensionality against fitness threshold is more rounded and less steep. As k increases, the slope becomes more steep (see **Figure 5**) and the plotted data more closely resembles that of connections and viable genotypes across the fitness threshold range.

A final useful descriptor for these networks is the average pairwise Hamming distance. The Hamming distance is the number of loci that differ between any two given genotypes. Averaging this variable for every pair of viable genotypes in a network produces a number that describes how spread out or compact the network is. The smaller the Hamming distance, the more concentrated the viable genotypes are and the larger the Hamming distance, the more dispersed they are. Like the others, this variable was measured for a set of 10 networks, generated for each value of k and analyzed across a fitness threshold range of 0.2–0.95 (**Figure 6**). In networks of k = 0, the Hamming distance remains constant until a threshold of 0.5, when it begins to decline (**Figure 6A**). From this point until there are no more viable genotypes remaining in the network, the overall rate of decrease of the Hamming distance accelerates. For networks of k =3, the Hamming distance follows a similar pattern, but first begins to decrease around a fitness threshold of 0.6 before accelerating the rate at which it shrinks (**Figure 6B**).

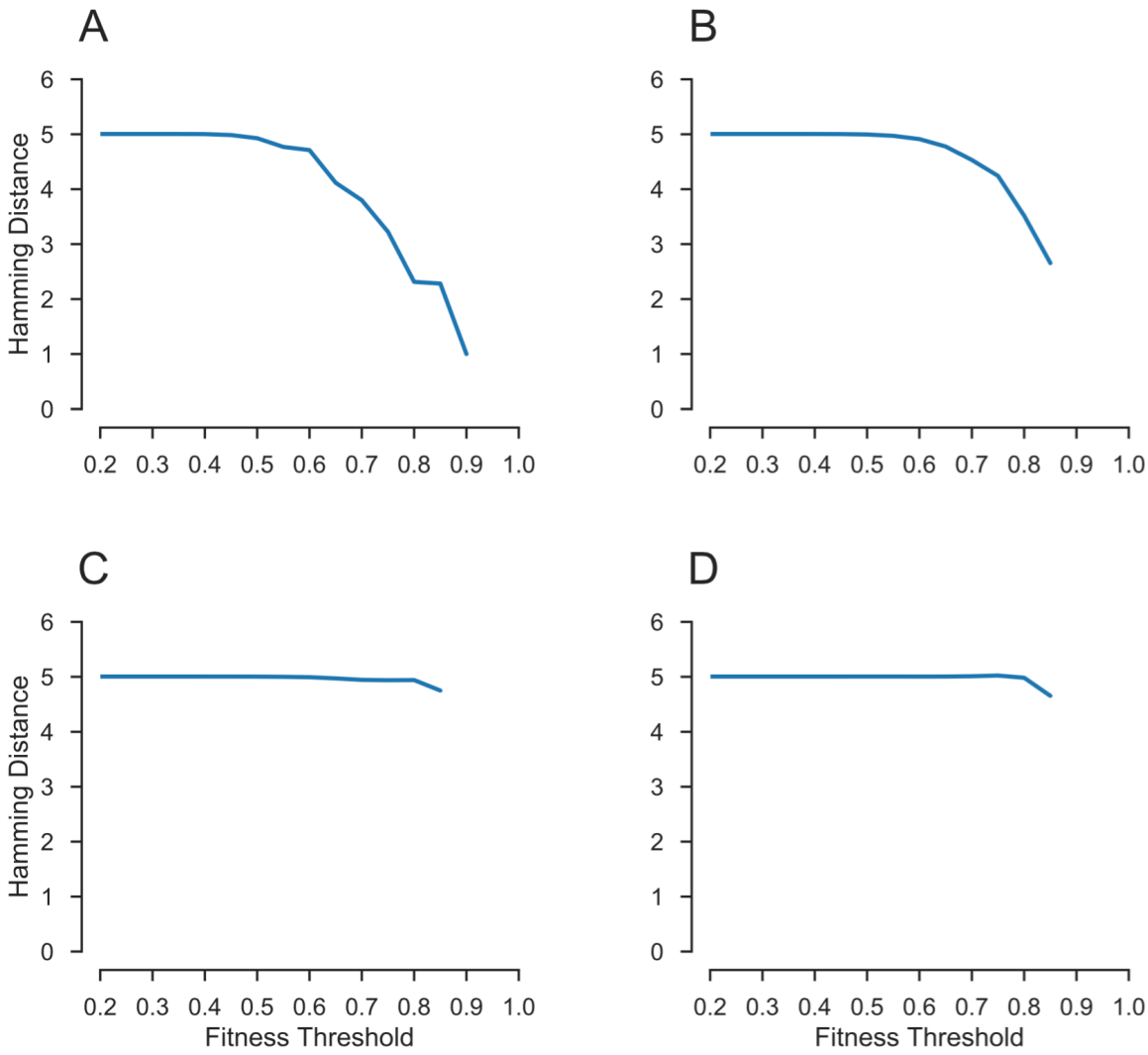

Figure 6. Network Hamming Distance. Average pairwise Hamming distance in a sample of 10 networks as the fitness threshold increases in landscapes of (A) k = 0, (B) k = 3, (C) k = 6 and (D) k = 9 are shown above.

In addition, the Hamming distance decreases overall by less than in the case of k = 0. With each increasing value of k, the Hamming distance begins to decrease at a higher fitness threshold and the overall decrease in Hamming distance across the fitness threshold range shrinks. Finally, in networks of k

= 9, the Hamming distance marginally increases before decreasing at a threshold on 0.8 (**Figure 6D**).

3.2 Ruggedness of the Landscape

One of the fundamental principles of the original Nk model was that, as k increased, the correlation between genotype fitness and proximity in genotype space decreased. It is because of this that landscapes produced by this model are referred to as "tunably rugged." To confirm that the model I propose here follows similar trends, I measured the ruggedness in trait values of landscapes produced with each value of k from 0 to 9. Specifically, I counted the average number of peaks—genotypes surrounded by only genotypes of lower fitness—in a set of 10 networks for each k value before imposing any fitness threshold (**Figure 7**). As expected, networks of k = 0 only ever had one single peak. As k increased, the number of peaks observed increased also, at a rate slightly faster than linear. Therefore, higher values of k had many more peaks than low value troughs.

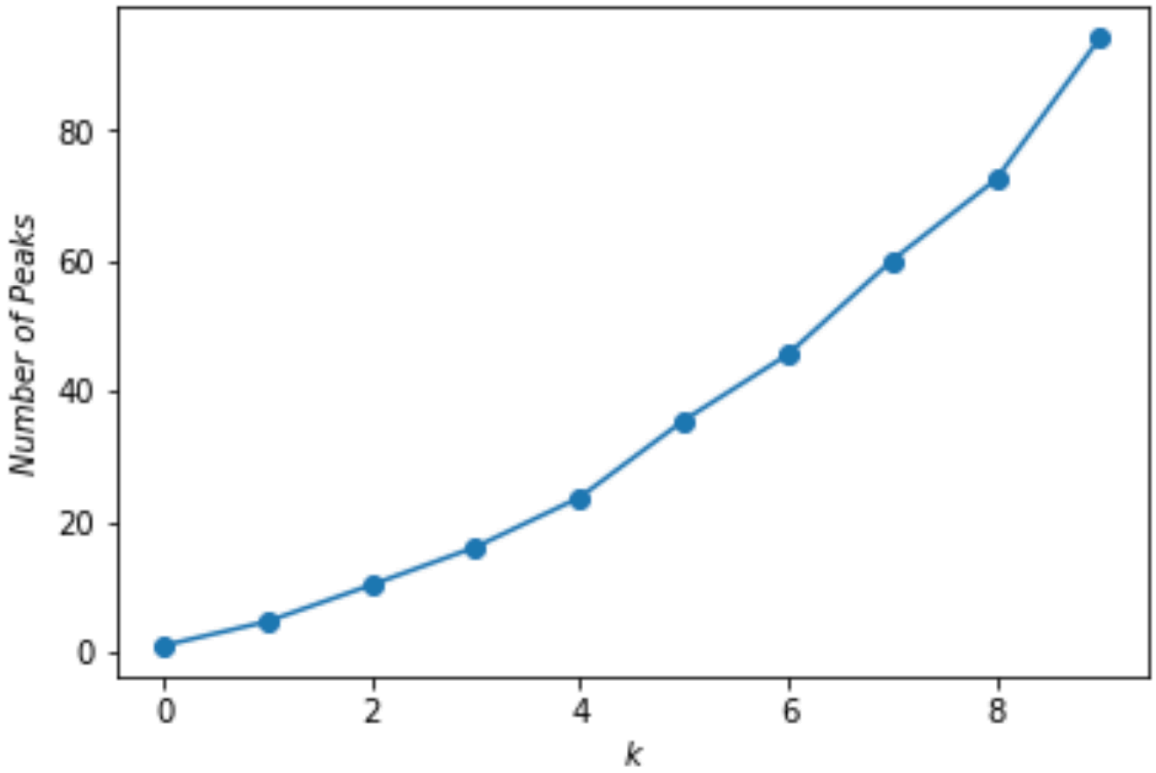

Figure 7. Network Peaks. Average number of fitness peaks in a sample of 10 networks for each k value from 0 to 9. For these networks, the fitness threshold was equal to 0 so that peaks at any trait value level would be included.

3.3 Network Components

While the number of connections in a network give a very general idea of the network's overall connectivity, it is more revealing to measure the properties of the network's components. Here, a component refers to a set of viable genotypes in a network that are connected to one another. In this work, even a single viable genotype, disconnected from the rest of the

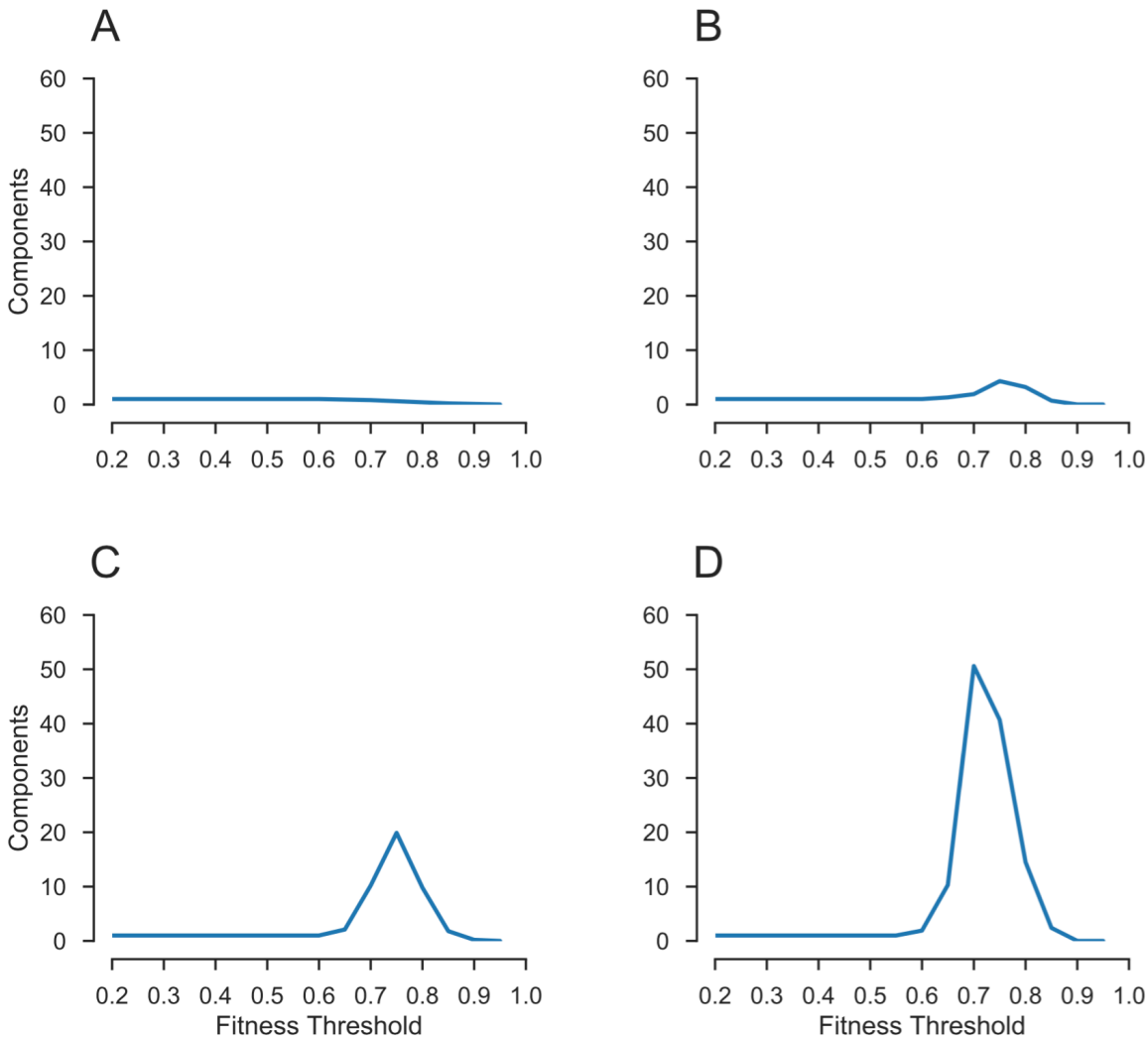

Figure 8. Network Components. Average number of components in a sample of 10 networks as the fitness threshold increases in landscapes of (A) k = 0, (B) k = 3, (C) k = 6 and (D) k = 9 are shown above.

network, is considered to be a component. I measured the average number of components for each fitness threshold increment across 10 replicate networks at every value of k (**Figure 8**).

For every value of k, the number of components at the lower end of the fitness threshold range began at 1, reflective of the fact that every or

almost every genotype was viable and part of the same, large component. For networks where k = 0, this large component endured until a fitness threshold value of 0.65. At this point, many of the 10 networks produced have no components and the average number of components quickly decreases from 1 to 0 over the next few fitness threshold increments. In networks where k = 3, instead of decreasing directly from one component to none, the networks' average number of components first increased to 4.3 and then dropped abruptly to 0. Each following set of networks of a higher k value produced a larger average number of components after the single large component broke apart, but before the average number of components dropped to 0. For networks where k = 6, the maximum average number of components was 19.9 and for networks where k = 9, the maximum average number of components was 50.6.

When there was only one component, it was made up of every viable genotype and its size was clear. However, as soon as there exist multiple components, their sizes were unknown. Therefore, after analyzing the quantity of components that exist under different threshold values, I investigated how large each of these components were in a few different ways. The first of these was a measurement of the average biggest and smallest component sizes in 10 networks produced for each value of k across

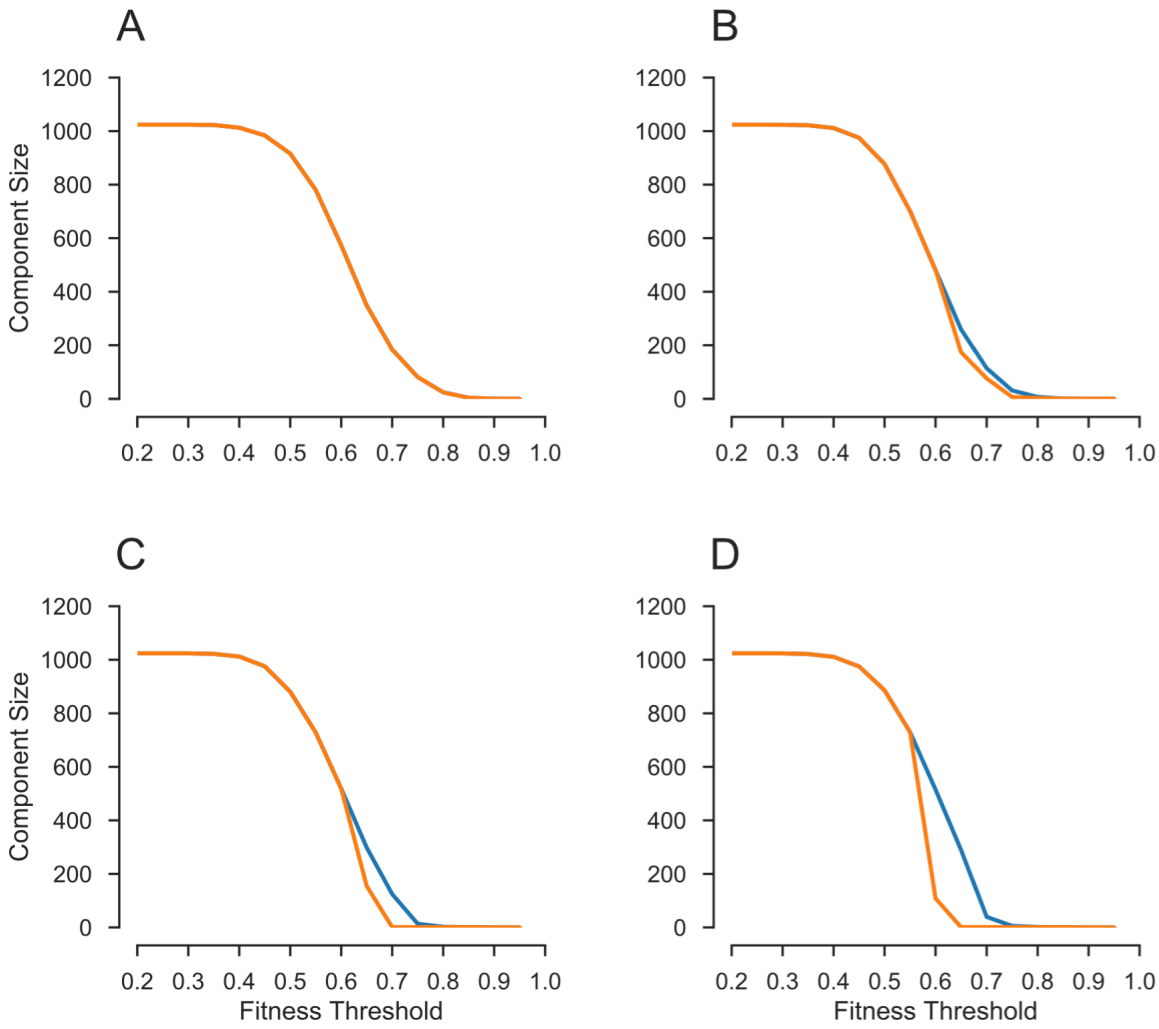

Figure 9. Networks' Biggest and Smallest Components. Average biggest and smallest component size in a sample of 10 networks as the fitness threshold increases in landscapes of (A) k = 0, (B) k = 3, (C) k = 6 and (D) k = 9 are shown above. The plot of biggest component size (blue) and smallest component size (orange) overlap entirely for the k = 0 network as there is only ever one component, which is therefore both the largest and smallest.

a range of fitness thresholds (0.2–0.9) (**Figure 9**). For k = 0, as there was either only one or no components in each network, the biggest and smallest component sizes were the same across all fitness threshold increments (**Figure 9A**). Seeing as this one component was comprised of every viable

genotype in the network, it followed the exact same trend as the number of viable genotypes did. With this in mind, further discussion of the k = 0 case for metrics of network connectivity that investigate the sizes of components is unnecessary. Instead, in figures of these metrics, the lowest k value depicted is k = 1. In the k = 3 case, the two plots begin to separate, with the average smallest component size decreasing at a faster rate per fitness threshold increment than the average biggest component size (**Figure 9B**). For each increasing value of k that followed, this divide between average biggest component size and average smallest component size widened. Networks produced with k = 9 were characterized by a much steeper decline in average smallest component size than average biggest component size (**Figure 9D**). Naturally, once fitness thresholds reached high enough values so that there were no components left in the network, the values for the biggest component size and smallest component size aligned once again (**Figure 9B–D**).

Seeing as networks of higher k values had many separate components at certain fitness threshold increments, measurements of only the largest and smallest component were not effective descriptors of how a single large component broke apart as the fitness threshold increased. Consequently, I measured the average mean component sizes for 10 networks produced for

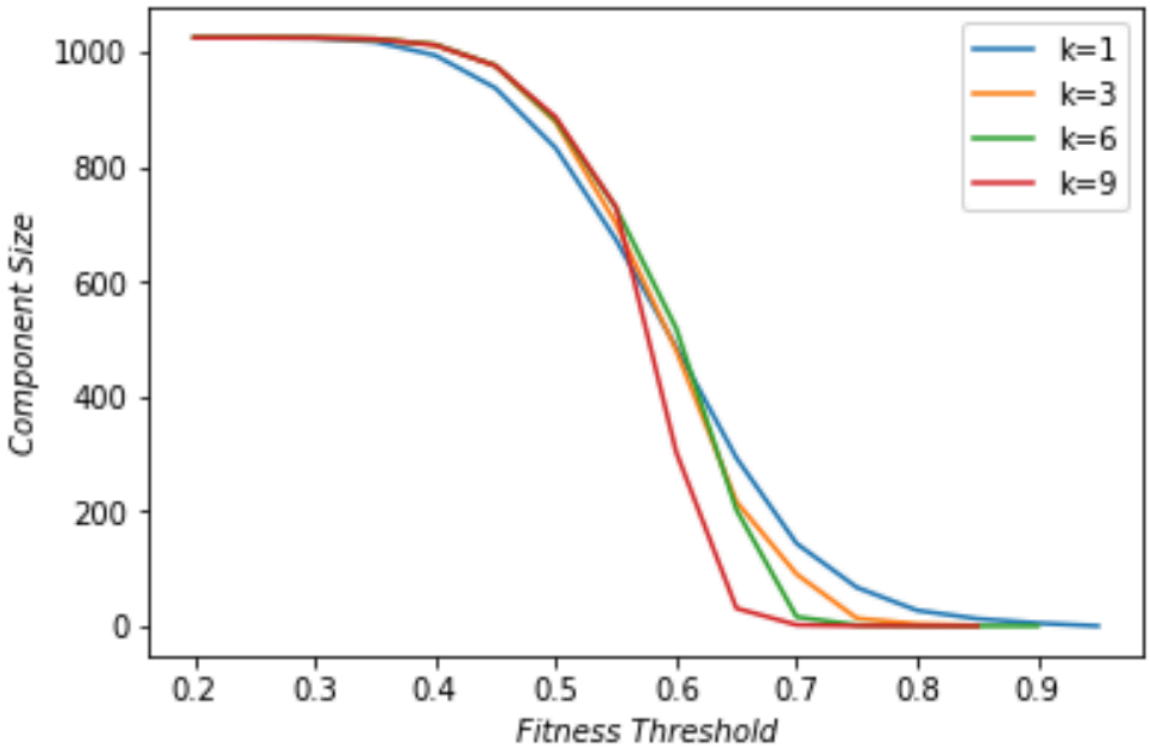

Figure 10. Mean Network Component Size. Average mean component size for each network in a sample of 10 as the fitness threshold increases in landscapes of k = 1, k = 3, k = 6 and k = 9 are shown above.

each value of k. Again, I analyzed these averages across the same range of fitness thresholds (**Figure 10**). For all networks, the initial mean component size begins at its maximum value, 1024. It remains here until close to 0.35, when it begins to decrease with every fitness threshold increment. For networks of k = 1, the mean component size reaches a value less than 10 at a fitness threshold of 0.9 (**Figure 10A**). For k = 3, k =6 and k =9, this point is reached at fitness thresholds of 0.8, 0.75, and 0.7 respectively (**Figure 10B–**

D). Therefore, as k increases, the mean component size is reduced to a value close to zero by progressively lower fitness thresholds.

Further, to account for the varying component sizes in networks being compared, I measured the average coefficient of variation of component size in networks of every value of k (**Figure 11**). Naturally, as the fitness threshold increased, the variation in component size remained constant until the network contained multiple components. At this point, the coefficient of variation increased as the number of components increased, and declined as the number of component declined (**Figure 11A–D**). In networks of higher k, the coefficient of variation reached a higher maximum value than in networks of lower k.

Lastly, I measured one final metric of network connectivity, the inclusivity of the largest component or, the proportion of all a single network's viable genotypes that are in that network's largest component. Like the other metrics, the average of this variable was calculated from 10 networks generated for each value of k for each increment in the fitness threshold range (**Figure 12**). Also like many of the of the other metrics, the proportion of viable genotypes in the big component remains at a constant value, in this case 1.00, for much of the lower fitness threshold range. After a fitness threshold value of 0.65, the proportion of viable genotypes in the

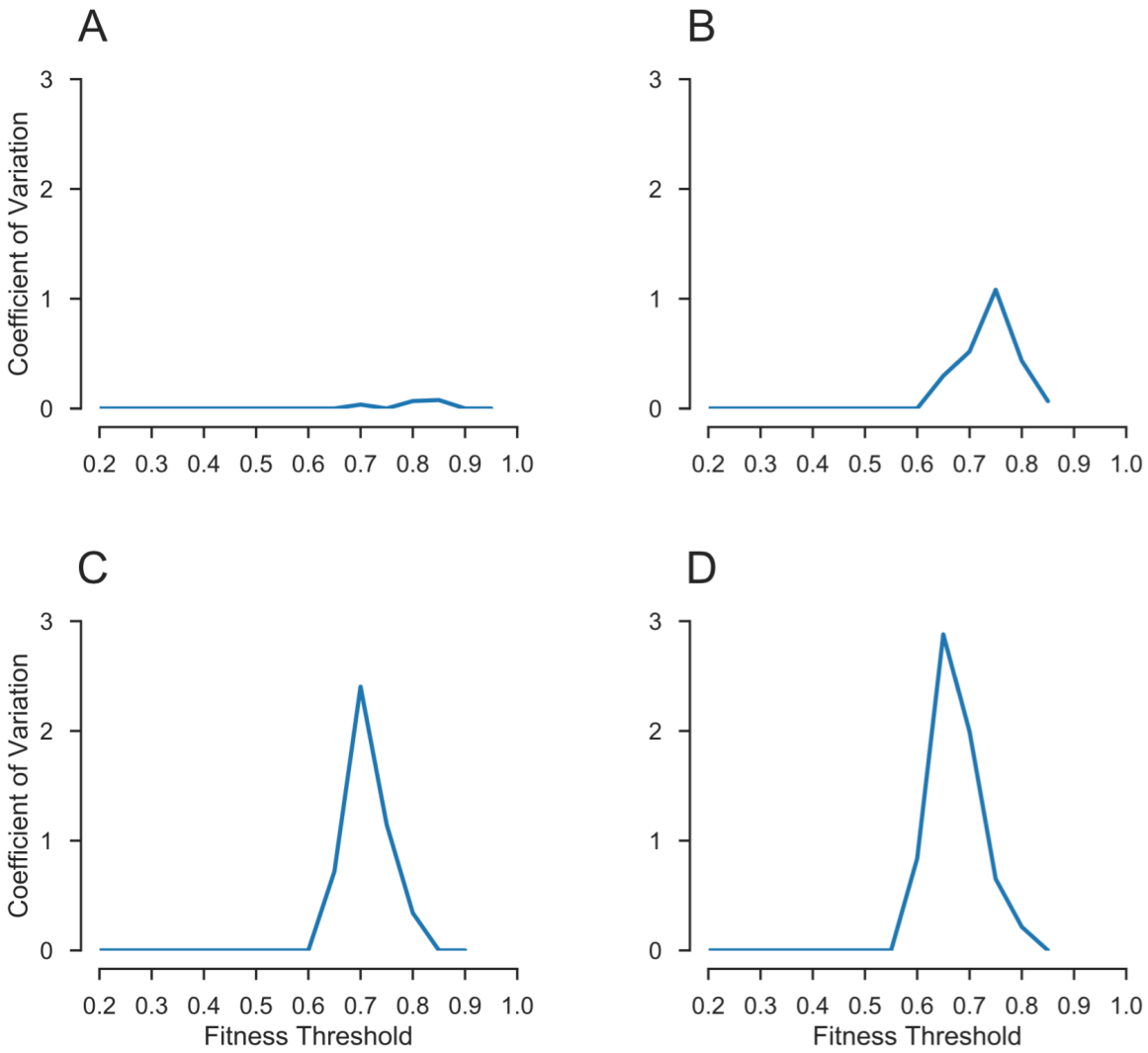

Figure 11. Network Coefficient of Variation. Average coefficient of variation for each network in a sample of 10 as the fitness threshold increases in landscapes of (A) k = 1, (B) k = 3, (C) k = 6 and (D) k = 9 are shown above.

biggest component decreases abruptly across all networks. In networks of k = 1 and k = 6, the proportion of viable genotypes in the biggest component returns to 1.00 after decreasing before the end of the fitness threshold range (**Figure 12A and 12C**). As the value of k increases, the abrupt decrease in

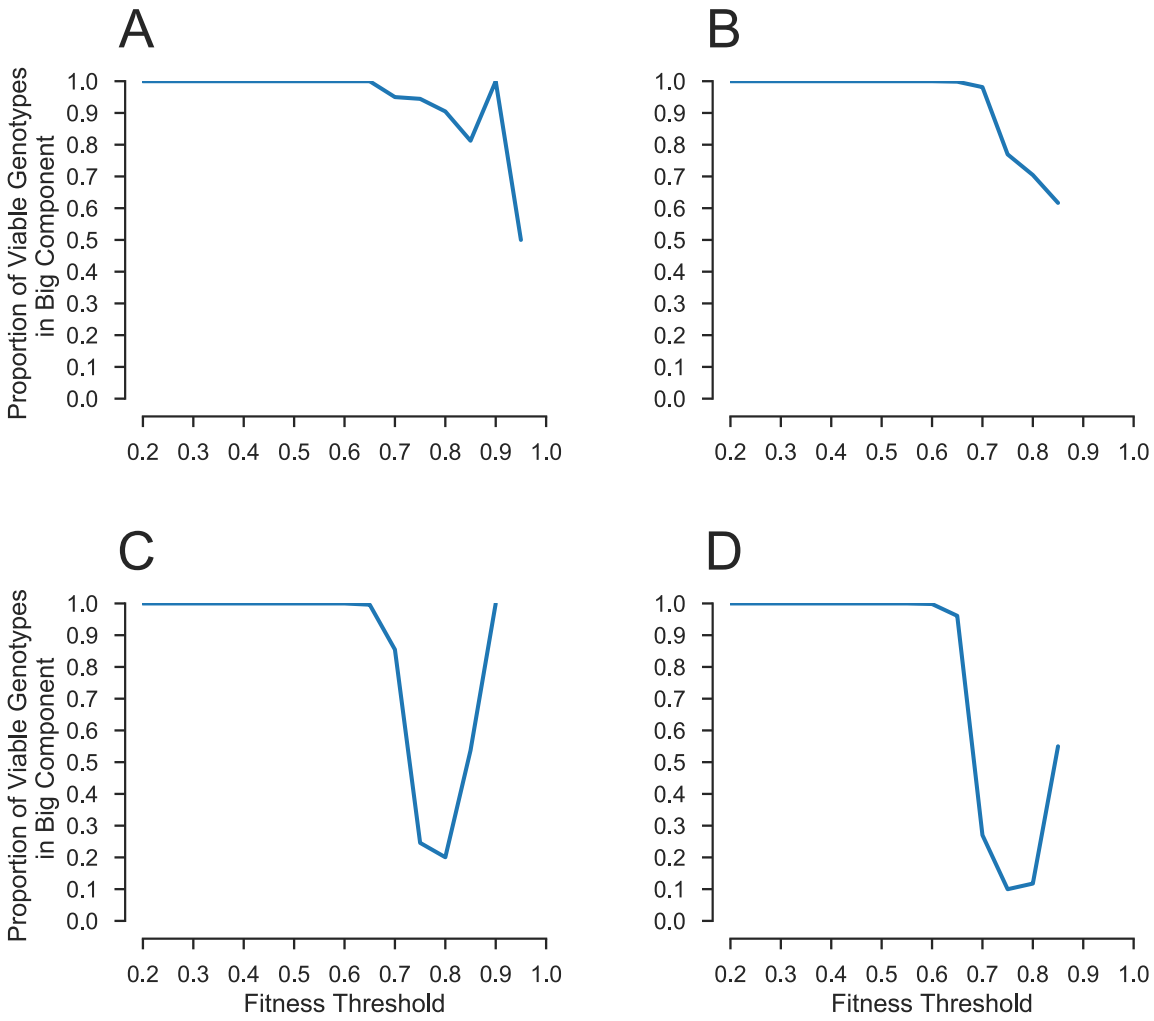

Figure 12. Network Big Component Inclusion. Average proportion of a sample of 10 networks' viable genotypes that fall within the largest component as the fitness threshold increases in landscapes of (A) k = 1, (B) k = 3, (C) k = 6 and (D) k = 9 are shown above.

the proportion of viable genotypes in the big component becomes more drastic. For networks of k = 3, k = 6, and k = 9, the proportion of viable genotypes in the big component drops to 0.62, 0.20, and 0.11 respectively (**Figure 12B–D**).

3.4 Properties of Percolation

The last property of this novel fitness landscape that I investigated was how each network quickly transformed from one large connected component into multiple as the fitness threshold increased. The threshold at which this occurs in any given network is known as the percolation threshold. Here, I define the threshold as the point at which the largest component in the network first contains less than 85% of the viable genotypes in the network. I measured this threshold for a set of 10 networks produced for each value of k (**Figure 13**). For k = 0, the viable genotypes in the biggest component never dropped below 85% as the networks never significantly divided into many different components before all genotypes became inviable. After this point, increasing values of k correlated with a lower percolation threshold.

At the percolation threshold, the number of components in each network increased quickly before decreasing again abruptly as the fitness threshold increased. While this pattern was similar among networks of k values 1–9, the maximum number of components seen at this peak was different. As k increased, the number of components the large components broke into also increased and at an accelerating pace (**Figure 14**).

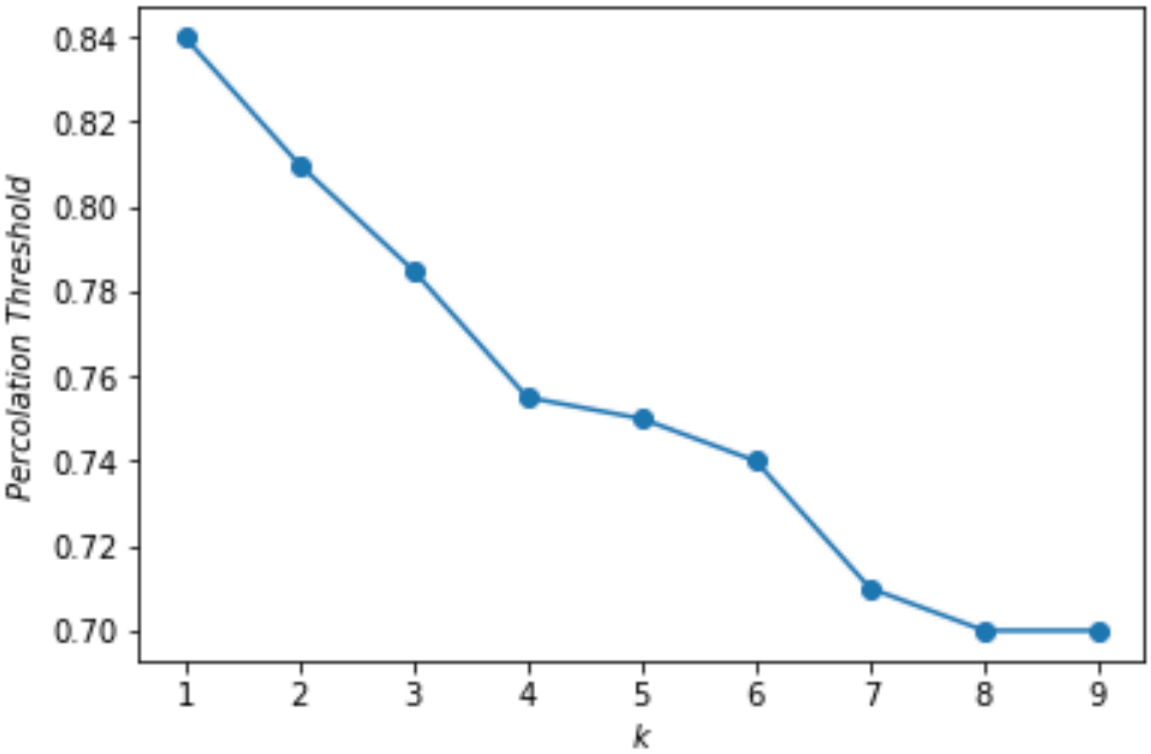

Figure 13. Network Percolation Threshold. Fitness threshold at which, for an average of 10 networks, the percentage of viable genotypes in the largest component first dropped below 85%. For networks of k = 0, the percentage of viable genotypes in the largest component never decreased below 85% so it was not included.

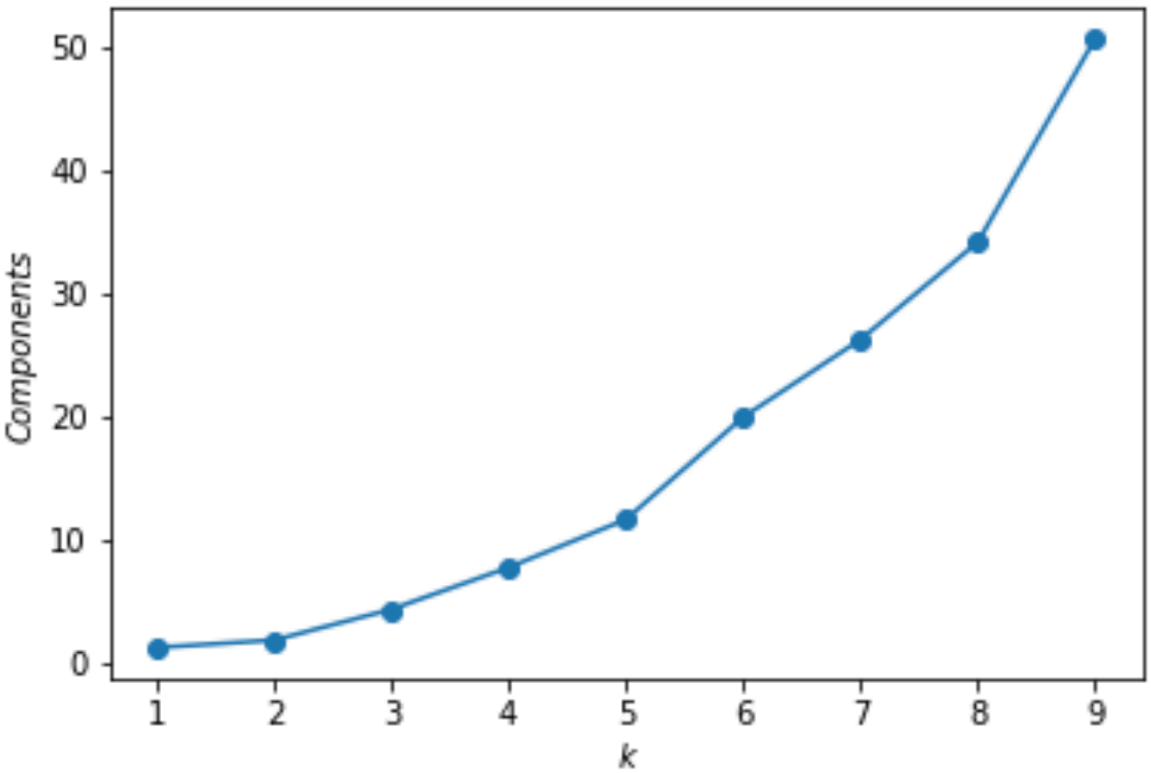

Figure 14. Network Maximum Components. Average maximum number of components to exist in a set of 10 networks for each value of k from 1–9. Networks of k = 0 were not considered as they only ever have one or no components.

DISCUSSION

4.1 The Fitness Behind the Landscape

Each of the results detailed above contributes to the analysis of the networks produced for this work. By imposing a range of fitness thresholds on these networks, we are able to elucidate how the pathways or ridges connecting different viable genotypes shift and eventually disappear as the fitness threshold is increased. This, in turn, provides a better understanding of the shape and breadth of networks produced. Further, by using a fitness landscape model in which epistasis is controlled, this research contributes to our understanding of how changes in a landscape's level of epistatic interaction play into the formation of neutral networks on which evolution by neutral mutations alone could occur.

My results highlighted the static nature of landscapes across all values of k for fitness thresholds on both the low and high end of the range. This characterized the number of viable genotypes, connections, and components, the size of both the biggest and smallest components, the hamming distance, and the dimensionality of networks analyzed. At first, it may seem

counterintuitive that as the fitness threshold below which all genotypes are inviable initially increases, no genotypes are rendered inviable. However, this is explained by the fact that the trait values of all the possible genotypes in each network form a normal distribution. Most genotypes in this model have an attribute value that falls somewhere in the middle of the attribute value range. At both the extreme high end and extreme low end of this range, there are very few genotypes. Because of this, the first fitness threshold increments exclude almost no genotypes, and by the final fitness threshold increments, there are few genotypes left to be lost. This concentration of attribute values in the middle of the attribute value range is reflective of the fitness ranges under which many biological systems operate.

4.2 The Effects of a Fitness Threshold

As the fitness threshold was increased, networks lost viable genotypes as more genotypes' associated trait values fell below the threshold. The grouping and dispersion of these lost genotypes in the landscape was dependent on the value of k associated with each network. **Figure 15** depicts a three dimensional representation of a rugged fitness landscape with an

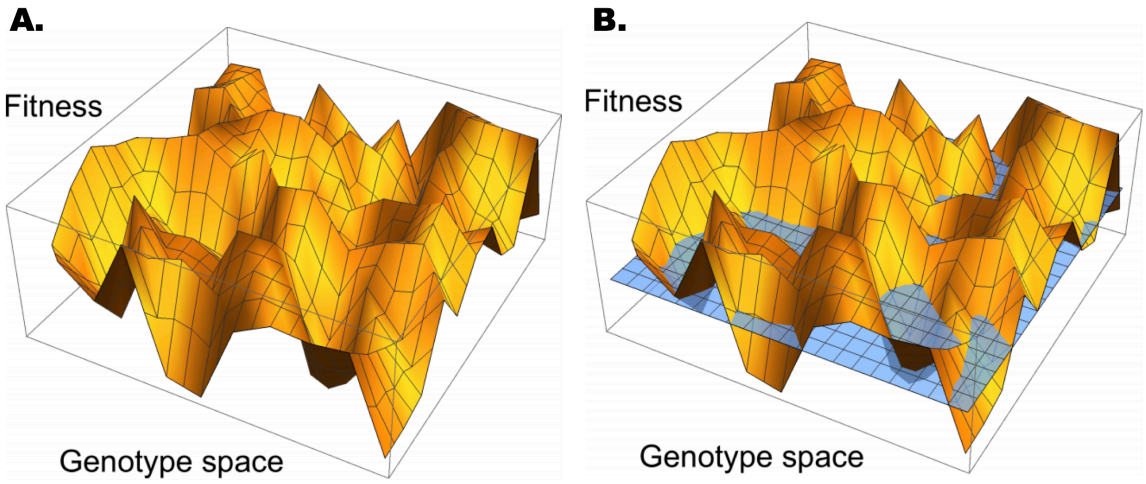

Figure 15. Rugged Fitness Landscape with and Without a Threshold. A three dimensional depiction of a fitness landscape is shown both (B) with and (A) without a fitness threshold. Every genotype in genotype space above the blue plane that indicates the level of this threshold is considered viable and every genotype beneath it is considered inviable.

imposed fitness threshold shown as a plane transecting the landscape (**Figure 15B**). In this depiction it is easy to see how, when the fitness threshold is raised, many peaks that were once connected by areas of lower trait value genotypes become isolated from each other. However, for low k, networks look much different. They are characterized by having only a few, large fitness peaks. At the extreme, k = 0, there is only one single peak, a genotype which has the highest trait value option at each of its loci. In this case, genotypes become inviable at the fringe of the network and, as the fitness threshold is raised, this fringe grows closer to the central peak. Networks with higher k values are more similar to the landscape shown in

Figure 15, with many different local trait value maxima. The concept of a set of fringe genotypes with trait values just above the fitness threshold moving upwards from the base of a peak towards its maximum can again be applied here. However, for networks with multiple peaks, this occurs for each peak independently instead of for the network as a whole.

After understanding that networks with higher k values have more peaks, the trend of seeing more components produced as k values increase is explained. In these networks, each peak represents a group of genotypes that becomes an independent component, separated from the rest of the network as the threshold rises. With each fitness threshold increment after the network ceases to be one connected component, new peaks are isolated. Finally, once the threshold surpasses the highest trait value of each peak, the component is lost entirely.

In all networks with multiple peaks analyzed in this study, the shift from having a single component to having the maximum average number of components occurred over a small fitness threshold range. This would suggest that many of the peaks begin to separate themselves from the surrounding low trait value genotypes at approximately the same level across the network. The inverse shift, from having the maximum average number of components to having none, occurs over a similarly short range of

fitness thresholds. This subsequently suggests that many of the peaks have their maximums at a similar trait value level as well. However, while many of the peaks may begin and end at similar fitness thresholds, they are not uniform in size. Investigation of the coefficient of variation showed that as soon as multiple components existed, there was variation in component size (**Figure 11**). In fact, as k was increased, the amount of variation between component sizes increased as well, a reflection of the increased ruggedness and disorder of the landscapes.

4.3 Impact of Trait Value Ruggedness

Networks in this study were further described by results showing mean component size (**Figure 10**). As k increased, the rate at which the average component size decreased across the fitness threshold range was higher. This results from the fact that in networks of higher k, with more peaks, there were many different, smaller components shrinking concurrently. In networks of lower k, the genotypes becoming inviable as the threshold increased occupied the fringes of only a few, large components, leading the average component size to decrease at a slower pace. Additionally, the fitness threshold value at which the mean component

size reached zero was lower in networks of lower k. This suggests that, in addition to being larger, the few peaks present in landscapes of low k had higher average maximum trait values as well.

As noted previously, in the k = 0 case, one large trait value maximum existed in the middle of the network, with lower value genotypes spanning out on all sides. If the many local maxima in a network at high k are imagined as tall peaks, the singular maximum in networks of k = 0 can be imagined as a large mountain. In this context, the trends seen for Hamming distance are better understood. For networks of k = 0, the Hamming distance decreased significantly as the fitness threshold increased. This is reflective of low trait value genotypes on the edges of the network becoming inviable and the remaining viable genotypes existing in an ever shrinking landscape.

For every increasing value of k, the Hamming distance decreased by a smaller overall amount. This suggests that the multiple peaks observed in landscapes above k = 0 were still somewhat centralized and that the general shape of the fitness mountain seen in the k = 0 network was not yet erased. Therefore, at higher fitness thresholds, the final peaks were more likely ones that existed in this area of higher fitness at the center of the network, explaining the continued decrease in Hamming distance. As the value of k increased further, these peaks became less and less clustered together as this

fitness mountain disappeared. In networks of k = 9, the Hamming distance barely dropped below its initial value. In fact, at high fitness thresholds, it increased marginally before decreasing at all. Because there were so many peaks in networks where k = 9, it is reasonable to imagine that the few genotypes making up the small, high trait value peaks left at high fitness thresholds may have been more spread out on average than every genotype was at the start. After this analysis, it is clear that networks with different levels of epistasis respond drastically differently to the imposition of a fitness threshold as well as to changes in that threshold's value.

CONCLUSION

5.1 Importance of the Model

The use of computational models to help understand evolution is becoming increasingly common. Producing variations of these models that allow for new examinations of evolutionary mechanisms is therefore crucial to furthering research efforts in the field of evolutionary biology. The research conducted in this thesis represents an investigation into a novel model of holey fitness landscape. Specifically, it analyzes the interplay between the level of epistasis and a range of fitness thresholds in a multidimensional network model. Other studies have developed and explored networks based on the Nk model and have even considered the percolation of neutral networks specifically (Newman and Engelhardt 1998, Gravner et al. 2007). However, none have used the imposition of a fitness threshold as I did to create or analyze these neutral networks.

Neutral networks observed in multidimensional models may play an essential role in determining evolutionary trajectories. In this study, percolation thresholds express the highest fitness threshold at which these

neutral networks can exist and span genotype space. This study elucidates and evaluates the properties of percolation thresholds in networks at every level of epistatic interaction. As a result, this work contributes to our understanding of how these neutral networks are organized and of the variables that impact their size and connectivity.

5.2 Future Research Directions

As mentioned previously, the scale and complexity of computational network models is nowhere near that of the real biological systems that they represent. For this reason, future work should focus on more efficient computational methods that allow for the creation and analysis of larger, more complex landscapes. Specifically, further research should be conducted on networks of genotypes where each locus has more than two alleles and, if possible, where the number of genes is much greater than 10, as used in this study.

Finally, while this research analyzed the properties of a novel network model, it did not analyze how populations would evolve across this landscape. Allowing genotypes to accumulate mutations and navigate landscapes produced by the model proposed here would provide valuable

insights into the mechanisms of evolution and speciation. It would contribute to a better comprehension of how different levels of ruggedness as well as different fitness thresholds influence these mechanisms. Additionally, research of this nature would complement the conclusions of this thesis and lend itself to an even better understanding of the novel landscapes analyzed in this study.

APPENDICES

6.1 Average Network Data

k = 0

Threshold	Big Component Inclusion	Big Component Size	Component Size CV	Components	Dimensionality	Edges	Hamming Distance	Mean Component Size	Median Component Size	Small Component Size	Vertices
0.20	1.0	1024.0	0.0	1.0	5.000000	5120.0	5.004888	1024.000000	1024.000000	1024.0	1024.0
0.25	1.0	1024.0	0.0	1.0	5.000000	5120.0	5.004888	1024.000000	1024.000000	1024.0	1024.0
0.30	1.0	1024.0	0.0	1.0	5.000000	5120.0	5.004888	1024.000000	1024.000000	1024.0	1024.0
0.35	1.0	1023.0	0.0	1.0	4.996450	5111.4	5.004861	1023.000000	1023.000000	1023.0	1023.0
0.40	1.0	1012.9	0.0	1.0	4.973384	5040.2	5.002817	1012.900000	1012.900000	1012.9	1012.9
0.45	1.0	984.0	0.0	1.0	4.919520	4866.2	4.985742	984.000000	984.000000	984.0	984.0
0.50	1.0	915.9	0.0	1.0	4.765457	4455.2	4.927821	915.900000	915.900000	915.9	915.9
0.55	1.0	780.4	0.0	1.0	4.431393	3654.8	4.769436	780.400000	780.400000	780.4	780.4
0.60	1.0	573.5	0.0	1.0	3.782708	2523.6	4.711758	573.500000	573.500000	573.5	573.5
0.65	1.0	348.5	0.0	0.9	3.481498	1431.4	4.117736	387.222222	387.222222	348.5	348.5
0.70	1.0	184.1	0.0	0.8	2.675197	701.8	3.798962	230.125000	230.125000	184.1	184.1
0.75	1.0	81.1	0.0	0.6	2.499811	271.0	3.224399	135.166667	135.166667	81.1	81.1
0.80	1.0	24.0	0.0	0.4	1.611135	64.5	2.313441	60.000000	60.000000	24.0	24.0
0.85	1.0	3.8	0.0	0.2	1.772727	6.6	2.283117	19.000000	19.000000	3.8	3.8
0.90	1.0	0.2	0.0	0.1	0.500000	0.1	1.000000	2.000000	2.000000	0.2	0.2
0.95	NaN	0.0	NaN	0.0	NaN	0.0	NaN	NaN	NaN	0.0	0.0

k = 1

Threshold	Big Component Inclusion	Big Component Size	Component Size CV	Components	Dimensionality	Edges	Hamming Distance	Mean Component Size	Median Component Size	Small Component Size	Vertices
0.20	1.000000	1024.0	0.000000	1.0	5.000000	5120.0	5.004888	1024.000000	1024.000000	1024.0	1024.0
0.25	1.000000	1024.0	0.000000	1.0	5.000000	5120.0	5.004888	1024.000000	1024.000000	1024.0	1024.0
0.30	1.000000	1023.1	0.000000	1.0	4.996268	5111.7	5.004875	1023.100000	1023.100000	1023.1	1023.1
0.35	1.000000	1017.0	0.000000	1.0	4.975726	5060.8	5.004525	1017.000000	1017.000000	1017.0	1017.0
0.40	1.000000	993.4	0.000000	1.0	4.909118	4880.4	5.001596	993.400000	993.400000	993.4	993.4
0.45	1.000000	936.5	0.000000	1.0	4.769637	4486.5	4.987127	936.500000	936.500000	936.5	936.5
0.50	1.000000	832.2	0.000000	1.0	4.527940	3815.5	4.946788	832.200000	832.200000	832.2	832.2
0.55	1.000000	672.8	0.000000	1.0	4.160749	2885.9	4.849998	672.800000	672.800000	672.8	672.8
0.60	1.000000	486.2	0.000000	1.0	3.688668	1901.4	4.681234	486.200000	486.200000	486.2	486.2
0.65	1.000000	293.3	0.000000	1.0	3.045409	1007.8	4.374846	293.300000	293.300000	293.3	293.3
0.70	0.950000	144.6	0.035355	1.2	2.268112	428.6	3.727025	144.333333	144.200000	144.2	145.4
0.75	0.944444	61.2	0.000000	1.0	1.728000	150.6	3.232176	68.000000	68.000000	61.2	61.3
0.80	0.904762	19.8	0.067686	1.0	1.376469	43.1	2.707319	28.035714	28.214286	19.2	21.2
0.85	0.812500	5.5	0.076547	1.0	1.041734	9.2	2.739194	13.535714	13.500000	5.4	6.1
0.90	1.000000	1.2	0.000000	0.2	1.125000	1.4	1.613095	6.000000	6.000000	1.2	1.2
0.95	0.500000	0.1	0.000000	0.2	0.000000	0.0	2.000000	1.000000	1.000000	0.1	0.2

k = 2

Threshold	Big Component Inclusion	Big Component Size	Component Size CV	Components	Dimensionality	Edges	Hamming Distance	Mean Component Size	Median Component Size	Small Component Size	Vertices
0.20	1.000000	1024.0	0.000000	1.0	5.000000	5120.0	5.004888	1024.000000	1024.000000	1024.0	1024.0
0.25	1.000000	1024.0	0.000000	1.0	5.000000	5120.0	5.004888	1024.000000	1024.000000	1024.0	1024.0
0.30	1.000000	1024.0	0.000000	1.0	5.000000	5120.0	5.004888	1024.000000	1024.000000	1024.0	1024.0
0.35	1.000000	1022.2	0.000000	1.0	4.991865	5102.7	5.004874	1022.200000	1022.200000	1022.2	1022.2
0.40	1.000000	1010.7	0.000000	1.0	4.949641	5002.9	5.004463	1010.700000	1010.700000	1010.7	1010.7
0.45	1.000000	971.4	0.000000	1.0	4.825392	4689.2	5.001021	971.400000	971.400000	971.4	971.4
0.50	1.000000	868.3	0.000000	1.0	4.546091	3951.4	4.983422	868.300000	868.300000	868.3	868.3
0.55	1.000000	695.6	0.000000	1.0	4.123206	2874.7	4.932035	695.600000	695.600000	695.6	695.6
0.60	1.000000	477.2	0.000000	1.0	3.552575	1707.3	4.825387	477.200000	477.200000	477.2	477.2
0.65	0.999184	258.2	0.139690	1.2	2.849960	751.0	4.609099	242.066667	234.000000	234.0	258.4
0.70	0.976956	111.5	0.153911	1.2	2.172882	259.1	4.234776	106.800000	106.800000	102.1	112.8
0.75	0.892389	37.5	0.468928	1.8	1.510277	70.8	3.658659	29.416667	27.850000	22.9	40.7
0.80	0.808642	8.0	0.368359	1.7	0.920576	12.0	2.907473	7.018519	6.611111	5.0	9.1
0.85	1.000000	1.2	0.000000	0.4	0.250000	0.9	2.111111	3.000000	3.000000	1.2	1.2
0.90	NaN	0.0	NaN	0.0	NaN	0.0	NaN	NaN	NaN	0.0	0.0
0.95	NaN	0.0	NaN	0.0	NaN	0.0	NaN	NaN	NaN	0.0	0.0

k = 3

Threshold	Big Component Inclusion	Big Component Size	Component Size CV	Components	Dimensionality	Edges	Hamming Distance	Mean Component Size	Median Component Size	Small Component Size	Vertices
0.20	1.000000	1024.0	0.000000	1.0	5.000000	5120.0	5.004888	1024.000000	1024.000000	1024.0	1024.0
0.25	1.000000	1024.0	0.000000	1.0	5.000000	5120.0	5.004888	1024.000000	1024.000000	1024.0	1024.0
0.30	1.000000	1023.8	0.000000	1.0	4.999022	5118.0	5.004888	1023.800000	1023.800000	1023.8	1023.8
0.35	1.000000	1022.0	0.000000	1.0	4.990891	5100.7	5.004877	1022.000000	1022.000000	1022.0	1022.0
0.40	1.000000	1011.6	0.000000	1.0	4.949373	5007.0	5.004620	1011.600000	1011.600000	1011.6	1011.6
0.45	1.000000	975.2	0.000000	1.0	4.819976	4701.9	5.003130	975.200000	975.200000	975.2	975.2
0.50	1.000000	877.1	0.000000	1.0	4.510620	3961.7	4.996213	877.100000	877.100000	877.1	877.1
0.55	1.000000	701.3	0.000000	1.0	4.010354	2821.8	4.971781	701.300000	701.300000	701.3	701.3
0.60	1.000000	480.8	0.000000	1.0	3.376333	1634.5	4.911998	480.800000	480.800000	480.8	480.8
0.65	0.998555	259.6	0.297111	1.3	2.631810	695.6	4.777582	217.050000	217.050000	174.5	260.0
0.70	0.981409	114.1	0.519302	1.9	1.924579	228.2	4.531615	91.016667	85.350000	75.9	115.5
0.75	0.769491	30.9	1.081936	4.3	1.155758	47.2	4.241830	14.285714	9.600000	6.5	37.7
0.80	0.704532	6.9	0.434034	3.2	0.585122	8.2	3.516937	4.360000	4.100000	2.7	10.4
0.85	0.616667	0.6	0.066667	0.7	0.366667	0.4	2.655556	1.833333	1.833333	0.5	1.1
0.90	NaN	0.0	NaN	0.0	NaN	0.0	NaN	NaN	NaN	0.0	0.0
0.95	NaN	0.0	NaN	0.0	NaN	0.0	NaN	NaN	NaN	0.0	0.0

k = 4

Threshold	Big Component Inclusion	Big Component Size	Component Size CV	Components	Dimensionality	Edges	Hamming Distance	Mean Component Size	Median Component Size	Small Component Size	Vertices
0.20	1.000000	1024.0	0.000000	1.0	5.000000	5120.0	5.004888	1024.000000	1024.0000	1024.0	1024.0
0.25	1.000000	1023.9	0.000000	1.0	4.999511	5119.0	5.004888	1023.900000	1023.9000	1023.9	1023.9
0.30	1.000000	1023.9	0.000000	1.0	4.999511	5119.0	5.004888	1023.900000	1023.9000	1023.9	1023.9
0.35	1.000000	1021.2	0.000000	1.0	4.987124	5092.9	5.004874	1021.200000	1021.2000	1021.2	1021.2
0.40	1.000000	1008.9	0.000000	1.0	4.937406	4981.7	5.004670	1008.900000	1008.9000	1008.9	1008.9
0.45	1.000000	968.6	0.000000	1.0	4.782756	4634.5	5.003388	968.600000	968.6000	968.6	968.6
0.50	1.000000	872.5	0.000000	1.0	4.456140	3894.9	4.997490	872.500000	872.5000	872.5	872.5
0.55	1.000000	708.0	0.000000	1.0	3.942760	2805.1	4.981197	708.000000	708.0000	708.0	708.0
0.60	1.000000	498.3	0.000000	1.0	3.258393	1643.4	4.944823	498.300000	498.3000	498.3	498.3
0.65	0.995909	286.3	0.474270	1.6	2.473227	728.9	4.875869	230.566667	214.8000	190.6	287.3
0.70	0.906448	122.9	1.475756	5.1	1.643809	230.3	4.766294	60.764206	42.2500	42.1	131.3
0.75	0.652097	34.2	1.429757	7.7	0.987814	49.8	4.576687	8.385284	1.4500	1.0	46.2
0.80	0.531396	4.9	0.569224	4.7	0.500333	6.4	4.105322	2.269444	1.6500	1.1	10.4
0.85	0.564583	1.1	0.129162	2.2	0.097917	0.3	4.050000	1.135417	1.0625	0.8	2.5
0.90	1.000000	0.1	0.000000	0.1	0.000000	0.0	NaN	1.000000	1.0000	0.1	0.1
0.95	NaN	0.0	NaN	0.0	NaN	0.0	NaN	NaN	NaN	0.0	0.0

k = 5

Threshold	Big Component Inclusion	Big Component Size	Component Size CV	Components	Dimensionality	Edges	Hamming Distance	Mean Component Size	Median Component Size	Small Component Size	Vertices
0.20	1.000000	1024.0	0.000000	1.0	5.000000	5120.0	5.004888	1024.000000	1024.00	1024.0	1024.0
0.25	1.000000	1024.0	0.000000	1.0	5.000000	5120.0	5.004888	1024.000000	1024.00	1024.0	1024.0
0.30	1.000000	1023.8	0.000000	1.0	4.999022	5118.0	5.004887	1023.800000	1023.80	1023.8	1023.8
0.35	1.000000	1022.3	0.000000	1.0	4.991863	5103.2	5.004886	1022.300000	1022.30	1022.3	1022.3
0.40	1.000000	1011.2	0.000000	1.0	4.942440	4998.2	5.004796	1011.200000	1011.20	1011.2	1011.2
0.45	1.000000	974.7	0.000000	1.0	4.793694	4674.5	5.004163	974.700000	974.70	974.7	974.7
0.50	1.000000	886.7	0.000000	1.0	4.461927	3961.6	5.001817	886.700000	886.70	886.7	886.7
0.55	1.000000	726.2	0.000000	1.0	3.889895	2835.0	4.994017	726.200000	726.20	726.2	726.2
0.60	1.000000	512.2	0.000000	1.0	3.143691	1623.9	4.973576	512.200000	512.20	512.2	512.2
0.65	0.997330	296.1	0.508846	1.7	2.328939	704.1	4.927915	233.741667	218.50	193.2	296.8
0.70	0.964357	140.5	1.466082	4.3	1.559374	231.1	4.861748	61.466389	42.75	34.8	145.2
0.75	0.492324	26.9	1.571014	11.6	0.887854	46.5	4.707079	4.734006	1.40	1.0	50.5
0.80	0.268275	3.7	0.532330	8.8	0.265566	4.4	4.623758	1.428698	1.00	1.0	13.0
0.85	0.635714	0.7	0.000000	1.6	0.000000	0.0	4.783333	1.000000	1.00	0.7	1.6
0.90	0.750000	0.2	0.000000	0.3	0.000000	0.0	2.000000	1.000000	1.00	0.2	0.3
0.95	NaN	0.0	NaN	0.0	NaN	0.0	NaN	NaN	NaN	0.0	0.0

k = 6

Threshold	Big Component Inclusion	Big Component Size	Component Size CV	Components	Dimensionality	Edges	Hamming Distance	Mean Component Size	Median Component Size	Small Component Size	Vertices
0.20	1.000000	1024.0	0.000000	1.0	5.000000	5120.0	5.004888	1024.000000	1024.00	1024.0	1024.0
0.25	1.000000	1024.0	0.000000	1.0	5.000000	5120.0	5.004888	1024.000000	1024.00	1024.0	1024.0
0.30	1.000000	1023.9	0.000000	1.0	4.999511	5119.0	5.004888	1023.900000	1023.90	1023.9	1023.9
0.35	1.000000	1022.0	0.000000	1.0	4.990597	5100.4	5.004883	1022.000000	1022.00	1022.0	1022.0
0.40	1.000000	1011.9	0.000000	1.0	4.943278	5002.2	5.004863	1011.900000	1011.90	1011.9	1011.9
0.45	1.000000	975.5	0.000000	1.0	4.781921	4665.1	5.004627	975.500000	975.50	975.5	975.5
0.50	1.000000	880.7	0.000000	1.0	4.391285	3869.8	5.003497	880.700000	880.70	880.7	880.7
0.55	1.000000	728.3	0.000000	1.0	3.807564	2776.8	4.999911	728.300000	728.30	728.3	728.3
0.60	1.000000	518.6	0.000000	1.0	2.986982	1553.7	4.992756	518.600000	518.60	518.6	518.6
0.65	0.996115	298.3	0.720632	2.1	2.079054	625.5	4.971085	203.225000	170.10	154.6	299.4
0.70	0.854746	124.5	2.402858	10.2	1.287879	185.5	4.944192	16.663232	1.35	1.0	141.6
0.75	0.245855	13.5	1.146009	19.9	0.633320	33.6	4.937618	2.658192	1.20	1.0	51.2
0.80	0.200497	2.4	0.341208	9.8	0.187095	2.6	4.940612	1.269625	1.00	1.0	12.4
0.85	0.537037	0.9	0.000000	1.8	0.000000	0.0	4.750000	1.000000	1.00	0.9	1.8
0.90	1.000000	0.2	0.000000	0.2	0.000000	0.0	NaN	1.000000	1.00	0.2	0.2
0.95	NaN	0.0	NaN	0.0	NaN	0.0	NaN	NaN	NaN	0.0	0.0

k = 7

Threshold	Big Component Inclusion	Big Component Size	Component Size CV	Components	Dimensionality	Edges	Hamming Distance	Mean Component Size	Median Component Size	Small Component Size	Vertices
0.20	1.000000	1024.0	0.000000	1.0	5.000000	5120.0	5.004888	1024.000000	1024.00	1024.0	1024.0
0.25	1.000000	1024.0	0.000000	1.0	5.000000	5120.0	5.004888	1024.000000	1024.00	1024.0	1024.0
0.30	1.000000	1023.9	0.000000	1.0	4.999511	5119.0	5.004888	1023.900000	1023.90	1023.9	1023.9
0.35	1.000000	1021.6	0.000000	1.0	4.988346	5096.1	5.004886	1021.600000	1021.60	1021.6	1021.6
0.40	1.000000	1008.8	0.000000	1.0	4.928656	4972.1	5.004867	1008.800000	1008.80	1008.8	1008.8
0.45	1.000000	973.3	0.000000	1.0	4.764978	4638.2	5.004738	973.300000	973.30	973.3	973.3
0.50	1.000000	879.2	0.000000	1.0	4.358342	3832.3	5.003967	879.200000	879.20	879.2	879.2
0.55	1.000000	714.4	0.000000	1.0	3.672786	2625.4	5.001619	714.400000	714.40	714.4	714.4
0.60	1.000000	503.9	0.000000	1.0	2.788454	1407.0	4.997770	503.900000	503.90	503.9	503.9
0.65	0.990546	297.0	1.394607	3.7	1.870918	563.5	4.988186	128.483333	78.60	64.0	299.7
0.70	0.748401	108.3	2.909725	17.4	1.067333	153.2	4.979244	9.008877	1.05	1.0	142.0
0.75	0.186063	10.0	1.017701	26.2	0.527762	28.2	4.938981	2.044667	1.00	1.0	52.7
0.80	0.196936	2.7	0.416912	10.5	0.210609	3.1	4.918600	1.288822	1.00	1.0	13.6
0.85	0.452381	0.9	0.098127	2.2	0.059524	0.2	5.316667	1.076190	1.00	0.7	2.4
0.90	1.000000	0.2	0.000000	0.2	0.000000	0.0	NaN	1.000000	1.00	0.2	0.2
0.95	NaN	0.0	NaN	0.0	NaN	0.0	NaN	NaN	NaN	0.0	0.0

k = 8

Threshold	Big Component Inclusion	Big Component Size	Component Size CV	Components	Dimensionality	Edges	Hamming Distance	Mean Component Size	Median Component Size	Small Component Size	Vertices
0.20	1.000000	1024.0	0.000000	1.0	5.000000	5120.0	5.004888	1024.000000	1024.00	1024.0	1024.0
0.25	1.000000	1024.0	0.000000	1.0	5.000000	5120.0	5.004888	1024.000000	1024.00	1024.0	1024.0
0.30	1.000000	1023.5	0.000000	1.0	4.997555	5115.0	5.004888	1023.500000	1023.50	1023.5	1023.5
0.35	1.000000	1020.2	0.000000	1.0	4.981560	5082.2	5.004885	1020.200000	1020.20	1020.2	1020.2
0.40	1.000000	1010.0	0.000000	1.0	4.931622	4981.0	5.004878	1010.000000	1010.00	1010.0	1010.0
0.45	1.000000	971.8	0.000000	1.0	4.751340	4617.6	5.004786	971.800000	971.80	971.8	971.8
0.50	1.000000	881.0	0.000000	1.0	4.330824	3815.8	5.004654	881.000000	881.00	881.0	881.0
0.55	1.000000	708.9	0.000000	1.0	3.547780	2515.7	5.004324	708.900000	708.90	708.9	708.9
0.60	0.999413	508.1	0.298827	1.3	2.652482	1350.4	5.001079	431.650000	431.65	355.2	508.4
0.65	0.977140	290.7	2.353526	7.0	1.638517	488.6	4.998945	46.209048	1.05	1.0	297.4
0.70	0.526155	73.2	2.750383	30.4	0.861897	119.2	4.992534	4.775339	1.20	1.0	137.4
0.75	0.106946	5.6	0.700086	34.1	0.352798	18.7	4.973739	1.546778	1.00	1.0	52.6
0.80	0.139407	1.9	0.253326	12.6	0.102759	1.5	4.854442	1.123042	1.00	1.0	14.1
0.85	0.531481	0.9	0.000000	2.4	0.000000	0.0	4.966667	1.000000	1.00	0.9	2.4
0.90	0.666667	0.2	0.000000	0.4	0.000000	0.0	4.666667	1.000000	1.00	0.2	0.4
0.95	NaN	0.0	NaN	0.0	NaN	0.0	NaN	NaN	NaN	0.0	0.0

k = 9

Threshold	Big Component Inclusion	Big Component Size	Component Size CV	Components	Dimensionality	Edges	Hamming Distance	Mean Component Size	Median Component Size	Small Component Size	Vertices
0.20	1.000000	1024.0	0.000000	1.0	5.000000	5120.0	5.004888	1024.000000	1024.00	1024.0	1024.0
0.25	1.000000	1024.0	0.000000	1.0	5.000000	5120.0	5.004888	1024.000000	1024.00	1024.0	1024.0
0.30	1.000000	1023.8	0.000000	1.0	4.999022	5118.0	5.004888	1023.800000	1023.80	1023.8	1023.8
0.35	1.000000	1021.4	0.000000	1.0	4.987263	5094.0	5.004889	1021.400000	1021.40	1021.4	1021.4
0.40	1.000000	1011.0	0.000000	1.0	4.937252	4991.6	5.004886	1011.000000	1011.00	1011.0	1011.0
0.45	1.000000	975.1	0.000000	1.0	4.762238	4643.8	5.004896	975.100000	975.10	975.1	975.1
0.50	1.000000	886.0	0.000000	1.0	4.323308	3831.0	5.004923	886.000000	886.00	886.0	886.0
0.55	1.000000	729.9	0.000000	1.0	3.557157	2597.0	5.005272	729.900000	729.90	729.9	729.9
0.60	0.998235	515.1	0.837844	1.9	2.526910	1304.7	5.004403	303.383333	286.55	108.5	516.0
0.65	0.960977	289.8	2.879520	10.3	1.465448	442.7	5.005165	31.756663	1.00	1.0	301.4
0.70	0.270270	39.3	1.993003	50.6	0.669985	96.0	5.010376	2.914851	1.00	1.0	142.4
0.75	0.099769	5.5	0.651725	40.7	0.263186	14.5	5.023017	1.351790	1.00	1.0	54.9
0.80	0.117581	1.8	0.214417	14.5	0.074303	1.2	4.982533	1.083184	1.00	1.0	15.7
0.85	0.550000	0.8	0.000000	2.4	0.000000	0.0	4.653333	1.000000	1.00	0.8	2.4
0.90	NaN	0.0	NaN	0.0	NaN	0.0	NaN	NaN	NaN	0.0	0.0
0.95	NaN	0.0	NaN	0.0	NaN	0.0	NaN	NaN	NaN	0.0	0.0

BIBLIOGRAPHY

Bazykin AD. 1969. Hypothetical Mechanism of Speciation. *Evolution* 23: 685–687.

Dickinson H, Antonovics J. 1973. Theoretical Considerations of Sympatric Divergence. *American Naturalist* 107: 256–274.

Dobzhansky TG. 1936. Studies on Hybrid Sterility. II. Localization of Sterility Factors in *Drosophila Pseudoobscura* hybrids. *Genetics* 21: 113–135.

Gavrilets S. 2004. Fitness Landscapes and the Origin of Species. Princeton (NJ): Princeton University Press.

Gravner J, Pitman D, Gavrilets S. 2007. Percolation on Fitness Landscapes: Effects of Correlation, Phenotype, and Incompatibilities. *Journal of Theoretical Biology* 248: 627–645.

Kauffman SA, Weinberger ED. 1989. The *NK* Model of Rugged Fitness Landscapes and its Application to Maturation of the Immune Response. *Journal of Theoretical Biology* 141: 211–245.

Lee Y, DSouza LM, Fox GE. 1997. Equally Parsimonious Pathways Through an RNA Sequence Space Are Not Equally Likely. *Journal of Molecular Evolution* 45: 278–284.

Maheshwari S, Barbash DA. 2011 The Genetics of Hybrid Incompatibilities. *Annual Review of Genetics* 45: 331–355.

Newman MEJ, Engelhardt R. 1998. Effects of Selective Neutrality on the Evolution of Molecular Species. *Proceedings of the Royal Society of London* 265: 1333–1338.

Orr HA. 1996. Perspectives Anecdotal, Historical and Critical Commentaries on Genetics. *Genetics* 144: 1331–1335.

Presgraves DC. 2010. The Molecular Evolutionary Basis of Species Formation. *Nature Reviews: Genetics* 11: 175–180.

Schuster P, Fontana W, Stadler P, Hofacker I. 1994. From Sequences to Shapes and Back: A Case Study in RNA Secondary Structures. *Proceedings of the Royal Society of London* 255: 279–284.

Servedio MR, Brandvain Y, Dhole S, Fitzpatrick CL, Goldberg EE, Stern CA, Cleve JV, Yeh DJ. 2014. Not Just a Theory – The Utility of Mathematical Models in Evolutionary Biology. *PLOS Biology* 12(12): 1–5.

Smith JM. 1962. Disruptive Selection, Polymorphism and Sympatric Speciation. *Nature* 195: 60–62.

Szendro IG, Schenk MF, Franke J, Krug J, Arjan GM de Visser. 2013. Quantitative Analyses of Empirical Fitness Landscapes. *Journal of Statistical Mechanics: Theory and Expreriment* 2013: P01005

Weinreich DM, Lan Y, Wylie CS, Heckendorn RB. 2013. Should Evolutionary Geneticists Worry About Higher-Order Epistasis. *Current Opinion in Genetics and Development* 23: 700–707.

Wright S. 1932. The Roles of Mutation, Inbreeding, Crossbreeding and Selection in Evolution. *Proceedings of the Sixth International Congress on Genetics* 1: 356–366.