UNDERSTANDING THE DISTRIBUTION OF FITNESS EFFECTS ACROSS ENVIRONMENTS: EXPLORING THE STATISTICAL PROPERTIES OF FITNESS LANDSCAPES USING FISHER'S GEOMETRIC MODEL

Shahnewaz Ahmed
Department of Physics and Astronomy
Rutgers University
sa1951@rutgers.edu

Michael Manhart
Center for Advanced Biotechnology and Medicine
Rutgers University
mmanhart@rutgers.edu

April 24, 2025

ABSTRACT

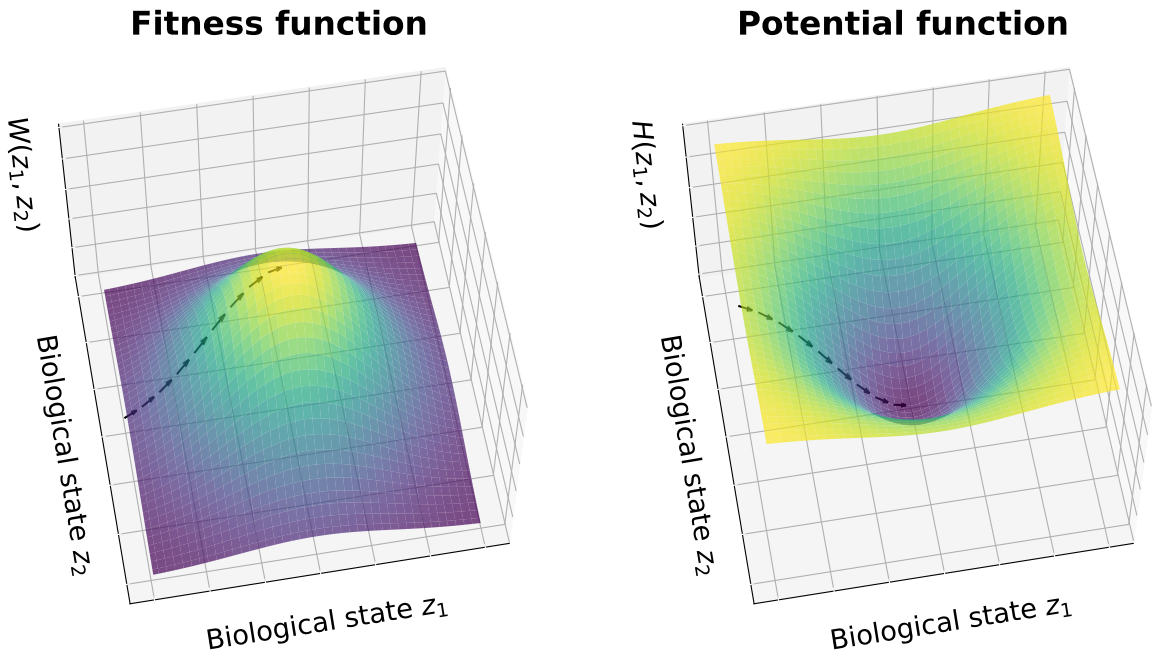
The distribution of fitness effects (DFE) is a fundamental concept in evolutionary biology; understanding the DFE is important for analyzing a variety of phenomena, including quantitative traits, complex diseases, and the evolution of antibiotic resistance. The DFE is the set of fitness effects from all spontaneous mutations accessible to an organism's genome in a specific environment. However, since environments in nature are constantly changing, understanding how the DFE varies across different environments is crucial for predicting evolutionary outcomes. In this study, we employ the Fisher Geometric Model (FGM) to explore the statistical properties of DFEs and their dependence on environmental changes. The FGM provides a theoretical framework to model fitness landscapes, where mutations are represented as random perturbations in a high-dimensional phenotypic space, and fitness is determined by the distance to an optimal phenotype.

We compare the predictions of the FGM with empirical data from a large-scale study of ~ 3800 gene knockout mutants of *E. coli* across ~ 100 environments. These environments include both stress conditions (e.g., antibiotics, metals, salts) and non-stress conditions (varying carbon and nitrogen sources). By simulating DFEs under various environmental conditions, we explore whether FGM can reproduce key statistical features observed in experimental data, such as the distributions of fitness means, variances, and correlation coefficients across different environments. Our results indicate that, while the FGM successfully captures certain aspects of the data, it falls short of fully explaining the observed patterns—particularly the detailed shape of the distribution of correlation coefficients between environments. This study highlights the limitations of the FGM in describing empirical datasets and emphasizes the need for alternative or extended models to better capture the complexity of biological adaptation.

1 Introduction

Scientists have long sought to unravel the inner workings of living organisms. In earlier decades, they made significant breakthroughs by dissecting biological systems into smaller pieces, focusing on individual elements such as genes and molecules. Today, we stand at the threshold of a new *phase—one* that demands a grasp of life's full complexity. To achieve this, researchers must gather and structure enormous amounts of biological data in a consistent format, create more effective tools to retrieve and link this information, and employ advanced mathematical models to interpret it all.

One promising approach to modeling biological complexity draws inspiration from principles in physics. The concept of energy minimization, fundamental to physical systems, also finds a parallel in biological adaptation, albeit with greater complexity. In physics, objects minimize potential energy as a function of degrees of freedom (DOFs), such as a



(a) (b)

Figure 1: Evolution is often visualized as populations navigating a landscape, represented in the picture with the black arrow lines in the figures. The same process can be represented in two equivalent ways: either 1a as a movement toward peaks on a fitness landscape or 1b as a descent toward valleys on a landscape of potential function.

ball rolling downhill or molecules settling into low-energy configurations. Similarly, biological populations evolve toward states that optimize fitness, often visualized as movement across fitness landscapes toward peaks of maximum adaptation. By reinterpreting this landscape as a "fitness potential," where evolutionary trajectories move "downhill," we preserve the intuitive metaphor while adhering to Darwinian principles, as illustrated in Figure 1. This framework helps bridge the conceptual gap between physics and biology, though it must be noted that biological systems introduce unique challenges due to their multidimensionality, lack of symmetry, and complexity across multiple scales [1, 2]. Importantly, while this analogy offers valuable insights into optimization dynamics, it should not be mistaken for a deeper biological law.

Biological adaptation operates in high-dimensional spaces shaped by countless biological and ecological interactions. Even simple cells contain thousands of proteins, each with mutable amino acids, creating a vast configuration space. These landscapes are rugged, with kinetic traps that can slow adaptation, but must also contain smooth pathways to enable evolutionary progress [3, 4]. Unlike physical systems, biological landscapes are densely packed with functional regions: minor mutations can unlock new capabilities, and promiscuous proteins allow the exploration of novel functions without losing existing ones. Such features enable both gradual refinement and rapid innovation, depending on the evolutionary context [5].

Adaptation dynamics shifts as systems evolve. Early stages often involve large fitness gains from single mutations, propelling populations up steep gradients. Later stages become more complex, relying on epistatic (presence of one gene modifies the expression of another gene) interactions between multiple mutations, where the order and combination of changes determine outcomes [6]. Stochastic fluctuations add another layer of complexity, sometimes driving the exploration of new states and other times opposing adaptive trends [7]. Researchers have tried to define some kind of simplified biological fitness landscape that might have mathematically tractable properties. Capturing these processes requires careful modeling choices, from selecting the right scale (molecular, cellular, or population) to defining relevant degrees of freedom and fitness metrics. Since biological optimization functions are rarely known, researchers often take a hypothesis-driven approach: proposing plausible landscape features, deriving predictions, and testing them against experimental data.

This project aims to utilize one of the earliest and simplest biological landscape models to understand how the distribution of fitness effects (DFE) of new mutations changes across different environmental conditions. DFE is a

fundamental concept in evolutionary biology, characterizing the spectrum of fitness consequences arising from new mutations within a population. It offers a localized perspective on the evolutionary landscape, detailing the likelihood of different mutations becoming fixed within a population over time based on their impact on fitness. Although DFE provides a localized view of how mutations impact an organism's fitness within a specific environment, natural environments are dynamic. Therefore, to make predictions about evolutionary trajectories in the real world, it is crucial to develop or employ a simple mathematical model in order to investigate how the characteristics of the DFE, such as its mean, variance, and overall shape, are modulated by environmental change and *qualitatively* compare with the experimental data.

Fisher's geometric model (FGM) seems to be a natural choice for this. In the 1930s, Robert A. Fisher, one of the three founding fathers of population genetics, proposed a geometric model of adaptation in his seminal book "The Genetical Theory of Natural Selection" [8]. Although there is some empirical support for the model's prediction [9, 10], this model is most successful at predicting the effect of epistasis on the distribution of fitness effects [11, 12]. However, the validity of this model has not being tested against a large mutational database subjected to different growth environments. Some studies did partake in observing some quantitative results from the model and comparing them with the experimental data [13, 14], but those are limited to a handful of environments over different species. In this work, we investigated the prediction of this model with empirical data from a large-scale study of 3789 single-gene knockout mutants of *E.coli* subjected to 101 different environments [15].

2 Description and Analysis of Experimental Data

Among the most thoroughly researched organisms is the bacterium *E. coli* K-12. Its entire genetic sequence has been mapped with remarkable precision, possibly the most detailed for any organism of its size [16]. This high-resolution blueprint has enabled scientists to revisit and clarify the roles of most of its genes. Given how much we already know about E. coli, and because it's a relatively simple single-celled organism, it serves as the perfect candidate for achieving a comprehensive understanding of a living system. The notion of fully deciphering E. coli isn't recent—it was even suggested as "Project K" in 1973. Back then, however, critical technologies like high-speed computing and the internet didn't yet exist [17].

Today, high-throughput omics technologies—such as DNA sequencing (genomics), RNA sequencing (transcriptomics), and mass spectrometry (proteomics)—are revolutionizing biological and clinical research [18]. In a recent large-scale study, researchers employed transposon mutagenesis (a technique where mobile genetic elements disrupt protein-coding genes) to generate mutant libraries for 32 diverse bacterial species [15]. Although transposons can change their location within the genome during initial library creation, they do not move independently after the initial library creation.

These mutants were then grown under hundreds of controlled conditions (e.g., varying nutrient availability, stressors, or environmental perturbations). Researchers first created a pool of mutant strains using randomly barcoded transposons (RB-TnSeq). Each transposon carries a unique DNA barcode and inserts randomly into the genome, allowing thousands to hundreds of thousands of mutant strains to be tracked in a particular experiment. After the mutagenesis the pool of mutant samples was recovered from the freezer in nutrient-rich media, then the cells were typically washed and sampled to establish a baseline to note the abundance at the beginning. The remaining cells were then grown under different experimental conditions. By comparing the abundance of each DNA barcode before and after the experiment, researchers determined which mutants thrived and which did not [19].

The change in abundance of mutants associated with a gene is expressed as a $\log_2$ ratio, which is the basis for the gene's fitness value. A fitness value of 0 suggests that mutants in that gene grew at rates similar to the overall population. A negative fitness value indicates that mutants were less abundant at the end, implying that the gene is important for growth or survival under those conditions. In contrast, a positive value suggests that the gene may hinder growth. To calculate gene-level fitness, the strain-level fitness values—each derived from the change in read counts of the corresponding barcodes—are averaged with a weighting system that accounts for variability in read depth. The data are also normalized so that the typical gene has a fitness value close to zero, and adjustments are made for genomic copy number variation.

$$\text{Mutant fitness} = \log_2\left(\frac{\text{abundance after}}{\text{abundance at the start}}\right) \tag{1}$$

In order to facilitate further analyses of these mutant phenotypes and protein sequences, researchers developed a website called fitness browser. From that website, we collected fitness data of *E. coli* single knockout gene mutation for 3789 genes, each exposed to 101 environments, where 54 environments are labeled as stress environments and 47 environments are identified to be nutrient-rich non-stress environments. Stress environments include different antibiotic solutions, for example, Tetracycline 0.0004 mg/ml, Bacitracin 0.5 mg/ml, Spectinomycin 0.0125 mg/ml,

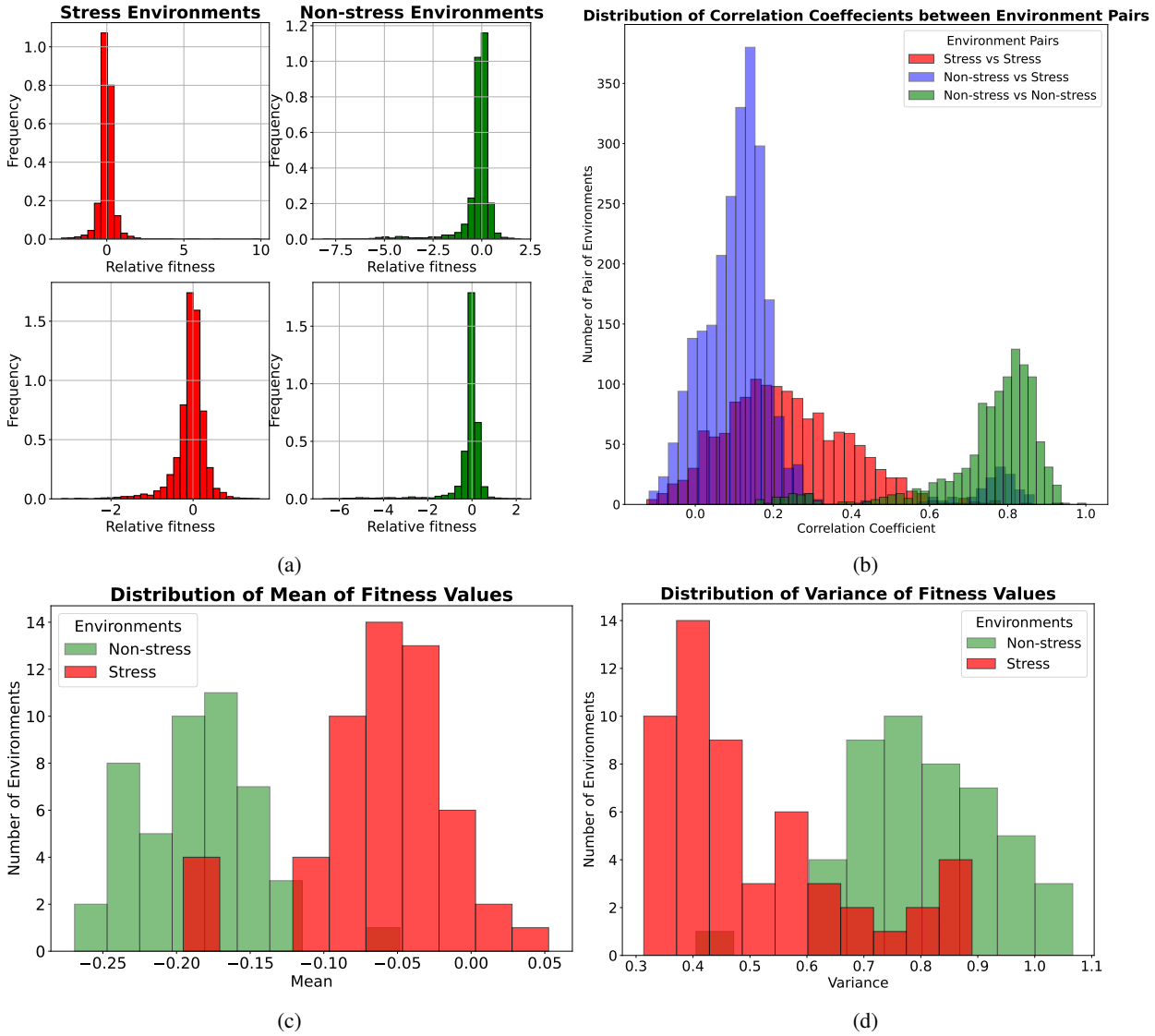


Figure 2: The four panels in Figure 2a illustrate sample Distributions of Fitness Effects (DFEs) for both stress and non-stress environments. These DFEs are plotted using the dataset referenced in the text. The distributions generally resemble a skewed Gaussian shape, highlighting the importance of analyzing key statistical parameters such as the mean, variance, and pairwise correlation coefficients. Figure 2b displays the correlation coefficient between different environment pairs. There are three possible pairings, which are highlighted separately. Figure 2c and 2d show the distribution of mean and variance, respectively, for stress and non-stress environments.

etc.; chemical solutions like Aluminum chloride 5 *mM*, Cobalt chloride 0.32 *mM*, Acetate 70 *mM*, etc. On the other hand, Carbon-rich environments like D-Glucose, L-Fucose, Glycerol etc, Nitrogen-rich environments D-Alanine, L-Asparagine, Cytidine, etc., are considered Non-stress environments in this study. These environments are not necessarily some absolute source of Carbon and Nitrogen; rather they vary the specific compound serving as the Carbon or Nitrogen source without additional stressors.

In Figure 2, several statistical behaviors of these fitness datasets are displayed concerning the environments. The figure 2a shows how the gene relative fitness would create different distributions in stress or non-stress environments. It is clear from these four figures that these distributions are highly centered around the relative fitness value of 0 and are almost evenly spread around that value. This suggests that the mean fitness values may be small for all of these environments. In contrast, stress environments exhibit small negative or even positive mean fitness values, as shown in Figure 2a. However, an opposite trend is observed in the variance: fitness values under stress conditions tend to have significantly lower variance—often close to zero—indicating that most mutations have similarly small effects on fitness in such environments.

Similarly, a strong pattern is apparent in the distribution of the correlation coefficients between the environment pairs displayed in the figure 2b. There is a strong correlation between non-stress environments, but stress environments exhibit a small correlation with each other and with the non-stress environments. This suggests that non-stress environments are more alike, and stress environments are very different from each other, and also not related to the non-stress ones. This behavior somewhat alludes to the Anna Karenina principle: "Happy families are all alike; every unhappy family is unhappy in its own way." A similar principle was also found in ecological risk assessment [20]. We applied this principle to guide the selection of several model parameters. The modeling framework and simulation results are detailed in the following sections.

3 Methods

To describe the whole premise of the model, we need to establish some concepts from biology and formalize those concepts using mathematical language for modeling. There are multiple ways of doing this, for this article, we are going to follow some standard definitions present in the literature [21, 22].

- **Genotype:** A genotype is the hereditary information carried by nucleic acid polymers DNA and/or RNA and passed on from a parent to its offspring. This can be modeled using a sequence σ with length L as below:

  $$\Sigma = (\Sigma_1, \Sigma_2, \ldots, \Sigma_L)$$

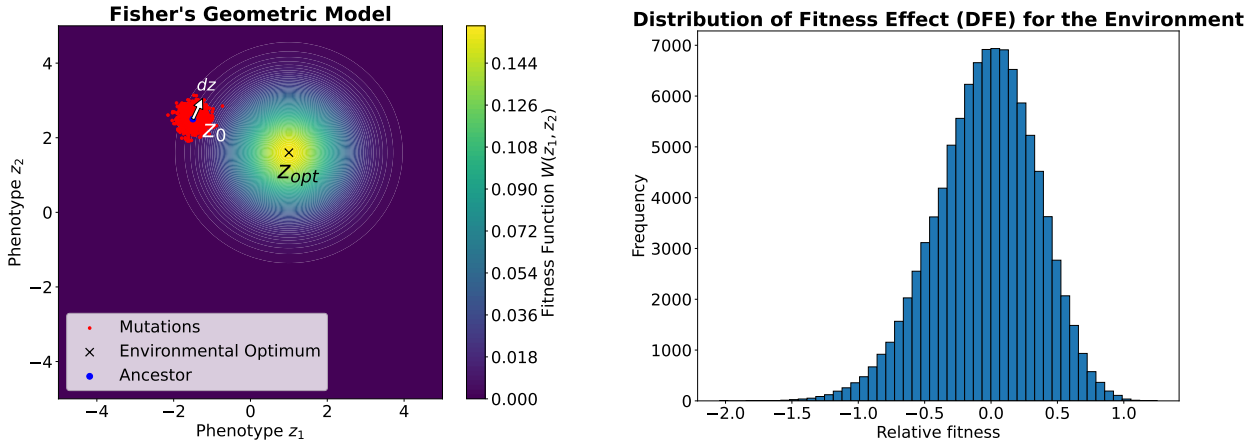
  where each element of a site is taken from an alphabet $\{0, 1, 2, \ldots, a-1\}$. Each element of an alphabet is called an **allele**.
- **Mutation:** In general, mutations are any form of changes in the genotype. For example one point mutations which simply change the allele at one site of a genotype Σ:

  $$\Sigma_i \to \Sigma_i', \text{ for any } i = 0, 1, \ldots, a-1.$$

- **Phenotype:** Phenotype of an organism is the observable characteristics or traits (e.g., plant height, color of leaves, abundance of a specific protein production inside a bacterium, etc.) produced by the interaction of the genotype and the environment.
- **Fitness:** A quantity that is proportional to the mean number of viable, fertile progeny produced by some genotype. Scientists often define this quantity based on their specific experimental setups and the variables they measure. An example is provided in equation (1). For the purposes of modeling, we will define it in terms of a fitness function, which is a real-valued function of phenotype and environment.

Fisher's geometric model is the paradigmatic representative of a phenotypic fitness landscape [23]; it bears a close relation to the antiferromagnetic Hopfield model with random continuous pattern vectors [24]. Nevertheless, in this article, we are not going to take advantage of that relation. Hence, we will use the following sets of assumptions to define this landscape and its relationships with the genotype and phenotype of a particular organism.

- **Phenotype space:** A phenotype space is a spanned by set of n phenotypic traits z_i represented by a column vector $|\mathbf{z}\rangle = \{z_i\}_{i\in[1,n]}$. A two-dimensional representation of the phenotype space is shown in Figure 3. The figure also contains the fitness function.
- **Genotype to Phenotype Relationship:** Typically, biological organisms have a large number of genes; we can assume the genotype Σ has a large length or $L \to \infty$. Therefore, we could ignore the discrete nature of



(a) (b)

Figure 3: Figure 3a illustrates Fisher's Geometric Model, where the ancestral species $|\mathbf{z}_0\rangle$ occupies a position in phenotype space, and the fitness function $W(z_1, z_2)$—representing a specific environment—has its optimum at $|\mathbf{z}_{\text{opt}}\rangle$. The red dots indicate small deviations in phenotype space resulting from mutations in the DNA of the ancestral species, and $|d\mathbf{z}\rangle$ is a corresponding vector in Phenotypic space. Figure 3b presents the corresponding distribution of fitness values for this scenario, commonly referred to as the Distribution of Fitness Effects (DFE).

mutation, and assign mutations that cause a phenotypic displacement $|d\mathbf{z}\rangle = \{dz_i\}_{i\in[1,n]}$ from an initial or ancestor phenotype $|\mathbf{z}_0\rangle$, and these vectors are distributed according to a multivariate normal distribution regardless of underlying environments $MVN(0, \sigma^2_{\text{mut}}\mathbf{I}_n)$. Here, σ^2_{mut} is the variance, and $\mathbf{I}_n$ is the $n \times n$ identity matrix. This hypothesis is called "universal pleiotropy" in the literature [25].

- **Phenotypic Fitness Landspace:** In this model, the fitness function is defined in terms of the phenotypic traits z_i. The fitness function monotonically decreases from the global maxima at $|\mathbf{z}_{\text{opt}}\rangle$ isometrically, which forms nonlinear fitness isoclines by which genotype–genotype interactions emerge [26]. The fitness of this ancestor phenotype $|\mathbf{z}_0\rangle$ is $W(|\mathbf{z}_0\rangle, |\mathbf{z}_{\text{opt}}\rangle)$ and a mutant has phenotype $|\mathbf{z}_0\rangle + |d\mathbf{z}\rangle$ and fitness $W(|\mathbf{z}_0\rangle + |d\mathbf{z}\rangle, |\mathbf{z}_{\text{opt}}\rangle)$. There would be one single optima in the landscape, and this hypothesis is referred as "stabilizing selection."
- **Environmental change:** The peak of $W(|\mathbf{z}\rangle, |\mathbf{z}_{\text{opt}}\rangle)$ defines the location of the optimal phenotype $|\mathbf{z}_{\text{opt}}\rangle$, which is going to change if the environment changes without changing the shape of the fitness landscape [13].
- **Relative Fitness or log-fitness:** The relative fitness or log-fitness function $H(|\mathbf{z}\rangle, |\mathbf{z}_{\text{opt}}\rangle)$ can be defined as follows:

$$H(\mathbf{z}) = \ln\left(\frac{W(\mathbf{z})}{W(\mathbf{z_0})}\right)$$

  This definition agrees with the notion of the typical behavior of a potential.
- **Small effect of mutation:** We will assume $|d\mathbf{z}\rangle$ remains small around the initial phenotype $|\mathbf{z}_{\text{opt}}\rangle$, so that $\ln W(|\mathbf{z}_0\rangle + |d\mathbf{z}\rangle, |\mathbf{z}_{\text{opt}}\rangle) - \ln W(|\mathbf{z}_0\rangle, |\mathbf{z}_{\text{opt}}\rangle)$ can be approximated by a second-order multivariate Taylor Series around $\mathbf{z}_0$, using Einstein's summation convention

$$H \approx \frac{\partial \ln W(|\mathbf{z}_0\rangle, |\mathbf{z}_{\text{opt}}\rangle)}{\partial z_i} dz_i + \frac{1}{2}\frac{\partial^2 \ln W(|\mathbf{z}_0\rangle, |\mathbf{z}_{\text{opt}}\rangle)}{\partial z_i \partial z_j} dz_i dz_j + o[dz^3]. \tag{3}$$

4 Results

The evolutionary fate of a population depends critically on the effects of individual mutations on organismal fitness. Mutations are broadly categorized as either deleterious or beneficial. To mathematically characterize these effects, one must consider the distribution of fitness values across all possible mutations, commonly referred to as the Distribution of Fitness Effects (DFE). Although determining the exact shape of the DFE is challenging, valuable insights can still be gained by examining key statistical parameters such as the mean, variance, and correlation coefficients. Using equation (3), we can compute the mean M and variance V of H over mutants $|d\mathbf{z}\rangle$. In literature, it has been observed that just an unbiased assumption could work, and there is no clear trend expected or observed for the effect of mutations on, for

example, morphological traits [27]. According to the genotype and phenotype relation in our model assumption $|d\mathbf{z}\rangle$ has a multivariate distribution, we can deduce that mutation effects on phenotypic traits ($|\mathbf{z}\rangle$) are unbiased, $E[dz_i] = 0$.

Mean: Now, keeping only terms up to the second order in dz, equation (3) yields the mean

$$M = E(H) \approx \frac{1}{2} \frac{\partial^2 \ln W(|\mathbf{z}_0\rangle, |\mathbf{z}_{\text{opt}}\rangle)}{\partial z_i \partial z_j} E[dz_i dz_j] + o[E[dz^3]], \tag{4}$$

From equation (4), the average relative fitness effect of mutations or M depends on the local curvature of the log-fitness function at the optimum, $\partial^2 \ln W(|\mathbf{z}_0\rangle, |\mathbf{z}_{\text{opt}}\rangle)/\partial z_i \partial z_j$. Indeed, symmetrical variation in dz can just introduce a bias in H if the phenotype-fitness relationship $\ln W(|\mathbf{z}_0\rangle, |\mathbf{z}_{\text{opt}}\rangle)$ is nonlinear. Now suppose, the logarithmic fitness is concave around $|\mathbf{z}_0\rangle$, that is, $J = \left[\partial^2 \ln W(|\mathbf{z}_0\rangle, |\mathbf{z}_{\text{opt}}\rangle)/\partial z_i \partial z_j\right]$ is a negative semi-definite matrix, then the mutations are deleterious on average ($M \leq 0$).

This claim can be proven by the first, utilizing the fact that the Hessian matrix $J = \lambda_i |\Lambda_i\rangle \langle \Lambda_j|$, can be decomposed into its eigenvalues $\lambda_i < 0$ ($\forall i = 1, \ldots, n$) and corresponding eigenvectors $|\Lambda_i\rangle$, and given a vector, $|dz\rangle = [dz_1, \ldots, dz_n]^T$ we can write

$$\frac{1}{2} \frac{\partial^2 \ln W(|\mathbf{z}_0\rangle, |\mathbf{z}_{\text{opt}}\rangle)}{\partial z_i \partial z_j} dz_i dz_j = \frac{1}{2} \langle dz| J |dz\rangle = \frac{1}{2} \lambda_i |\langle dz|\Lambda_i\rangle|^2 \leq 0.$$

Combining with eq(4) we get,

$$M \approx E\left[\frac{1}{2} \frac{\partial^2 \ln W(|\mathbf{z}_0\rangle, |\mathbf{z}_{\text{opt}}\rangle)}{\partial z_i \partial z_j} dz_i dz_j\right] + o[E[dz^3]] \leq 0.$$

From this result, we can imply that if $|\mathbf{z}_0\rangle$ is close to the global optima $|\mathbf{z}_{\text{opt}}\rangle$ where the log-fitness may be more concave than other places, the mean M is going be more negative, hence more deletarious. On the other hand, since $E[dz_i] = E[dz_j] = 0$, the terms $E[dz_i dz_j]$ are the variances and covariances of the effects of mutation accumulation on the underlying phenotypic traits. Furthermore, since dz distributed according to $MVN(0, \sigma^2_{\text{mut}} \mathbf{I}_n)$,

$$E[dz_i dz_j] = \delta_{ij} \sigma^2_{\text{mut}}.$$

This value quantify globally how and how much mutation accumulation affects the phenotype distribution among mutants. Now applying this multivariate normal distribution and the isotropic fitness function assumption i.e., $W(|\mathbf{z}_0\rangle, |\mathbf{z}_{\text{opt}}\rangle) = W(||\mathbf{z}_0 - \mathbf{z}_{\text{opt}}||)$ for the mutation vectors assumption, mean M becomes

$$M \approx \frac{1}{2} \frac{\partial^2 \ln W(|\mathbf{z}_0\rangle, |\mathbf{z}_{\text{opt}}\rangle)}{\partial z_i \partial z_j} \delta_{ij} \sigma^2_{\text{mut}} + o[E[dz^3]]$$

$$\approx \frac{\sigma^2_{\text{mut}}}{2} \left(\frac{\partial^2 \ln W(||\mathbf{z}_0 - \mathbf{z}_{\text{opt}}||)}{\partial (||\mathbf{z}_0 - \mathbf{z}_{\text{opt}}||)^2} + \frac{n-1}{||\mathbf{z}_0 - \mathbf{z}_{\text{opt}}||} \frac{\partial \ln W(||\mathbf{z}_0 - \mathbf{z}_{\text{opt}}||)}{\partial (||\mathbf{z}_0 - \mathbf{z}_{\text{opt}}||)} \right) + o[E[dz^3]]. \tag{5}$$

Here, the last line is derived using an identity related to transformation of the N-dimensional Laplace operator from cartesian to spherical coordinates.

Variance: As for the variance

$$V = E[H^2] - M^2 \approx E\left[\frac{\partial \ln W(|\mathbf{z}_0\rangle, |\mathbf{z}_{\text{opt}}\rangle)}{\partial z_i} \frac{\partial \ln W(|\mathbf{z}_0\rangle, |\mathbf{z}_{\text{opt}}\rangle)}{\partial z_j} dz_i dz_j + o[dz^3]\right] - M^2$$

$$\approx \frac{\partial \ln W(|\mathbf{z}_0\rangle, |\mathbf{z}_{\text{opt}}\rangle)}{\partial z_i} \frac{\partial \ln W(|\mathbf{z}_0\rangle, |\mathbf{z}_{\text{opt}}\rangle)}{\partial z_j} E[dz_i dz_j] + o[E[dz^3]]. \tag{6}$$

Let's impose the multivariate normal distribution for the mutation vectors. As a consequence, we get,

$$V \approx \left(\frac{\partial \ln W(|\mathbf{z}_0\rangle, |\mathbf{z}_{\text{opt}}\rangle)}{\partial z_i} \frac{\partial \ln W(|\mathbf{z}_0\rangle, |\mathbf{z}_{\text{opt}}\rangle)}{\partial z_i} \right) \sigma^2_{\text{mut}} + o[E[dz^3]]. \tag{7}$$

The mutational variance in relative fitness (V) is proportional to the product of first derivatives of $\ln W(|\mathbf{z}_0\rangle, |\mathbf{z}_{\text{opt}}\rangle)$ taken at $|\mathbf{z}_0\rangle$. This result is intuitively simple: the variance in the underlying phenotypic traits (z_i) translates into the variance in fitness according to the local slope of the fitness function (to $\partial \ln W(|\mathbf{z}_0\rangle, |\mathbf{z}_{\text{opt}}\rangle)/\partial z_i$) irrespective of its sign (hence the square).

Now, if the fitness function W monotonic with the distance to the optimum $||z_0 - \zeta_0||$. From which we essentially arrive at the following formula from equation (7)

$$V \approx \left(\frac{\partial \ln W(||\mathbf{z}_0 - \mathbf{z}_{\text{opt}}||)}{\partial(||\mathbf{z}_0 - \mathbf{z}_{\text{opt}}||)} \sigma_{\text{mut}} \right)^2 + o[E[dz^3]] \tag{8}$$

Correlation coefficient: If we consider two different environments with two different optimum $|\mathbf{z}_{\text{opt}}\rangle, |\mathbf{z}'_{\text{opt}}\rangle$, with two fitness landscapes $W(|\mathbf{z}_0\rangle, |\mathbf{z}_{\text{opt}}\rangle)$ and $W(|\mathbf{z}_0\rangle, |\mathbf{z}'_{\text{opt}}\rangle)$ respectively, we can calculate the Pearson's correlation coefficients,

$$\rho_{H,H'} = \frac{E(HH') - E(H)E(H')}{\sqrt{E(H^2) - E(H)^2}\sqrt{E(H'^2) - E(H')^2}}$$

where H, H' are the log-fitness for mutations with respect to $|\mathbf{z}_0\rangle$ for the environments with phenotypic optimums at $|\mathbf{z}_{\text{opt}}\rangle, |\mathbf{z}'_{\text{opt}}\rangle$ respectively. We can rewrite the formula of the correlation coefficients $\rho_{H,H'}$ in terms of means and variances as follows:

$$\rho_{H,H'} = \frac{E(HH') - MM'}{\sqrt{VV'}}.$$

Now, we can focus on $E[HH']$ by rewriting it more explicitly,

$$\begin{aligned} E[HH'] \approx\ & E\Bigg[\Big(\frac{\partial \ln W(|\mathbf{z}_0\rangle, |\mathbf{z}_{\text{opt}}\rangle)}{\partial z_i} dz_i + \frac{1}{2} \frac{\partial^2 \ln W(|\mathbf{z}_0\rangle, |\mathbf{z}_{\text{opt}}\rangle)}{\partial z_i \partial z_j} dz_i dz_j + o[dz^3] \\ & \times \Big(\frac{\partial \ln W(|\mathbf{z}_0\rangle, |\mathbf{z}'_{\text{opt}}\rangle)}{\partial z_p} dz_p + \frac{1}{2} \frac{\partial^2 \ln W(|\mathbf{z}_0\rangle, |\mathbf{z}'_{\text{opt}}\rangle)}{\partial z_p \partial z_q} dz_p dz_q + o[dz^3] \Big) \Bigg] \\ \approx\ & \frac{\partial \ln W(|\mathbf{z}_0\rangle, |\mathbf{z}_{\text{opt}}\rangle)}{\partial z_i} \frac{\partial \ln W(|\mathbf{z}_0\rangle, |\mathbf{z}'_{\text{opt}}\rangle)}{\partial z_j} E[dz_i dz_j] + o[E[dz^3]]. \end{aligned}$$

Next, we can impose $E[dz_i dz_j] = \delta_{ij}\sigma^2_{\text{mut}}$, which gives us

$$\begin{aligned} E[HH'] &\approx \frac{\partial \ln W(|\mathbf{z}_0\rangle, |\mathbf{z}_{\text{opt}}\rangle)}{\partial z_i} \frac{\partial \ln W(|\mathbf{z}_0\rangle, |\mathbf{z}'_{\text{opt}}\rangle)}{\partial z_i} \sigma^2_{\text{mut}} + o[E[dz^3]] \\ &\approx \Big\langle \frac{\partial \ln W(|\mathbf{z}_0\rangle, |\mathbf{z}_{\text{opt}}\rangle)}{\partial |\mathbf{z}\rangle}, \frac{\partial \ln W(|\mathbf{z}_0\rangle, |\mathbf{z}'_{\text{opt}}\rangle)}{\partial |\mathbf{z}\rangle} \Big\rangle \sigma^2_{\text{mut}} + o[E[dz^3]] \end{aligned}$$

where $\langle\ ,\ \rangle$ is just usual Euclidean dot product in $\mathbb{R}^n$ space. Moreover, we combine the above result, definition of correlation coefficient and using the observation that $MM' \approx o[E[dz^4]]$ to arrive to the following result:

$$\rho_{H,H'} \approx \frac{\Big\langle \frac{\partial \ln W(|\mathbf{z}_0\rangle, |\mathbf{z}_{\text{opt}}\rangle)}{\partial |\mathbf{z}\rangle}, \frac{\partial \ln W(|\mathbf{z}_0\rangle, |\mathbf{z}'_{\text{opt}}\rangle)}{\partial |\mathbf{z}\rangle} \Big\rangle}{\sqrt{\Big\langle \frac{\partial \ln W(|\mathbf{z}_0\rangle, |\mathbf{z}_{\text{opt}}\rangle)}{\partial |\mathbf{z}\rangle}, \frac{\partial \ln W(|\mathbf{z}_0\rangle, |\mathbf{z}_{\text{opt}}\rangle)}{\partial |\mathbf{z}\rangle} \Big\rangle \Big\langle \frac{\partial \ln W(|\mathbf{z}_0\rangle, |\mathbf{z}'_{\text{opt}}\rangle)}{\partial |\mathbf{z}\rangle}, \frac{\partial \ln W(|\mathbf{z}_0\rangle, |\mathbf{z}'_{\text{opt}}\rangle)}{\partial |\mathbf{z}\rangle} \Big\rangle}} + o[E[dz^3]].$$

At this point we invoke the radially symmetric assumption for the fitness function, i.e., $\ln W(|\mathbf{z}_0\rangle, |\mathbf{z}_{\text{opt}}\rangle) = \ln W(||\mathbf{z}_0 - \mathbf{z}_{\text{opt}}||)$. Hence,

$$\frac{\partial \ln W(||\mathbf{z}_0 - \mathbf{z}_{\text{opt}}||)}{\partial z_i} = \frac{W'(||\mathbf{z}_0 - \mathbf{z}_{\text{opt}}||)}{W(||\mathbf{z}_0 - \mathbf{z}_{\text{opt}}||)} \frac{(z_{0,i} - \zeta_{0,i})}{||\mathbf{z}_0 - \mathbf{z}_{\text{opt}}||}.$$

We can plug this into the formula of $E[HH']$, yielding

$$E[HH'] \approx \frac{W'(||\mathbf{z}_0 - \mathbf{z}_{\text{opt}}||)W'(||\mathbf{z}_0 - \mathbf{z'}_{\text{opt}}||}{W(||\mathbf{z}_0 - \mathbf{z}_{\text{opt}}||)W(||\mathbf{z}_0 - \mathbf{z'}_{\text{opt}}||)} \frac{\langle \mathbf{z}_0 - \mathbf{z}_{\text{opt}}, \mathbf{z}_0 - \mathbf{z'}_{\text{opt}} \rangle}{||\mathbf{z}_0 - \mathbf{z}_{\text{opt}}||\ ||\mathbf{z}_0 - \mathbf{z'}_{\text{opt}}||} \sigma^2_{\text{mut}} + o[E[dz^3]].$$

Finally, we can use this result in the formula of $\rho_{H,H'}$, we get that

$$\rho_{H,H'} \approx \cos\left(\theta_{\mathbf{z}_{\text{opt}}, \mathbf{z}'_{\text{opt}}}\right) + o[E[dz^3]] \tag{9}$$

where $\theta_{\mathbf{z}_{\text{opt}}, \mathbf{z}'_{\text{opt}}}$ is the angle between $\left|\mathbf{z}_{\text{opt}}\right\rangle - \left|\mathbf{z}_0\right\rangle, \left|\mathbf{z'}_{\text{opt}}\right\rangle - \left|\mathbf{z}_0\right\rangle$. This would imply that a high correlation between two environments would mean in the phenotypic space the optimums of the environments are located close to a straight line emanating from the ancestor phenotype and on the same side of the ancestor phenotype. On the contrary, a small correlation would indicate that the environmental optimums are forming almost a right angle with the ancestor phenotype.

Example: As an example, we will use a special functional form for the fitness function W, where the log-fitness would look like a power function, that is

$$W(||\mathbf{z} - \mathbf{z}_{\text{opt}}||) = e^{\alpha ||\mathbf{z} - \mathbf{z}_{\text{opt}}||^Q} \tag{10}$$

where $\alpha, s \in \mathbb{R}$ is some constant. One important reason for choosing this function is that if fitness function has this form, the equilibrium fitness function is not affected [2]. This type of function has been studied in this context before [28], but most widely used version is a quadratic function ($Q = 2$) with $\alpha < 0$ [29]. Next, we find the mean from the equation (5),

$$\begin{aligned} M \approx & \frac{\alpha \sigma^2_{\text{mut}}}{2} Q\,(Q + n - 2) \\ & \times ||\mathbf{z}_0 - \mathbf{z}_{\text{opt}}||^{Q-2} + o[E[dz^3]] \end{aligned} \quad , \tag{11}$$

and from the Equation (9) we get variance,

$$\begin{aligned} V \approx & \alpha^2 \sigma^2_{\text{mut}} Q^2 \\ & \times ||\mathbf{z}_0 - \mathbf{z}_{\text{opt}}||^{2Q-2} + o[E[dz^3]] \end{aligned} \quad . \tag{12}$$

It is apparent from equation (11) that mean M would grow linearly with the dimension of phenotypic space n, this parameter is also referred as the phenotypic complexity. Furthermore, since the model requires that there is a single maxima of the function, the log-fitness function is concave; hence $M < 0$, and that requires α to have a negative value. Unlike the mean M, variance is positive and does not depend on the phenotypic complexity up to the second order of σ_{mut}. The parameter Q plays a crucial role in determining the qualitative agreement between the model and experimental observations. With these results in hand, we are now prepared to analyze the experimental data and assess how well the Fisher's Geometric Model qualitatively matches the empirical findings.

5 Qualitative Comparison with Experimental Data

Figure 4a shows a cartoon picture of the phenotypic space where the ancestor phenotype $|\mathbf{z}_0\rangle$ is placed in the center of the phenotype space. From the pattern in the experimental data in Figure 2b we learned that the non-stress environments are alike (having a high correlation with each other), their optima are placed on a single axis. Furthermore, the optimum of each stress environment is placed on all the other perpendicular axes. According to equation (9), this configuration might give us a theoretically calculated distribution of correlation coefficients that is closer to the experimental one. However, since not all the correlation between stress environments and stress vs non-stress environments are not at the same value, we have introduced small noise values to the coordinates of the optima of the environments so that it produces a distribution more like the experimental one.

Moreover, in the last section we established the fact that, if the log-fitness function is more concave around some area, the value of mean is going to be more negative. So, if the neighborhood of environmental optima is more concave than the points far away from the optima, then the ancestor phenotype close to the environmental optima would have a more negative mean, and hence more deletarious mutation would appear in that environment. To test this hypothesis, the optima of stress environments are placed far from the ancestor phenotype, and for non-stress environments, the optima

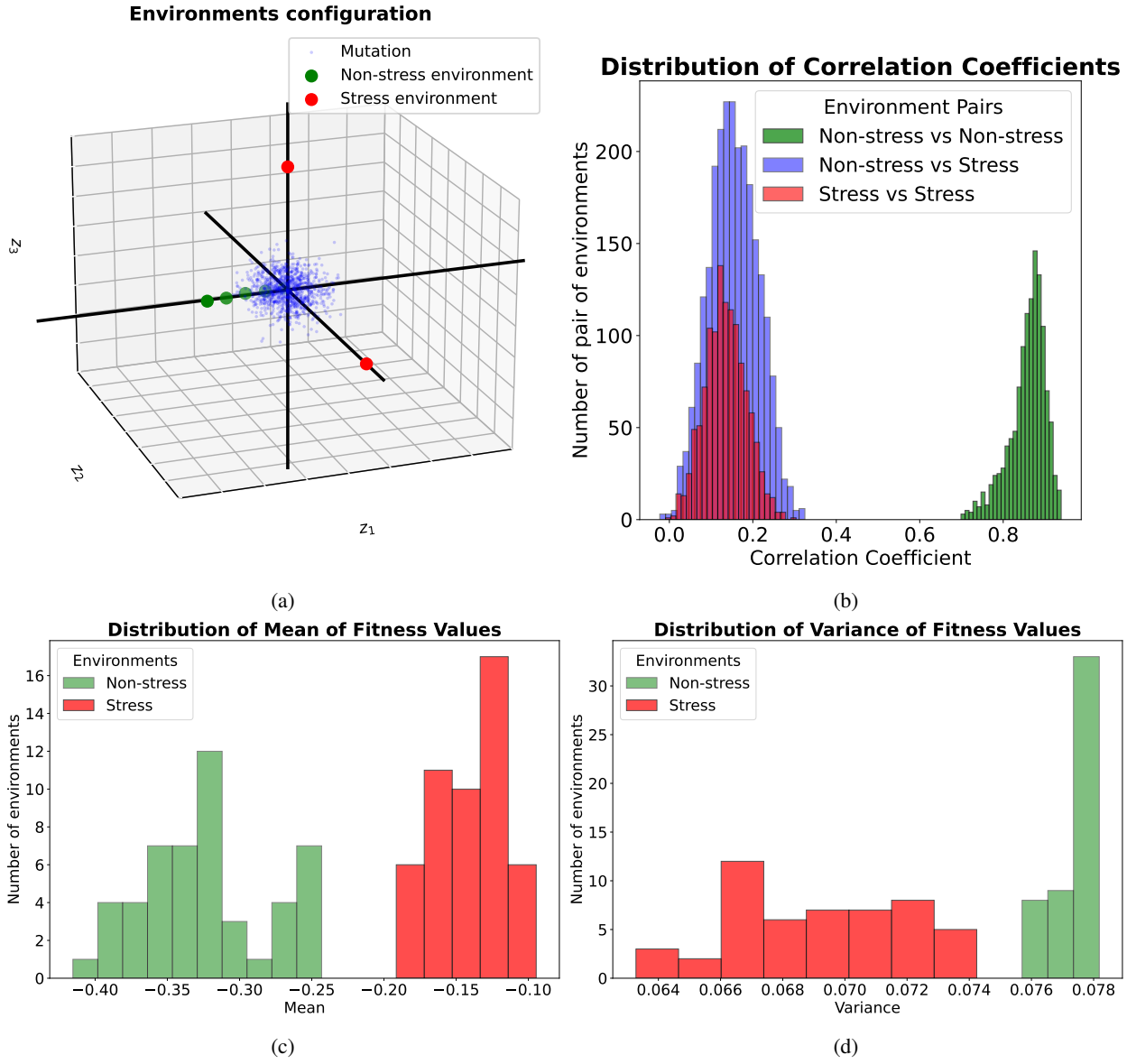


Figure 4: Figure 4a is a schematics diagram of the configuration of phenotypic space alongside with ancestor phenotype $|\mathbf{z}_0\rangle$ at the center of the coordinate system, and the blue dots representing new phenotypes after mutation. The red and green dots signify the phenotypic optimum of stress and non-stress environments, respectively. The simulation was performed using $10,000$ mutants and 100 environments within a 51 dimensional phenotypic space, which results in the distribution of mean, variance, and correlation coefficient depicted in the Figure 4c, 4d, and 4b, repectively.

are positioned close to the center of the phenotype space. These placements make sense because if the environment is most suitable for the ancestor, the optimum phenotype of that environment must coincide with the ancestor phenotype. If the optimum phenotype is closer to the ancestor phenotype, it would take a small number of steps to reach a stable situation. However, the opposite would happen with regard to optima being far from the ancestor phenotype.

In most of the literature, the Gaussian function is used to represent the fitness function, which amounts to using $Q = 2, \alpha < 0$ in the equation (10). However, this choice would not provide the correct qualitative statistical representation of the data presented in Figure 2. From equation (12), it is very easy to observe that if Q is equal to 2, the value of the mean M is independent of the distance from the environmental optimum phenotype to the ancestor phenotype $||\mathbf{z}_0 - \zeta_0||$, which contradicts with the experimental result shown in figure 2c. Moreover, $Q = 2$ implies that V would increase as $||\mathbf{z}_0 - \zeta_0||$ increases according to formula (11); it is completely opposite to the behavior displayed in figure 2d. Hence, we can conclude that traditional FGM with a Gaussian fitness function would not qualitatively reproduce the experimental data.

To achieve a more accurate qualitative match with the experimental data, we want both $|M|$ and V to decrease as $||\mathbf{z}_0 - \zeta_0||$ increases, despite $W(|\mathbf{z}_0\rangle, |\mathbf{z}_{\text{opt}}\rangle) = \exp\left(\alpha||\mathbf{z}_0 - \zeta_0||^Q\right)$ to decrease monotonically from the optimum, provided $\alpha < 0$. The monotonically decreasing condition can be achieved if Q is positive and according to equations (11) and (12), both M and V would decrease if $Q < 1$. Therefore, the suitable choice for Q is just between 0 to 1. In Figure 4 $Q = 1/2$ is used to generate the numerical simulation result for correlation coefficients, mean, and variance of relative fitness values of 50 stress and 50 non-stress environments. The phenotype space is chosen to be 51-dimensional, and all non-stress environments are placed near 50 different orthogonal axes. All stress environments are located along the final axis orthogonal to the other 50 axes.

Besides these specifications, we used $\sigma_{\text{mut}} = 0.1,\ \alpha = -1$ and, $10,000$ mutation vectors are being used for this simulation. As predicted by the theoretical calculation, in figures 4c and 4d, M and V behave as expected; however, we can see there is a clear separation between the values M and V of stress and non-stress environments which is not present in the experimental data. The mean and variances values are more spread out in experimental data compared to numerical simulation ones. It is also very apparent that the values of V in numerical simulation is far off from the experimental results. Furthermore, in Figure 2b the correlation coefficients between stress environments are more dispersed compared to the simulation in Figure 4b. There are also several experimental data points exhibiting negative correlations, which are not captured by the simulation results. Overall, the numerical simulation qualitatively reproduces some of the key statistical features of the empirical data. However, a more detailed analysis of the experimental data, and further refinement of the theoretical model are necessary to achieve a more accurate comparison.

6 Conclusion

In this work, we have utilized a simple set of assumptions within the Fisher's Geometric Model framework to derive theoretical results and, based on these, designed a specific configuration of the phenotypic space with different elements incorporated. Using this configuration, we computed key statistical parameters of the Distribution of Fitness Effects (DFE) and qualitatively compared them with experimental results. While the model successfully captures some features of the empirical data, it fails to account for finer nuances, highlighting its limitations. Additionally, some of our parameter choices were somewhat ad hoc. It may be possible to extract model parameters directly from experimental data, enabling a more concrete and quantitative comparison between the model and empirical observations. For instance, rather than introducing random noise to perturb coordinates, one could use equation (9) to more precisely determine the direction of the environment's optimum location. Subsequently, equation (12) could be employed to estimate the distance to the optimum based on experimental data. Finally, using these newly acquired coordinates for the environmental optimum, a more accurate numerical comparison could be envisioned by applying equation (11).

Furthermore, the agreement between the model and empirical data can potentially be improved by revising some of the underlying hypotheses. For example, instead of relying solely on the stabilizing selection hypothesis, one could design a fitness landscape with multiple optima or incorporate a flat region along certain directions. In addition to modeling the environmental optimum, other parameters that vary with the environment could be introduced, allowing the model to distinguish between stress and non-stress environments even when they exhibit similar means and variances. Similarly, relaxing the universal pleiotropy assumption and adopting a more complex probability distribution for the genotype-to-phenotype mapping could yield a richer, more realistic model. In fact, random matrix theory has previously been employed to capture this relationship more effectively [27]. Nonetheless, the strength of FGM lies in its simplicity, which has enabled it to make insightful predictions across many areas of evolutionary genetics—including the study of DFEs, epistasis, dominance, adaptive trajectories, and the distribution of fixed mutations. Our current analysis indicates that there is still considerable room for refinement; in particular, the emergence of a power-law log-fitness function in our study suggests opportunities to revisit and extend previously established results derived under the Gaussian fitness

assumption. Finally, this modeling framework could be applied to other microorganisms as well, offering a way to further investigate how environmental factors shape the distribution of fitness effects.

References

[1] Luca Agozzino, Gábor Balázsi, Jin Wang, and Ken A Dill. How do cells adapt? stories told in landscapes. *Annual review of chemical and biomolecular engineering*, 11(1):155–182, 2020.

[2] Guy Sella and Aaron E Hirsh. The application of statistical physics to evolutionary biology. *Proceedings of the National Academy of Sciences*, 102(27):9541–9546, 2005.

[3] Susanna Manrubia, José A Cuesta, Jacobo Aguirre, Sebastian E Ahnert, Lee Altenberg, Alejandro V Cano, Pablo Catalán, Ramon Diaz-Uriarte, Santiago F Elena, Juan Antonio García-Martín, et al. From genotypes to organisms: State-of-the-art and perspectives of a cornerstone in evolutionary dynamics. *Physics of Life Reviews*, 38:55–106, 2021.

[4] Lesley T MacNeil and Albertha JM Walhout. Gene regulatory networks and the role of robustness and stochasticity in the control of gene expression. *Genome research*, 21(5):645–657, 2011.

[5] Olga Kuchner and Frances H Arnold. Directed evolution of enzyme catalysts. *Trends in biotechnology*, 15(12):523–530, 1997.

[6] Sasha F Levy, Jamie R Blundell, Sandeep Venkataram, Dmitri A Petrov, Daniel S Fisher, and Gavin Sherlock. Quantitative evolutionary dynamics using high-resolution lineage tracking. *Nature*, 519(7542):181–186, 2015.

[7] Cheemeng Tan, Philippe Marguet, and Lingchong You. Emergent bistability by a growth-modulating positive feedback circuit. *Nature chemical biology*, 5(11):842–848, 2009.

[8] RA Fisher. The genetical theory of natural selection oxford, uk: Oxford univ. *Press Complet. variorium ed*, 1930.

[9] RC MacLean, GG Perron, and Andy Gardner. Diminishing returns from beneficial mutations and pervasive epistasis shape the fitness landscape for rifampicin resistance in pseudomonas aeruginosa. *Genetics*, 186(4):1345–1354, 2010.

[10] Ana Sousa, Sara Magalhaes, and Isabel Gordo. Cost of antibiotic resistance and the geometry of adaptation. *Molecular biology and evolution*, 29(5):1417–1428, 2012.

[11] Guillaume Martin, Santiago F Elena, and Thomas Lenormand. Distributions of epistasis in microbes fit predictions from a fitness landscape model. *Nature genetics*, 39(4):555–560, 2007.

[12] Daniel M Weinreich and Jennifer L Knies. Fisher's geometric model of adaptation meets the functional synthesis: data on pairwise epistasis for fitness yields insights into the shape and size of phenotype space. *Evolution*, 67(10):2957–2972, 2013.

[13] Guillaume Martin and Thomas Lenormand. The fitness effect of mutations across environments: a survey in light of fitness landscape models. *Evolution*, 60(12):2413–2427, 2006.

[14] Guillaume Martin and Thomas Lenormand. The fitness effect of mutations across environments: Fisher's geometrical model with multiple optima. *Evolution*, 69(6):1433–1447, 2015.

[15] Morgan N Price, Kelly M Wetmore, R Jordan Waters, Mark Callaghan, Jayashree Ray, Hualan Liu, Jennifer V Kuehl, Ryan A Melnyk, Jacob S Lamson, Yumi Suh, et al. Mutant phenotypes for thousands of bacterial genes of unknown function. *Nature*, 557(7706):503–509, 2018.

[16] Monica Riley, Takashi Abe, Martha B Arnaud, Mary KB Berlyn, Frederick R Blattner, Roy R Chaudhuri, Jeremy D Glasner, Takashi Horiuchi, Ingrid M Keseler, Takehide Kosuge, et al. Escherichia coli k-12: a cooperatively developed annotation snapshot—2005. *Nucleic acids research*, 34(1):1–9, 2006.

[17] FHC Crick. Project k:" the complete solution of e. coli". *Perspectives in Biology and Medicine*, 17(1):67–70, 1973.

[18] Rui Vitorino. Transforming clinical research: the power of high-throughput omics integration. *Proteomes*, 12(3):25, 2024.

[19] Kelly M Wetmore, Morgan N Price, Robert J Waters, Jacob S Lamson, Jennifer He, Cindi A Hoover, Matthew J Blow, James Bristow, Gareth Butland, Adam P Arkin, et al. Rapid quantification of mutant fitness in diverse bacteria by sequencing randomly bar-coded transposons. *MBio*, 6(3):10–1128, 2015.

[20] Dwayne RJ Moore. The anna karenina principle applied to ecological risk assessments of multiple stressors. *Human and Ecological Risk Assessment*, 7(2):231–237, 2001.

[21] Sakshi Pahujani and Joachim Krug. Complexity and accessibility of random landscapes. *arXiv preprint arXiv:2502.05896*, 2025.

[22] Olivier Tenaillon. The utility of fisher's geometric model in evolutionary genetics. *Annual review of ecology, evolution, and systematics*, 45(1):179–201, 2014.

[23] H Allen Orr. The genetic theory of adaptation: a brief history. *Nature reviews genetics*, 6(2):119–127, 2005.

[24] Su-Chan Park, Sungmin Hwang, and Joachim Krug. Distribution of the number of fitness maxima in fisher's geometric model. *Journal of Physics A: Mathematical and Theoretical*, 53(38):385601, 2020.

[25] Annalise B Paaby and Matthew V Rockman. The many faces of pleiotropy. *Trends in genetics*, 29(2):66–73, 2013.

[26] François Blanquart, Guillaume Achaz, Thomas Bataillon, and Olivier Tenaillon. Properties of selected mutations and genotypic landscapes under fisher's geometric model. *Evolution*, 68(12):3537–3554, 2014.

[27] Guillaume Martin and Thomas Lenormand. A general multivariate extension of fisher's geometrical model and the distribution of mutation fitness effects across species. *Evolution*, 60(5):893–907, 2006.

[28] Olivier Tenaillon, Olin K Silander, Jean-Philippe Uzan, and Lin Chao. Quantifying organismal complexity using a population genetic approach. *PloS one*, 2(2):e217, 2007.

[29] Russell Lande and Stevan J Arnold. The measurement of selection on correlated characters. *Evolution*, pages 1210–1226, 1983.