A General Statistical Method for Identifying Adaptations by Parameterizing Trait Space

Drew Blount*
Reed College

Keywords
Adaptation, adaptationism, natural selection, spandrel, trait

Abstract It is obviously useful to think of evolved individuals in terms of their adaptations, yet the task of empirically classifying traits as adaptations has been claimed by some to be impossible in principle. I reject that claim by construction, introducing a formal method to empirically test whether a trait is an adaptation. The method presented is general, intuitive, and effective at identifying adaptations while remaining agnostic about their adaptive function. The test follows directly from the notion that adaptations arise from variation, heritability, and differential fitness in an evolving population: I operationalize these three concepts at the trait level, formally defining measures of individual traits. To test whether a trait is an adaptation, these measures are evaluated, locating the trait within a three-dimensional parameterized trait space. Within this space, I identify a region containing all adaptations; a trait's position relative to this adaptive region of trait space describes its status as an adaptation. The test can be applied in any evolving system where a few domain-specific statistical measures can be constructed; I demonstrate the construction of these measures, most notably a measure of an individual's hypothetical fitness if it were born with a different trait, in Packard's Bugs ALife model. The test is applied in Bugs, and shown to conform with our intuitive classification of adaptations.

1 Introduction

Many students of evolving systems, both biological and man-made, think of evolved organisms in terms of adaptations—those traits that, by virtue of benefiting their host individual's fitness, have spread through the population via natural selection. This is the frame of mind of the zookeeper, explaining how the giraffe evolved its long neck to eat from the tops of trees; or the keeper of an ALife model, who describes features of the evolved beings in terms of their selective function. Seeing organisms in terms of their adaptations allows you to think of them functionally, to make sense of their parts, and to imagine their evolutionary history. Yet there has been controversy in the philosophy of biology over whether this so-called adaptationist approach is scientifically tenable.

Traditionally, to claim that something is an adaptation, you specify what purpose that adaptation serves—its selective function in organisms. I'll call this type of claim a *particular adaptive hypothesis*

* Artificial Life Lab, Reed College, Portland, OR 97202. E-mail: me@drew.computer

(PAH), and each has the form "Trait T is an adaptation with function f." One of the largest criticisms of adaptationism is that there is no general empirical method of identifying the function f, or of comparing several possible functions, and that even searching for f in the first place assumes that T is in fact an adaptation.

Yet many traits are not adaptations. As first emphasized in Gould and Lewontin's seminal critique of adaptationism, "The spandrels of San Marco and the Panglossian paradigm" [7], there are many causal forces at play in evolution beyond strict natural selection, and these forces are often better explanations for why a certain trait is prominent in a population. For example, someone could conceivably ask, "Why does the gorilla have thicker fingers than any other primate?" We could spend years studying gorilla behavior, trying to imagine why their fingers adapted to be so robust. But perhaps the best explanation is, "gorillas are the biggest primates, and the thickness of their fingers is appropriate for their size." This is called an *allometric* explanation of a trait, an explanation by morphological proportion. Allometry is one of several explanations that Gould and Lewontin cite as commonly causing *spandrels*, or traits that seem like they might have an adaptive explanation, but are actually caused by forces or constraints other than natural selection.

Gould and Lewontin's most important point, and their strongest criticism of adaptationism, is that there is generally no way to test if a trait is an adaptation or a spandrel. I refute this notion by construction. In this article, I present a general-purpose methodology to empirically identify adaptations, without making any claims as to their adaptive functions. I show that *general adaptive hypotheses* (GAHs)—that is, claims of the form "trait T is an adaptation" that are agnostic about T's adaptive function—are empirically testable claims. This shows that adaptationism is perfectly compatible with empiricism; there is nothing special about adaptations such that they cannot be identified methodologically.

The greatest strength of this test is its conceptual simplicity and coherence with our intuitive understanding of adaptations. It is inspired by three basic features that are necessary for natural selection to occur in a population:

1. variation between individuals,
2. heritability of those variations, and
3. differential fitness between individuals, affected by those variations.

It is essentially tautological that, in an evolving population displaying those three features, some degree of natural selection should occur. Though these requirements are usually framed in terms of an entire population, it is possible to think of them on a trait-specific level. Much of this article is a formalization of that idea, as I define traitwise measures of each of the above characteristics. Thus, I present three parameters defined for a generic trait T: its historical *variability* across the population, $V(T)$; its *heritability* from one generation to the next, $H(T)$; and the average effect of T on its host individual's fitness, T's *differential fitness* $dF(T)$.

Inspired by Peter Godfrey-Smith's multidimensional analysis of Darwinian populations [6], I use these measures to parameterize a three-dimensional trait space. Evaluating V, H, and dF for a certain trait locates it within this space. Because only traits with certain values of these parameters are adaptations, there is a region of this trait space where all adaptations are located. Thus, to test whether a trait is an adaptation, I plot it in this space and check its position relative to the adaptive region.

I describe how to parameterize the trait space of a general evolving system, but of course different types of evolving systems might consider different types of things to be traits. One of the central pillars of artificial life is the idea that natural selection can be instantiated in wildly diverse systems. There is almost no limit to what could be considered a trait in some context—some traits are numeric, like height; some traits are patterns, like a leopard's spots; and in some artificial contexts, objects as strange as twitter hashtags or strings of computer code can be interpreted through the trait lens.

To account for this, while still defining my test in the most general terms, I define V, H, and dF in terms of what I call *domain-specific measures*. The purpose of these measures is to quantify features that are common to all evolving systems, but can only be measured by taking into account system-specific details. For example, it is natural to talk about the amount of variance among the trait values at a certain locus in a population. The variance of height will be measured differently than the variance of a behavior: There is a simple distance metric between two height values (i.e., absolute difference), but quantifying distance between two behavior values would likely require some knowledge of the domain at hand.

My test is then applicable in any domain where you could define and evaluate the four domain-specific measures I use to construct V, H, and dF. I demonstrate this process in a simple ALife model, Packard's Bugs [11], where the test is shown to accurately differentiate between adaptations and non-adaptive traits.

Thus, my test evaluates general adaptive hypotheses, identifying adaptations in a way agnostic to their potential adaptive function. Working from the idea that natural selection arises from three features in an evolving system—variation, heritability, and differential fitness—I show how those features can be operationalized as measures of individual traits.

The next section contains a more in-depth discussion of the relationship between adaptationism and testability. In Section 3 I formally define the parameters V, H, and dF in terms of several domain-specific parameters. The case study of Packard's Bugs is presented in Section 4.

2 Motivation

Dawkins's seminal book *The Selfish Gene* is perhaps the most famous example of adaptationism, a stance that venerates the role of adaptation in evolving systems. To Dawkins, every biological organism is a "survival machine" run by "replicators," and "natural selection favors replicators that are good at building survival machines" [4, p. 24]. In other words, individuals are primarily vehicles for adaptations. This is a methodological stance as well as a philosophical one: The adaptationist analyzes evolved organisms in terms of the adaptations that compose them, and the selective functions served by those adaptations.

Laypeople think about evolution in terms of adaptations, too. The morphologies of culturally prominent species, for example, are usually explained as adaptive solutions to problems of survival: The giraffe's long neck is adapted to eat the leaves of tall trees, zebras' stripes are an adaptation that confuses their lion predators, and every feature of a great white shark is perfectly optimized by natural selection to find and kill prey.

These Animal Planet examples illustrate perhaps the most common justification of the claim "this trait is an adaptation," the claim I call a general adaptive hypothesis. The GAH is most often supported by a particular adaptive hypothesis, a claim that specifies the supposed adaptive function. For example, we assume that the giraffe's neck is an adaptation because the story about high leaves sounds intuitive and probable.

The process of proving the GAH with a PAH need not be purely an argument by plausibility, however, as PAHs can often be tested empirically. Adaptationists have typically judged the validity of a PAH by finding other hypotheses implied by that PAH, and testing those [5, 3]. For example, the PAH concerning the giraffe's neck implies a few things: that tall trees were present in the environment where giraffe ancestors developed their neck, that a considerable portion of a giraffe's diet comes from leaves from tall trees, and that other animals throughout history that evolved eating habits like the giraffe's would have similar morphologies as well (rather than just being tall). This is the argument behind Dawkins' paradoxically titled "Adaptationism was always predictive and needed no defense" [3].

Yet testing particular adaptive hypotheses is no general method for identifying adaptations: It can only be applied in cases where you have formed a complex, plausible conjecture about a trait's adaptive function, critically limiting the set of testable traits to those we already think we understand.

Further, this strategy can conclusively confirm only that a trait is an adaptation. Because a negative result might only indicate that you were wrong about the conjectured adaptive function, there's a sort of halting problem in that you cannot distinguish "no adaptive explanation exists" from "we have not found the right explanation quite yet." This is the motivation behind Gould and Lewontin's epistemic criticism in [7]. In their words, "We would not object so strenuously to the adaptationist program if its invocation, in any particular case, could lead *in principle* to its rejection for want of evidence" ([7, p. 153]; emphasis mine).

By construction, this article refutes Gould and Lewontin's claim that adaptations cannot in principle be empirically identified. Since [7], adaptationism has matured greatly, and I am not the first to defend its empirical foundation. Two tests of general adaptive hypotheses that have been proposed deserve special attention: one that emerged in the biological sciences, and another from artificial life.

The HKA test, a much-cited and often-used test in molecular biology, allows for adaptive genetic loci to be identified in cases where genomes of two related species can be analyzed [8]. The test works by assuming the neutral evolutionary theory that genetic drift is a constant evolutionary force. Because mutations occur in each biological reproduction event, an inactive genetic locus (one that does not affect fitness) will become increasingly divergent between two related populations, as they drift apart genetically. Within one of the populations, you similarly expect inactive loci to be polymorphic. The HKA test then recognizes adaptive loci as those that are divergent between the two populations, but not correspondingly polymorphic within one of them. Thus the HKA test works by recognizing features that have diverged between related species, but stayed relatively homogeneous within at least one of them. The relative homogeneity indicates selection pressure working against neutral drift, and thus identifies loci where adaptation has occurred.

The HKA test has influenced other genetic methods for identifying adaptations, notably the McDonald–Kreitman test, which works by a similar logic and is considered HKA's successor [10]. The study of genetics has proven very useful for identifying adaptations—clearly, Gould and Lewontin's complaints about sloppy adaptationist methodology would be much weaker if they were made today. Both HKA and McDonald–Kreitman require genetic analysis of two closely related species to identify adaptations in one, so they are not completely general tests of the GAH—it might be debatable whether these tests refute Gould and Lewontin's in-principle claim about adaptive hypotheses.

Bedau's measure of so-called evolutionary activity [1], however, attacks the in-principle objection to adaptationism head on. Evolutionary activity was introduced as a measure of the amount of adaptation occurring within a population. Roughly, the measure works by the assumption that features that persist in an evolving population do so because they are resisting selective pressure—like HKA, evolutionary activity tests against the assumption of neutral drift. Evolutionary activity is indeed quite general, in large part because each application of it requires a domain-specific operationalization of the notion of activity. It could be argued fairly easily that evolutionary activity plots [2] could be analyzed to identify particular adaptations: When measuring the activity of traits, prominent waves of trait activity correspond to adaptations.

Because evolutionary activity can identify adaptations without specifying their adaptive functions, and because it is a generally applicable test, I consider it an in-principle refutation of Gould and Lewontin's thesis that general adaptive hypotheses are untestable. Several aspects of evolutionary activity have influenced the construction of my test, especially that it is defined in terms of some generic functions that will have different definitions depending on the system being studied.

My goal here is to offer another test of the general adaptive hypothesis, complementary to evolutionary activity. The value of the test presented here is its conceptual simplicity, in that it follows directly from the notion that natural selection occurs when variation, heritability, and differential fitness are all present in a population. By constructing an empirical test so that its components match as closely as possible with a nontechnical understanding of adaptation, I hope to soundly refute Gould and Lewontin's claim that adaptationist premises are inherently untestable—and,

of course, to develop a method that will be useful for identifying adaptations, both conceptually and in practice.

3 Variability, Heritability, Differential Fitness

This section presents my empirical test of the general adaptive hypothesis in general terms, so that the methods described here can be applied in a wide range of settings. I will define each of the traitwise adaptive parameters (variability V, heritability H, and differential fitness dF) in terms of domain-specific measures like trait distance and fitness. First, I will explain some assumptions behind the test.

This entire test, like most discussions surrounding adaptationism, assumes that individuals have readily identifiable, discrete traits. The division of an organism into traits is not always obvious, nor obviously justified, but that is not a discussion I will go into here.

I will assume that traits exist and can be measured, and speak of them in terms of loci. I use *trait locus* (or just *locus*, plural *loci*) to mean a set of alternative traits in a population, such that each individual has exactly one trait from the set. In these terms, the human eye color locus contains the traits for blue, brown, green, and so on. I will speak in terms of sets, saying "The locus containing T," "a trait at locus L," or "$T \in L$" to associate the two.

Some loci are finite sets of discrete traits, like the eye color locus or the presence or absence of a certain mutation. Other loci contain a continuum of trait values, for example, if you consider an organism's height to be a trait. Traits can also be integer-valued, like an Avidian's reproductive period, measured in time units. In less traditional evolutionary contexts—say, if you are studying an evolving corpus of text like Twitter through an ALife lens—you can imagine traits that might be strings of symbols. Many ALife systems, Avida being a prominent example, have traits made out of computer code [9].

It should be clear that the concept of trait, and by extension locus, can be instantiated wildly differently in different evolving systems. Similarly plastic is the notion of individual fitness, which can be a function explicitly defined by a programmer, as in genetic algorithms; an endogenous function, emergent from the system's dynamics, as in biological life; or perhaps something in between.

Because the concepts of traits and fitness can be so different in disparate evolving systems, there are some statistical analyses of traits that I can discuss in only general terms. For example, the seemingly simple statistic of the variance of all the trait values in the population at a given locus: If we're talking about height, this is quite simple; we calculate the statistical variance of the set of a population's height values. But what about traits of non-numeric types? What is the variance of a set of strings? A set of subroutines inside a "for" loop? A set of behaviors? Determining such metrics will often require domain-specific knowledge of the evolving system under study. For this reason, V, H, and dF are defined in terms of what I'll call *domain-specific measures*. These measures, like the variance example, quantify features that are common to all evolving systems, yet might be defined or measured differently depending on the domain. I will point out these measures as they are used to define V, H, and dF, and summarize them in Section 3.4.

A consequence of the generality of this approach is that the geography of the trait space parameterized by V, H, and dF will display the same trends in any application, but the details will of course be domain-specific. As I define each parameter, I will describe generally what values correspond with adaptations rather than non-adaptive traits, and I will consider the adaptive region to be the intersection of these adaptive ranges. It is important to note that the boundary of this region is not sudden, as there is not an absolute threshold between adaptations and non-adaptations in any of the dimensions of trait space. That is not to say that there are not regions of the space that contain only adaptations or only non-adaptations, just that the boundary between the two regions is fuzzy. This is a feature, not a flaw, as any test of adaptations should be sensitive to the fact that different traits can have different amounts of "adaptation-ness."

3.1 Variability

The necessity of variability to adaptation should be obvious—natural selection would have nothing to work with in a population of perfectly identical clones. More precisely, a trait can only be an adaptation if the locus containing that trait has been historically heterogeneous in the population. The purpose of the measure $V(T)$ is to quantify this heterogeneity, to ensure that trait values at T's locus have been nonuniform in the historical population.

In some settings, it might be sufficient to consider $V(T)$ to be a binary function, simply a yes–no answer to the question, "Has the population always been fixed at T's locus?" where "always" means the period of time where we ask whether natural selection acted on T. This basic question is enough to distinguish some spandrels from adaptations. For example, if we asked why humans evolved to have an appendix, measuring $V(T)$ for the trait "has an appendix" over the past several million years might indicate that none of our recent ancestors varied in their having of appendices, and therefore that natural selection never could have selected any alternative to the appendix trait.

It is intuitive that loci that are highly variable might adapt differently than loci that are almost entirely uniform. Therefore $V(T)$ is not binary but real-valued. There are several intuitive ways historical variance could be defined, and I wish to make no strong claim that what I propose here is the very best, but rather that it is sufficient for the purposes of this test. In order to present my formulation of V, I must introduce the first domain-specific measure necessary for applying my adaptive test. This is the measure of *phenotypic distance*, d_p:

$$d_p(T, T') = \text{phenotypic distance between traits } T \text{ and } T'. \tag{1}$$

For numeric traits like height, d_p will most often be distance on the number line. But in general, this measure must be sensitive to details of the traits being inspected: It is much easier to quantify the difference between my height and my friend's than it is to quantify the difference between our faces.

I will assume that d_p is one-dimensional and real-valued, which requires that, if traits differ in more than one dimension, those differences can be combined into a single value—for example, as Euclidean distance is calculated from distance in orthogonal dimensions. With this measure in hand, it is obvious to define $V(T)$ as the expected distance between two individuals' traits at T's locus, for individuals in some historical population P_h. Choosing P_h will be another domain-specific consideration. In some cases, it will make sense to define P_h as the union of all living populations over some period. In others, it will make sense to ask how variable the initial population was, so P_h will be the population at time 0. Writing $\tau(I, L_T)$ to denote I's trait at T's locus, and using the expected value function E, I then define V:

$$V(T) = E[d_p(\tau(I, L_T), \tau(I', L_T))] \qquad \text{for } I, I' \in P_h. \tag{2}$$

In this formulation, $V(T)$ is equal for all traits at the same locus. Thus, it is useful for distinguishing adaptive loci rather than specific adaptive traits at a locus. Low values of $V(T)$ indicate that a population has barely varied at T's locus, and thus T is likely not an adaptation. Thus, $V(T)$ has some lower threshold for adaptations, the value of which will of course be domain-specific.

3.2 Heritability

Heritability, like variability, is an obvious requirement for Darwinian adaptation, and it is even more straightforward to formalize. Obviously, a trait can only be an adaptation if it is heritable—if a parent with that trait is likely to have children with it, too. Therefore, $H(T)$ measures the inverse expected difference between parents' and children's traits at the locus L containing trait T, where

the parent has T. Thus a large value of $H(T)$ indicates that trait T is highly heritable. To formalize this measure, I again use the domain-specific measure of phenotypic distance, d_p:

$$H(T) = E[d_p(\tau(I, L_T), \tau(I', L_T))] \qquad \text{where } I \text{ has } T \text{ and is the parent of } I'. \tag{3}$$

Again, $H(T)$ measures the inverse expected phenotypic distance between a parent's and a child's traits at T's locus, for all parents with T. Like $V(T)$, there is a lower threshold on H-values that allow for adaptation, the numeric details of which will be domain- and population-specific.

3.3 Differential Fitness

Variation and heritability within a population are not enough to ensure natural selection: The differences between individuals, measured by V (with inheritance measured by H), must influence reproductive fitness. If variations in a population do not affect fitness, then they are irrelevant to the process of natural selection.

The third dimension of interest, then, after V and H, is a measure of a trait's average influence on its host individuals' fitness. In the spirit of Dawkins' *Selfish Gene*, this can be considered equivalent to the fitness of an individual trait: The fittest traits make for the fittest individuals, which reproduce the most and spread their traits throughout the world. The question is, how can you empirically measure a trait's fitness if you only have a fitness function for individual organisms? To achieve this, I introduce the *differential fitness* of a trait T, $dF(T)$: a measure of how much the fitness of individuals with trait T depends on their having it, rather than another trait at the same locus.

As we are concerned with fitness, $dF(T)$ is only measurable in domains where individual fitness is measurable. Therefore I speak in terms of the domain-specific measure of individual I's fitness, $F(I)$. Depending on the system, this could be either a predefined fitness function (in certain artificial evolving systems) or a statistical measure of reproductive effectiveness. In my example of Packard's Bugs, for example, I am looking at adaptations over a 10,000-time-step period, so I consider each individual's fitness to be its number of living descendants 10,000 time steps after its birth. As has consistently been the case in this discussion, specifics of the measurement of $F(I)$ are likely to be dependent on the evolutionary system at hand.

Now, because we are interested in the effect of T upon I's fitness, we naturally ask what would happen to $F(I)$ if everything else were equal, but I did *not* have trait T. Of course, it does not usually make sense to think of an individual with one of its traits removed—we must think about what would happen if the individual had a different trait instead. Thus, we need a traitwise counterfactual measure of fitness: For an individual I with trait T, and a competing trait at the same locus, T', we wish to answer "What would be the fitness of I if it had T' instead?" I call this measure the *counterfactual fitness of I with T'*, or simply the counterfactual fitness, and write it

$$F_{CF}(I, T \rightarrow T'). \tag{4}$$

F_{CF} is the first of two counterfactual domain-specific measures that I will use. Because counterfactual scenarios can be explicitly explored in some evolving systems (e.g., by manually setting the state of an artificial life simulation), but not in others, this measure must be domain-specific.

Counterfactual fitness is the crucial tool that allows for a trait's influence on an individual's fitness to be measured. First, it allows us to straightforwardly measure the *relative fitness advantage* of T over T' in an individual I, that is, the amount by which I's fitness would be different if it had T' rather than T:

$$\delta F(I, T \rightarrow T') = F(I) - F_{CF}(I, T \rightarrow T'), \tag{5}$$

where $\delta F(I, T \rightarrow T')$ is the relative fitness advantage of trait T over T' in individual I, which is assumed to have trait T.

I am working towards a measure of T's influence on $F(I)$, and so far have considered scenarios where I's value at T's locus is switched with a single other trait, T'. Then, if L is T's locus, we are interested in an aggregate measure of $\delta F(I, T \rightarrow T')$ over all $T' \in L$. This value reflects individual I's average fitness advantage over all hypothetical individuals I' that are identical to I in all respects but the trait T.

To construct such an aggregate measure of $\delta F(I, T \rightarrow T')$ over all $T' \in L$, it is important to allow for the fact that not all alternative traits T' are equally likely alternatives: Some traits at a locus might be very common mutations, while others might be exceedingly rare. A trait's aggregate fitness advantage over other traits at its locus should incorporate this fact, leading me to introduce the final domain-specific measure necessary to apply this test, $Prob_{CF}$. Whereas $\delta F(I, T \rightarrow T')$ measures the fitness of I in the counterfactual scenario where it were born with trait T' rather than T, $Prob_{CF}(I, T \rightarrow T')$ measures the prior likelihood that that scenario would have occurred in the system:

$$Prob_{CF}(I, T \rightarrow T') = \text{prior probability that the reproduction event that produced I would have produced an individual with } T'. \tag{6}$$

In most cases, an individual's trait at a given locus is determined by a random function of its parents' traits (or parent's trait) at that locus. Given I's parents' traits, $Prob_{CF}(I, T \rightarrow T')$ could be straightforwardly calculated by repeatedly evaluating or simulating the component of the reproduction function responsible for setting the trait at T's trait locus.

To arrive at a measure of how much I's fitness is affected by the presence of T rather than the probable alternative traits I might have been born with, I simply average T's relative fitness advantage over $T' \in I$ (Equation 5), weighted by the counterfactual probability (Equation 6) that I had been born with T', for all T' at T's locus. I call this the *individualized differential fitness* of T in I, because it expresses the expected difference in I's fitness if it did not have T. The individualized differential fitness dF_{ind} is then defined,

$$dF_{ind}(I, T) = \sum_{T' \neq T \in L_T} Prob_{CF}(I, T \rightarrow T') \delta F(I, T \rightarrow T'). \tag{7}$$

The symbol dF_{ind} is deliberately evocative of the parameter dF, and the astute reader will guess how I construct the latter from the former. Recall that I intend for dF to measure how a trait compares to other traits, in determining individuals' fitness in a population. dF_{ind} makes exactly this comparison, but for only one individual who carries the relevant trait. Thus, dF is simply the average of dF_{ind} for all individuals who carry the trait in question. In population P, dF is defined by

$$dF(T) = \frac{1}{|P_T|} \sum_{I \in P_T} dF_{ind}(I, T), \quad \text{where} \quad P_T = \{I \in P \mid I \text{ has trait } T\}. \tag{8}$$

In terms of the domain-specific measures F (the fitness function), F_{CF} (Equation 4) and $Prob_{CF}$ (Equation 6), $dF(T)$ is a pair of nested averages:

$$dF(T) = \frac{1}{|P_T|} \sum_{I \in P_T} \sum_{T' \in L} Prob_{CF}(I, T \rightarrow T') \times (F(I) - F_{CF}(I, T \rightarrow T')), \tag{9}$$

where P_T is as defined in Equation 8.

Comparing the above with Equation 5, you see that $dF(T)$ is a weighted average of T's relative fitness advantage over all other traits and all individuals in the population that have T. Because this average is weighted by the likelihood that an individual with T had another trait instead (i.e., $Prob_{CF}$), $dF(T)$ considers T in the context of the system's reproductive dynamics, where some alternative traits might be likelier alternatives than others. Traits that are not likely alternatives to T have little effect on T's differential fitness, and conversely, those with high $Prob_{CF}$ have a large effect on $dF(T)$. The traits with the highest differential fitness will be those such that an average individual with that trait has a much higher fitness than it would, were it to have any likely alternative trait.

3.4 Domain-Specific Measures

I have formally defined V, H, and dF in terms of several so-called domain-specific measures—functions that, despite quantifying notions common to all evolving systems, can do so only by including domain-specific details that are not shared by evolving systems in general. The test is only practicable in systems when these measures are defined and evaluatable, though it is important to note that the measures used to define dF need not necessarily be evaluated as often as they appear in its definition (Equation 9). As I show in Packard's Bugs, where dF is defined in terms of average values of both F_{CF} and $Prob_{CF}$, it is sometimes possible to determine these average values analytically or by some significant computational shortcut, though this can only be done with specific knowledge of the domain at hand. In any case, the domain-specific measures summarized in Table 1 must be defined for my test to be applied.

3.5 The Parameterized Trait Space

By measuring $V(T)$, $H(T)$, and $dF(T)$, one situates the trait T within a parameterized trait space. Within this space, there is a region where adaptations lie. If T is found to lie within this region, it is likely an adaptation. Though the detailed geography of this adaptive region will of course be domain-specific, dependent on dynamics particular to individual evolving systems, there are some obvious general features that the region will display in all domains.

Because adaptations occur only when there is some amount of variability, heritability, and differential fitness, the adaptive region is distant from all three axes in trait space. As mutations often enable evolution, and the presence of mutation in a population implies imperfect heritability, there is likely an upper threshold for H-values of adaptations as well. These considerations lead to the schematic in Figure 1.

Table 1. The domain-specific measures necessary to measure V, H, and dF.

Notation	Description	Definition	Used to define
$F(I)$	Fitness of I	Native fitness function for individuals in the evolving system at hand	dF
$d(T, T')$	Difference between T and T'	A real-valued measure, quantifying the phenotypic distance between traits T and T'.	V, H
$F_{CF}(I, T \rightarrow T')$	Counterfactual fitness of I with trait T replaced by T'	The hypothetical fitness of I, if I had been born with T' instead of T	dF
$Prob_{CF}(I, T \rightarrow T')$	Counterfactual probability that I had T' instead of T	Probability that the reproduction event which conceived individual I with trait T, had instead conceived an individual with trait T'.	dF

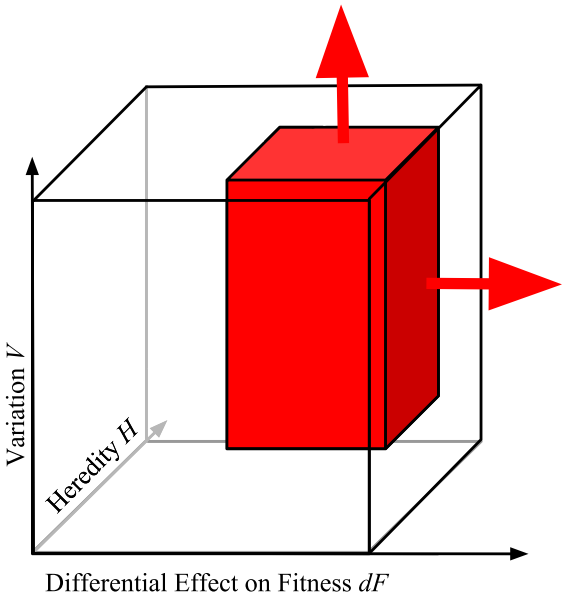

Figure 1. A dramatically simplified diagram of the adaptive region (the highlighted block) within the trait space parameterized by *V*, *H*, and *dF*. There is a lower threshold for adaptations along all three dimensions, and an upper threshold in the *H* dimension. In reality, the shape of this region would be expected to be more nuanced, and to vary according to the domain.

One important feature of this trait space is that in most evolving systems its boundaries are fuzzy rather than abrupt; there is not a sharp edge between an area of trait space where all traits are adaptations and an area where none of them are. Rather, there is likely a gradient between wholly non-adaptive and wholly adaptive regions of trait space. This is a feature, not a flaw, as the concept of adaptation is itself fuzzy, especially in the most complex evolving systems. Of course, all boundary details of the adaptive region of trait space will be domain-specific, and can potentially be explored through empirical survey, as I begin to do in the following example.

4 Proof of Concept in Packard's Bugs

Because measuring the differential fitness dF is significantly more involved than measuring V or H, I will demonstrate its measurement and usefulness in an evolving population where V and H are controlled, distinguishing adaptations from non-adapted traits in Packard's artificial life model Bugs [11]. In my experiments, V and H are the same for every trait in the system (as I will explain), so dF alone distinguishes adaptations from non-adaptive traits.

4.1 The Rules of Bugs

Creatures in Packard's Bugs are called, of course, bugs (not capitalized). The bugs live in a 2D grid world, and their genome is an instruction set, which deterministically maps a bug's sensory input to its movement throughout the world. A bug occupies one cell in the grid world, and its sensory input is only five bits: whether or not there is food in the bug's current position and in each of the adjacent cells. I'll call this information the *food pattern* at the bug's current position. Food is binary in each cell—either there, or not—and the global food distribution is permanent, that is, food is never created, destroyed, or moved.

There are $2^5 = 32$ food patterns that a bug could possibly see, and a bug's genome contains one movement rule for each pattern. At each time step, each bug moves according to the rule in its genome indexed by its local food pattern. These rules simply tell the bug to move 1–15 units in one of the eight cardinal directions. Thus, each of a bug's genes encodes a rule of the form "If you see this, move this many steps in that direction."

Like many evolving creatures, bugs need to eat in order to live and reproduce. Each bug has a numeric quantity of internal energy. Each time step, a bug's energy increases if it is in a cell with food (the bug eats), and decreases otherwise. Whenever a bug moves, it exerts an amount of energy that is proportional to the distance moved. When a bug's energy drops below zero, it dies. If a bug's gene ever instructs it to move to an already occupied position, the bug begins a random walk at that position to find an unoccupied space. Because this walk costs energy, and there are only so many cell in the world, overcrowding often kills bugs, and individuals compete over space.

Bugs reproduce whenever their internal energy goes above a certain threshold. One child is made asexually, a clone of its parent except for random pointwise mutations. This child occupies a random cell adjacent to its parent, and takes half of the parent's energy.

In my simulations, the Bugs world always had a fixed global food map: a foodless desert with a square oasis of food covering a quarter of the world's area. In this setup, only 14 of the 32 loci in the bug genome could ever be expressed, as most five-cell food patterns are not realized anywhere in the world map. I applied the test on a large population with random initial genomes, to ensure enough variability for selection to occur. This is a common starting scenario in artificial life experiments, and allows one to easily examine every trait in the trait space over repeated simulations.

4.2 Measuring *dF*

Measuring dF requires domain-specific measures of the fitness F, the counterfactual fitness F_{CF}, and the counterfactual probability P_{CF} (Equation 6), as illustrated in Equation 9. Fitness is the simplest to define. In evolving systems without explicitly encoded fitness functions, something similar to fecundity will likely capture the notion of fitness, and Packard's Bugs is one such case. I do not define individual fitness exactly as fecundity, or number of children, because two bugs that are equally successful at feeding themselves, and therefore reproducing, often differ greatly in the number of their children that survive to themselves reproduce. For this reason, I chose to define an individual's fitness as the number of living descendants it has at some point after its birth. Because I was interested in identifying adaptations over a 1,000-time-step period, I chose to define fitness in terms of that period as well:

$$F(I) = \text{number of living descendants of } I \text{ 1,000 time steps after } I\text{'s birth.} \tag{10}$$

Though I could measure the two counterfactual parameters by explicitly simulating counterfactual scenarios, as described in Section 3.3, doing so would be quite computationally intensive. Rather, the simplicity of Packard's Bugs allows me to get good estimates of the weighted sum in Equation 9 without explicitly evaluating each $F_{CF}(I, T \to T')$ and $P_{CF}(I, T \to T')$.

First, consider the measure of counterfactual probability P_{CF}. The motivation behind this measure, described in Section 3.3, is to weight the average of a trait T's relative fitness advantage over all other traits T', so that those traits that are more likely alternatives to T are considered more heavily. Yet in Packard's Bugs, all possible mutations at a locus occur with exactly equal probability. This allows me to essentially ignore P_{CF}, because the probability of mutation events is independent of the traits involved—the $P_{CF}(I, T \to T')$ terms in the sum in Equation 9 will amount to a constant factor in dF that is independent of T.

Thus, in Packard's Bugs, $dF(T)$ is a function of the average difference between $F(I)$ and $F_{CF}(I, T \to T')$, for all I with T. Because my experiments began with a random initial population,

the average value of $F_{CF}(I, T \rightarrow T')$ is simply the average value of $F(I')$, where I' has T'. Thus, letting P denote the population,

$$dF(T) \propto \sum_{I \in P_T} \sum_{T' \in L} F(I) - F_{CF}(I, T \rightarrow T') = mean(\{F(I) \mid I \text{ has } T\}) - mean(\{F(I') \mid I' \in P\}). \quad (11)$$

This is the domain-specific version of Equation 9, which I will use to measure *dF*. Interpreted in the context of the descendants-count definition of fitness, $dF(T)$ measures the expected number of descendants an individual with T will have, as a deviation from the average number of descendants for a generic bug. This captures the notion of a trait's average effect on host individuals' fitness in Packard's Bugs.

4.3 Results

I measured $mean(\{F(I) \mid I \text{ has } T\})$ for all traits at all loci over 100,000 simulations of Packard's Bugs, making the evaluation of Equation 11 very straightforward. As expected, traits that we intuitively believe to be adaptations have the largest empirical values of $dF(T)$. Good adaptations are distinguishable from decent ones, and maladaptations (traits that seem to actually harm their hosts' fitness) have negative differential fitness values. Further, selectively active and inactive loci can be easily identified by the range of *dF*-values of traits at that locus. These results are summarized in Table 2.

Because the data are interesting mostly for conforming with how people intuitively understand Packard's Bugs, I'll spend a moment describing standard bug behavior, and the most common strategies that bugs evolve.

Recall that in my simulations, all food is concentrated in a solid square area at the center of the bug world. Because bugs only see their five-cell neighborhood, a bug can only sense a few scenarios: whether it is in the interior of the food square, in the foodless desert, adjacent to an edge and inside the square, adjacent and outside, or in one of the corners on the inside of the square (bugs sense each corner and edge separately, as their genome ignores any kind of symmetry).

Because no other cell in the bug world has the same local food map as the interior NW corner of the food square, the space in a bug's genome encoding behavior in that scenario only rarely affects the bug's behavior. Because any successful bug spends most of its time surrounded by food, that locus on the genome is very important. Most loci in bugs remain dormant in the food-square world, because they contain instructions for scenarios that never occur there, for example, when a bug's current cell has food but none of the neighboring ones do. Traits at these loci have no effect whatsoever on the bugs' phenotypes. Thus you would expect all traits at these dormant loci to have empirical *dF* near zero, and for *dF*s to vary most at the loci where there is the most selective pressure, such as the locus encoding behavior in the food square.

In most individual simulation runs, one or two species of bug eventually dominate the world, generally occupying most of the central food square. Almost always, these dominant species will follow a simple strategy:

- While in the food square, move slowly in some direction (I will call this direction forward).
- When at the edge of the square (ideally the interior of the edge), make a large move in the opposite direction (backward).

In a trait space as simple as Bugs, and the world with the central food square, it is clear that this is a very good evolutionary strategy. First, it optimizes the amount of time the bug spends in the food square, ensuring that it reproduces as often as possible. Second, the large backward move in the second step ensures that the species' carrying capacity is large. After moving backwards, the bug must make many small forward steps before it reaches the edge again. This allows many bugs of the same species to march single-file along the same path.

Table 2. The extreme $dF(T)$ values at very active (top two), slightly active (middle), and inactive loci (bottom) in the bugs' genome.

Locus	Trait	$dF(T)$
	1 S	26.70
	15 NW	−1.90
	12 N	0.65
	9 S	−0.37
	7 E	0.13
	14 SW	−0.15
	15 SW	6.18
	1 W	−0.70
	12 N	0.42
	7 E	−0.22
	11 SW	0.17
	14 SE	−0.15

Notes. At each locus, I list the traits with the highest and lowest dF-values. Also note that for the active loci, the highlighted traits correspond with extremely good and bad strategies, given the environment's food distribution. All values were scaled by a constant to make them legible.

Note that the above strategy is represented in the data for active loci in Table 2 (top, middle). At these loci, the traits with the lowest dF-values are also intuitively bad responses to the stimulus encoded by the locus. For example, the trait with the lowest $dF(T)$ at the locus corresponding with the northern wall of the global food square instructs its host bug to jump out into the middle of the foodless desert.

Another important feature of the data is the range between the largest and smallest dF-values at each locus, which makes apparent which loci are most active. At completely dormant loci (Table 2, bottom), the difference between minimum and maximum dF is due only to statistical noise. Thus, in Packard's Bugs, $dF(T)$ identifies adaptive loci and the adaptations within them, as well as maladaptations.

5 Conclusion

A new general test for general adaptive hypotheses has been presented. In this test, a trait is measured along three dimensions: variability, heritability, and differential effect on fitness. Measuring these values for a single trait locates it within a parameterized trait space. In this trait space, all traits that are adaptations lie within a certain region. Thus, you find if trait T is an adaptation by locating it within trait space and seeing if it lies in the adaptive region.

This test is desirable for its conceptual simplicity. It is often said that natural selection occurs wherever a reproducing population meets three criteria:

- There is variation between individuals.

- Those variations are generally passed from parents to children.
- The variations between individuals affect their reproductive fitness.

One of the beautiful things about natural selection is that it follows logically from those simple features. In order to identify particular adaptations, which are traits that are prominent because they have benefited individual fitness and thus driven natural selection, I consider those three features in the context of a single trait rather than an entire population. It is those traits that (1) have historically varied in the population, (2) are able to be propagated by reproduction, and (3) benefit individuals' reproductive fitness that are most likely adaptations. These three features are formalized as the trait parameters V, H, and dF, for variability, heritability, and differential fitness.

To maintain generality my test is defined in terms of four domain-specific measures (listed in Table 1), which must be defined and evaluatable in a system in order to apply my test. Two of these are quite straightforward: the measure of trait similarity discussed above, and a measure of individual fitness. I am confident that satisfactory definitions of these two functions could be constructed straightforwardly in most evolving systems. The other two domain-specific measures are not as simple, as they deal with counterfactual scenarios. Specifically, they quantify (1) the probability that an individual with one trait might have had another instead, and (2) its fitness in that hypothetical scenario. I do not expect that these measures will be as easy to define as the other two; in biological life, for example, a trustworthy measure of counterfactual fitness would not be simple to construct. Nonetheless, in many artificial life contexts, counterfactual scenarios can be perfectly well explored by direct manipulation of the evolving system.

Further, as I demonstrate in measuring dF in an implementation of Packard's Bugs [11], it is possible to reason about the system at hand to find computational shortcuts, and avoid explicitly evaluating the large number of counterfactual scenarios mentioned in the definition of dF. I was able to measure dF of all traits at all loci in Bugs, with nothing more than fecundity data by genome over a huge number of simulations with random initial populations—still a bit of a computational task, but an ordinary personal computer was able to handle it in a few hours overnight.

It is worth noting that Bugs is too simple to have truly confounding spandrels of the type Gould and Lewontin were most concerned with. Thus, the results presented here should be considered a proof of concept, a demonstration that my test can be implemented and that empirical measurements of $dF(T)$ agree with our intuitions about adaptations. Because I measured dF in Bugs and held V and H constant at rates amenable to natural selection, a clear goal of further research is to measure traits in Bugs that vary along all three dimensions, to map trait space more thoroughly. Another obvious extension of this work is to implement the presented test in more, and more complex, evolving systems.

The framework I have presented here, my early results, and the previous works of those like Bedau decisively show one thing: General adaptive hypotheses are testable. Students of evolving systems can rest assured that adaptations are real things, identifiable by empirical methods, able to be discovered procedurally. Like Bedau's evolutionary activity measure [1], implementing the test presented here requires intimate statistical knowledge of the population at hand, and so it will be much easier applied in artificial than in biological life. Yet as artificial evolving systems become increasingly complicated, with increasingly many causes and constraints influencing evolution besides natural selection, spandrels will only become more prominent. The test presented here should help us make sense of the most complex, emergent models of artificial life.

References

1. Bedau, M. A. (1992). Measurement of evolutionary activity. In C. G. Langton, C. Taylor, J. D. Farmer, & S. Rasmussen (Eds.), *Artificial life II: Proceedings of the workshop on artificial life* (pp. 431–461). Redwood City, CA: Addison-Wesley.

2. Bullock, S., & Bedau, M. A. (2006). Exploring the dynamics of adaptation with evolutionary activity plots. *Artificial Life*, *12*(2), 193–197.

3. Dawkins, R. (1983). Adaptationism was always predictive and needed no defense. *Behavioral and Brain Sciences*, *6*, 360–361.

4. Dawkins, R. (2006). *The selfish gene*. Oxford, UK: Oxford University Press.

5. Dennett, D. (1998). The Leibnizian paradigm. In D. L. Hull & M. Ruse (Eds.), *The philosophy of biology*. Oxford, UK: Oxford University Press.

6. Godfrey-Smith, P. (2009). *Variation, selection, and origins*, Chapter 3. Oxford, UK: Oxford University Press.

7. Gould, S. J., & Lewontin, R. C. (1979). The spandrels of San Marco and the Panglossian paradigm: A critique of the adaptationist programme. *Proceedings of the Royal Society of London B*, *205*, 581–598.

8. Hudson, R. R., Kreitman, M., & Aguad, M. (1987). A test of neutral molecular evolution based on nucleotide data. *Genetics*, *116*(1), 153–159.

9. Lenski, R. E., Ofria, C., Pennock, R. T., & Adami, C. (2003). The evolutionary origin of complex features. *Nature*, *423*(6936), 139–144.

10. McDonald, J. H., & Kreitman, M. (1991). Adaptive protein evolution at the Adh locus in Drosophila. *Nature*, *351*(6328), 652–654.

11. Packard, N. H. (1989). Intrinsic adaptation in a simple model for evolution. In C. G. Langton (Ed.), *Artificial life* (pp. 141–155). Redwood City, CA: Addison-Wesley.