Evolution of Evolvability In Rapidly Adapting Populations

James T. Ferrare[1], Benjamin H. Good[2,3,4]

[1]Biophysics Program, Stanford University, Stanford, CA 94305
[2]Department of Applied Physics, Stanford University, Stanford, CA 94305
[3]Department of Biology, Stanford University, Stanford, CA 94305
[4]Chan Zuckerberg Biohub – San Francisco, San Francisco, CA 94158

* bhgood@stanford.edu (B.H.G)

Abstract

Mutations can alter the short-term fitness of an organism, as well as the rates and benefits of future mutations. While numerous examples of these evolvability modifiers have been observed in rapidly adapting microbial populations, existing theory struggles to predict when they will be favored by natural selection. Here, we develop a mathematical framework for predicting the fates of genetic variants that modify the rates and benefits of future mutations in linked genomic regions. We derive analytical expressions showing how the fixation probabilities of these variants depend on the size of the population and the diversity of competing mutations. We find that competition between linked mutations can dramatically enhance selection for modifiers that increase the benefits of future mutations, even when they impose a strong direct cost on fitness. However, we also find that modest direct benefits can be sufficient to drive evolutionary dead-ends to fixation. Our results suggest that subtle differences in evolvability could play an important role in shaping the long-term success of genetic variants in rapidly evolving microbial populations.

Main Text

The benefits of new mutations can manifest over multiple timescales. Some mutations alter the short-term fitness of an organism, while others can also affect the rates and fitness benefits of future mutations. Examples are common in the microbial world. Mutations in DNA repair genes can generate mutator strains with dramatically elevated mutation rates (1–3). These same variants also alter the molecular spectrum of new mutations, which can shift the relative probabilities of adaptive variants in addition to their overall rates (4–6). Other classes of mutations can open or close novel adaptive pathways through epistatic interactions with other genes (7–12). Striking examples of these "evolvability modifiers" have been observed in laboratory evolution experiments (8, 10, 11, 13; Fig. 1A), and they are thought to play a critical role in cancer (14–16) and the evolution of antibiotic resistance (17–19). But despite their potential importance, it is difficult predict when these long-term benefits should be favored by natural selection.

Classical arguments from modifier theory suggest that asexual populations will select for mutations that maximize their long-term mean fitness (20, 21; SI Section 1). This simple result applies for infinite populations near mutation-selection equilibrium. Both conditions are frequently violated in adapting populations, since a beneficial variant can fix before its long-term costs or benefits are fully realized. This greediness creates an inherent tension between short-term and long-term fitness (22, 23).

While some of these trade-offs can be understood in simple cases where mutations accumulate one-by-one (24–27; SI Section 2), most microbial populations reside in a qualitatively different regime. In many cases of interest, from laboratory evolution experiments (10, 28–30) to natural populations of viruses (31–34), bacteria (35–37), and certain cancers (14, 15), multiple beneficial mutations can arise and compete within the population at the same time. The competition between these linked mutations ("clonal interference"; 38) ensures that a successful variant must often generate multiple additional mutations to fix, which amplifies the indirect selection for evolvability. While recent work has started to explore these effects for mutation-rate modifiers alone (24), we currently lack an analytical framework for understanding more general differences in evolvability in the high-diversity regimes most relevant for microbes.

Our limited understanding of these dynamics leaves many basic questions unresolved: How does natural selection balance the short-term costs or benefits of a mutation with its long-term impact on evolvability? Which evolvability differences matter most for determining a mutation's long-term fate? And how do these answers depend on external factors like the size of the population or the diversity of competing mutations?

Modeling second-order selection in rapidly adapting populations

To address these questions, we analyze the fates of general evolvability modifiers in a population genetic model that explicitly accounts for interference among competing beneficial mutations. We consider an asexual population of N individuals that can acquire beneficial mutations at a large number of linked genetic loci. The mutations accessible to each genotype $\vec{g}$ can be summarized by their distribution of fitness effects, $\mu(s|\vec{g})ds$, which represents the per generation rate of producing mutations with fitness effects $s \pm ds/2$ (Fig. 1B). For most of this work, we will assume that the beneficial sites are sufficiently numerous and epistasis sufficiently weak that $\mu(s)$ remains approximately

constant over the relevant timescales (which we determine self-consistently below). This implies that the rate of adaptation of the population will approach a steady-state value $v(\mu, N)$ that depends on the values of $\mu(s)$ and N (39, 40). Given these assumptions, the simplest possible evolvability modifier is one that changes the distribution of fitness effects to a new value, $\mu_m(s)$, which is maintained for several additional substitutions (Fig. 1B). This can be viewed as the leading-order term in a more general expansion in the amount of "macroscopic epistasis" (41) in $\mu(s|\vec{g})$ (Fig. S1; SI Section 3.1). It generalizes the notion of a mutator allele to capture more subtle changes in evolvability, while bypassing the tremendous complexity of the underlying fitness landscape.

If the modifier takes over the population, the rate of adaptation will shift to a new value $v(\mu_m, N)$ that reflects the altered supply of mutations (e.g. Fig. 1A). To reach these high frequencies, the modifier must initially grow from a single founding individual, where stochastic fluctuations play a critical role. The outcome of this competition can be described by the fixation probability, $p_{\text{fix}}(\mu \to \mu_m)$, which provides a quantitative measure of the mutant's long-term reproductive value.

The fate of the modifier will strongly depend on its initial genetic background. While the distribution of backgrounds is often complicated at the genetic level, previous work has shown that progress can be made by grouping individuals by their relative fitness and modeling the resulting dynamics in fitness space (39, 40, 43). The distribution of fitnesses in the background population will approach a steady-state shape $f(x|\mu, N)$ that increases in fitness at rate $v(\mu, N)$ (Fig. 1C). This distribution has a characteristic width, x_c, which roughly coincides with the fittest individuals in the population. A new modifier mutation will arise on a genetic background with fitness x, and will compete with the wildtype population while acquiring further mutations from $\mu_m(s)$ (Fig. 1C, right). The outcome of this competition can be summarized by the conditional fixation probability $w(x|\mu \to \mu_m)$ (Fig. 1C, left). In rapidly adapting populations, the fate of the modifier will often be determined while it is still at a low frequency in the population (SI Section 3). When this assumption holds, the conditional fixation probability can be approximated by a standard branching process recursion,

$$\begin{aligned} 0 \approx & \underbrace{x \cdot w(x|\mu \to \mu_m)}_{\text{selection}} + \int \underbrace{\mu_m(s)\left[w(x+s|\mu \to \mu_m) - w(x|\mu \to \mu_m)\right]}_{\text{mutation}} ds \\ & - \underbrace{v(\mu, N) \cdot \partial_x w(x|\mu \to \mu_m)}_{\text{adaptation of the wildtype}} - \underbrace{\frac{1}{2} \cdot w(x|\mu \to \mu_m)^2}_{\text{genetic drift}}, \end{aligned} \tag{1}$$

which represents a balance between (i) the growth of the lineage due to selection, (ii) the adaptation of the background population, (iii) the stochastic effects of genetic drift when the lineage is at low copy-number, and (iv) the production of further mutations (SI Section 3). The overall fixation probability of the modifier can then be obtained by averaging over its random genetic background,

$$p_{\text{fix}}(\mu \to \mu_m; s_m) = \int w(x + s_m|\mu \to \mu_m) \cdot f(x|\mu, N) dx\,, \tag{2}$$

where we have also allowed the modifier to have a direct cost or benefit s_m. Together, Eqs. (1) and (2) provide a quantitative framework for understanding the trade-offs

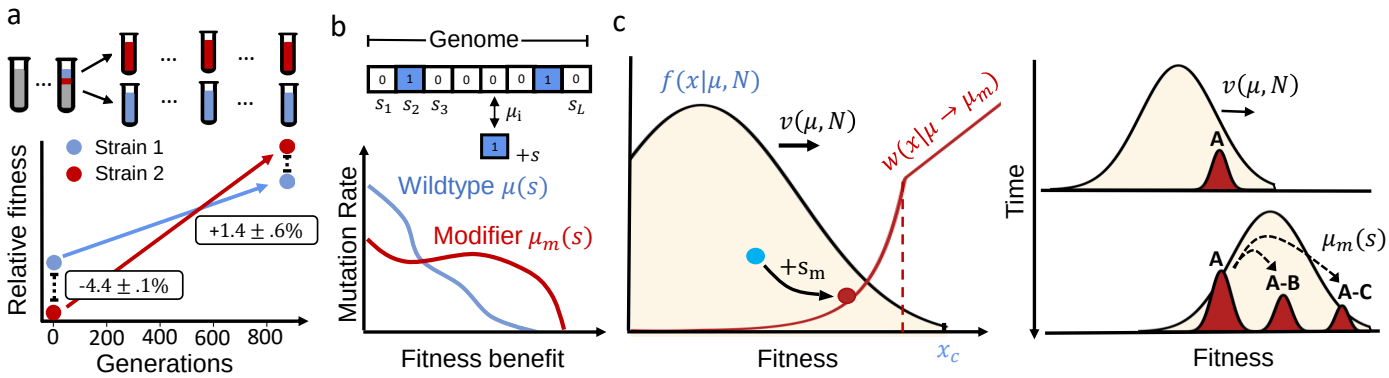

Fig 1. Modeling selection on evolvability in rapidly adapting asexual populations. (a) An empirical example of an evolvability modifier from Ref. (8). Two *E. coli* strains isolated from a long-term evolution experiment (42) had an initial fitness difference of ∼4%. The less-fit strain exhibited a higher rate of adaptation in replay experiments, which allowed it to consistently overcome its initial disadvantage after ∼900 generations of evolution. Fitness differences denote mean ± s.e.m. from 20 independent replays. **(b)** These long-term benefits can be modeled by the accumulation of beneficial mutations at a large number of linked genomic loci. The fitness benefits of the mutations are summarized by their distribution $\mu(s)ds$, which denotes the total rate of producing mutations with fitness effects $s \pm ds/2$. An evolvability modifier changes this distribution to a new value, $\mu_m(s)$. **(c)** A modifier with a direct cost or benefit (s_m) arises on a genetic background from the wildtype fitness distribution, $f(x|\mu, N)$, which has a maximum relative fitness x_c (left). The modifier lineage competes with the wildtype population as they both acquire further mutations (right). The outcome of this competition is described by the conditional fixation probability $w(x + s_m|\mu \to \mu_m)$, which exhibits a sharp transition at a critical initial fitness $x_{cm} \neq x_c$.

between "first-order" selection on fitness and "second-order" selection on evolvability that generalizes the mutator model in Ref. (24).

The fixation probability of a neutral mutation must always satisfy $p_{\text{fix}}(\mu \to \mu; 0) = 1/N$ (44). This provides a natural scale for interpreting the fixation probability in Eq. (2). Variants with $p_{\text{fix}}(\mu \to \mu_m; s_m) \gg 1/N$ are strongly favored by natural selection, while variants with $p_{\text{fix}}(\mu \to \mu_m; s_m) \ll 1/N$ are effectively purged. For this reason, it will be convenient to examine the scaled fixation probability, $\tilde{p}_{\text{fix}}(\mu \to \mu_m; s_m) \equiv N p_{\text{fix}}(\mu \to \mu_m; s_m)$, so that the sign of $\log \tilde{p}_{\text{fix}}$ coincides with the net "direction" of natural selection (45).

Selection for evolvability in a simple fitness landscape

To understand the second-order selection pressures encoded in Eqs. (1) and (2), we start by considering a simple model for the distribution of fitness effects, where all mutations confer the same characteristic fitness benefit s_b and occur at a total rate U_b. The most general evolvability modifier will either change the selection strength ($s_b \to s_b'$), the mutation rate ($U_b \to U_b'$), or both (Fig. 2A). This simplified model allows us to obtain an analytical solution for the fixation probability that is valid for empirically relevant scenarios where $x_c \gg s_b \gg \sqrt{v}$ (SI Section 4).

For a pure selection-strength modifier ($s_b \to s_b'$), we find that the fixation probability

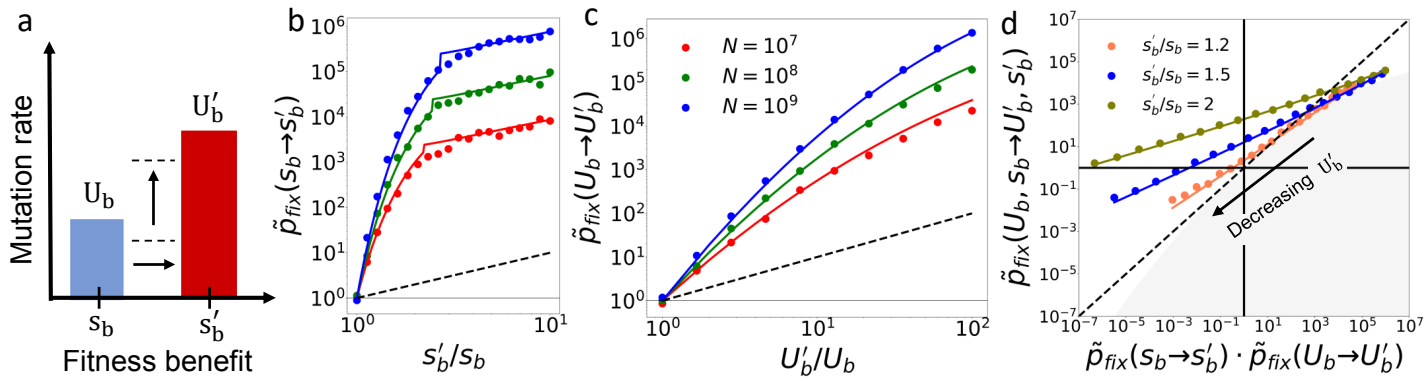

Fig 2. Interference between linked mutations enhances second-order selection for evolvability. **(a)** A simplified model for the distribution of fitness effects (Fig. 1B), where all mutations confer the same characteristic fitness benefit. An evolvability modifier will change the overall mutation rate ($U_b \to U_b'$), the overall selection strength ($s_b \to s_b'$), or both. **(b, c)** The fixation probability (scaled by the neutral expectation, $1/N$) for a selection-strength modifier (b) or a mutation-rate modifier (c). Symbols denote the results of forward-time simulations (SI Section 8.1) for $s_b = 10^{-2}$, $U_b = 10^{-5}$, and $N = 10^7 - 10^9$. Solid lines denote our theoretical predictions (SI Section 8.2), while the dashed lines denote the null expectation in the absence of clonal interference (SI Section 2). **(d)** Fixation probability of a compound modifier ($s_b \to s_b'$ and $U_b \to U_b'$) compared to an additive null model. Symbols denote the results of forward-time simulations for $s_b = 10^{-2}$, $U_b = 10^{-5}$, and $N = 10^8$; the y-coordinates are obtained from simulations, while the x-coordinates are obtained from the theoretical predictions in panels b and c. Solid lines denote our theoretical predictions, which deviate substantially from the additive expectation (dashed lines). The gray region indicates forbidden combinations that cannot arise in our theory (SI Section 8.2).

initially grows as an exponential function of the new selection strength,

$$\log \tilde{p}_{\text{fix}}(s_b \to s_b') \sim \left[1 - \left(\frac{s_b}{s_b'}\right)^2\right] \log\left(\sqrt{(Ns_b)\,(NU_b)}\right), \tag{3}$$

before saturating to a linear scaling when $s_b' \gtrsim \sqrt{2x_c s_b}$ (Fig. 2B). This behavior is qualitatively different than that observed for mutation-rate modifiers (24),

$$\log \tilde{p}_{\text{fix}}\left(U_b \to U_b'\right) \sim \left(\frac{x_c}{s_b}\right) \log\left(\frac{U_b'}{U_b}\right), \tag{4}$$

whose benefits increase more slowly with the mutation rate (Fig. 2C). In both cases, the fixation probabilities in Fig. 2B,C are significantly larger than the linear scaling expected when mutations accumulate via discrete selective sweeps (Fig. 2B,C; SI Section 2). This gap grows increasingly large as a function of NU_b, demonstrating that the competition between linked mutations enhances second-order selection for evolvability.

The origin of this behavior has a simple heuristic explanation. For small changes in the selection coefficient ($s_b' - s_b \ll s_b$), successful mutations typically arise in the high-fitness "nose" of the population ($x \approx x_c \gg s_b$; Fig. 1C) and must acquire $\sim x_c/s_b$ additional mutations before they reach appreciable frequencies. In each of these steps ($j = 1 \ldots x_c/s_b$), a selection-strength modifier produces $\sim \exp\left[\frac{s_b}{v} \cdot j(s_b' - s_b)\right]$ more mutations than a wildtype individual with the same fitness, leading to the exponential scaling observed in Eq. (3). The linear saturation at larger values of s_b' occurs when a

single additional mutation is sufficient to ensure fixation. However, since successful modifiers still arise on anomalously fit genetic backgrounds ($x \sim \sqrt{x_c s_b}$), they are able to hitchhike to higher frequencies than they would on their own, thereby increasing their overall probability of producing a successful mutation. Both effects become increasingly important for larger values of x_c/s_b, which highlights how they critically depend on the genetic diversity that is maintained within the population.

The strength of second-order selection can be intuitively quantified by comparing Eqs. (3) and (4) with the corresponding fixation probability of a first-order mutation,

$$\log \tilde{p}_{\text{fix}}(s|\mu) \sim \begin{cases} \frac{x_c s}{v} & \text{if } s \ll x_c, \\ \log(2Ns) & \text{if } s \gg x_c, \end{cases} \tag{5}$$

that has been studied in previous work (39, 40, 46; SI Section 4.1.3). Comparing this expression with Eq. (3) shows that even fractional changes in the selection coefficient ($\Delta s_b/s_b \sim \sqrt{s_b/x_c} \lesssim 1$) can generate fixation probabilities as large as a typical beneficial mutation ($s \approx s_b$). This contrasts with the behavior observed for a mutation-rate modifier in Eq. (4), where the mutation rate must increase by several orders of magnitude [$\log(\Delta U_b/U_b) \sim s_b^2/v \gg 1$] to achieve the same effect. This shows that large populations can more efficiently select on changes to the fitness benefits of future mutations, compared to the overall rate at which they occur.

The fixation probability of a compound modifier ($s_b \to s_b'$, $U_b \to U_b'$) can be understood using these basic building blocks. We find that the leading-order contributions can be expressed as a linear combination,

$$\log \tilde{p}_{\text{fix}}(U_b, s_b \to U_b', s_b') \sim \log \tilde{p}_{\text{fix}}(s_b \to s_b') + \alpha \cdot \log \tilde{p}_{\text{fix}}(U_b \to U_b'), \tag{6}$$

where the weighting factor α is given by $\alpha \approx \max\{(s_b/s_b')^2, s_b/x_c\}$. This additional weighting factor implies that mutation-rate and selection-strength modifiers do not additively combine: an increase in s_b will act to temper selection on the mutation rate, while decreases in s_b will amplify it. This non-additivity arises because larger selection-strength modifiers eventually lower the number of mutations required to fix ($x_c/s_b \longrightarrow x_c s_b/s_b'^2$), which diminishes the compounding effects of the altered mutation rate. These differences can be large, and can alter the sign of selection on the modifier (Fig. 2D). Moreover, since the individual components in Eq. (6) depend on the underlying parameters in different ways, the sign of selection can also vary as a function of the population size and the overall mutation rate (24, 26). These examples illustrate how selection-strength modifiers can lead to qualitatively different behavior than expected for mutation-rate changes alone, and that even modest shifts in s_b can frequently overpower order-of-magnitude differences in U_b.

Trade-offs between first- and second-order selection

We are now in a position to understand how natural selection balances the short-term costs and benefits of a mutation with its long-term impact on evolvability. When the evolvability deficits are not too large (SI Section 4.1.5), we find that the leading-order behavior can be decomposed into contributions from first- and second-order selection,

$$\log \tilde{p}_{\text{fix}}(U_b, s_b \to U_b', s_b'; s_m) \approx \log \tilde{p}_{\text{fix}}(U_b, s_b \to U_b', s_b') + \gamma \cdot \log \tilde{p}_{\text{fix}}(s_m|\mu), \tag{7}$$

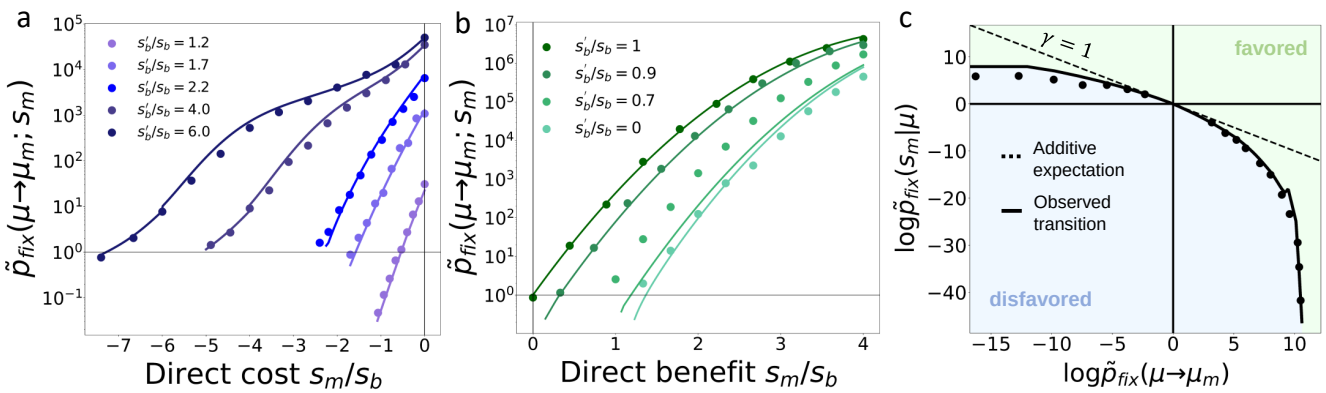

Fig 3. Trade-offs between first- and second-order selection. (a, b) Fixation probability of a selection-strength modifier with a direct cost (a) or benefit (b). Symbols denote the results of forward-time simulations for $s_b = 10^{-2}$, $U_b = U_b' = 10^{-5}$, and $N = 10^8$, while solid lines denote our theoretical predictions (SI Section 8.2). **(c)** Phase diagram illustrating the transition between favored (green) and unfavored modifiers (blue). Symbols denote results of forward time simulations for the selection-strength modifiers in panels a and b; x-coordinates show the measured fixation probabilities in the absence of a direct cost or benefit (y-intercepts in panels a and b), while the y-coordinates show the predicted fixation probability of a first-order mutation inferred from the x-intercept in panels a and b [$\tilde{p}_{\text{fix}}(\mu \to \mu_m; s_m) = 1$]. Solid lines show our theoretical predictions (SI Section 8.2), which exhibit large deviations from the additive expectation ($\gamma = 1$, dashed line).

where $\tilde{p}_{\text{fix}}(U_b, s_b \to U_b', s_b')$ is the fixation probability of the modifier from Eq. (6), $\tilde{p}_{\text{fix}}(s_m|\mu)$ is the fixation probability of a first-order mutation from Eq. (5), and γ is an additional weighting factor satisfying $\gamma \approx (s_b/s_b')\left[1 - (x_c s_b/v)^{-1}\log(U_b'/U_b)\right]$. This decomposition shows that differences in evolvability will generally modulate the effects of first-order selection on fitness.

A modifier that is strongly beneficial in the absence of a direct cost or benefit [$\tilde{p}_{\text{fix}}(\mu \to \mu_m; 0) \gg 1$] will always reduce the relative contribution from s_m ($\gamma \lesssim 1$). As a result, strongly beneficial modifiers can remain positively selected even when they impose a large direct cost on fitness ($|s_m| \gg s_b$; Fig. 3A,C). This contrasts with the classical behavior observed for discrete selective sweeps, where direct costs larger than s_b will typically prevent a modifier from fixing (SI Section 2). Larger populations are therefore better able to sacrifice short-term fitness for long-term gains in evolvability.

The opposite behavior occurs when short-term fitness benefits are linked to long-term reductions in evolvability. In this case, strongly deleterious modifiers [$\tilde{p}_{\text{fix}}(\mu \to \mu_m; 0) \ll 1$] will generally amplify the relative contribution of a short-term fitness benefit (Fig. 3B,C). A striking example of this effect occurs in the extreme case of an evolutionary "dead end", where further beneficial mutations are not available. These local fitness peaks cannot be captured by the linear approximation in Eq. (1), since their long-term success depends on their ability to influence the adaptation of the wildtype.

Remarkably, the strength of this feedback is still dominated by chance events that occur while the modifier lineage is rare. This allows us to generalize Eq. (1) to account for this non-linear behavior (SI Section 4.1.6). We find that moderate direct benefits as small as $\approx 0.4x_c$ are sufficient to cause an evolutionary dead-end to be positively selected [$\tilde{p}_{\text{fix}}(\mu \to \emptyset; s_m) \gg 1$], even though they cause the long-term rate of adaptation to drop to zero if they fix (Fig. 3B). This critical fitness benefit is often larger than the

size of a single driver mutation (s_b), but it is also smaller than the total fitness variation maintained within the population (x_c) and is only weakly dependent on NU_b. These examples illustrate how the evolutionary foresight of natural selection can be highly asymmetric: larger populations can greedily select for mutations that lower their rate of adaptation, even while they are better able to endure short-term fitness costs to realize long-term evolutionary benefits.

Continuous distributions of fitness effects

We have so far focused on a simplified model of the fitness landscape, where all beneficial mutations confer the same characteristic fitness benefit. However, most organisms produce mutations with a range of different fitness benefits. A general evolvability modifier will therefore correspond to a continuous perturbation of the distribution of fitness effects, $\delta\mu(s) = \mu_m(s) - \mu(s)$, representing the addition or subtraction of mutations with a range of fitness benefits (Fig. 4A). How does second-order selection on these more general evolvability differences relate to the selection-strength and mutation-rate and axes in Fig. 2?

We can extend our solution of Eq. (1) to a large class of wildtype mutation distributions that have been studied in previous work (SI Section 5). In these settings, the fitness benefits of fixed mutations are strongly peaked around a characteristic fitness benefit $s_b(\mu, N)$ [with a corresponding mutation rate $U_b(\mu, N)$], even when $\mu(s)$ is more broadly distributed [e.g. an exponential distribution] (39, 40, 47, 48; SI Section 5.1). In terms of these parameters, we find that selection on a general evolvability modifier [$\mu(s) \to \mu(s) + \delta\mu(s)$] sensitively depends on the integral,

$$I(\delta\mu) \equiv \int \frac{\delta\mu(s)\, ds}{U_b} \exp\left[\log \tilde{p}_{\text{fix}}(s|\mu_m, N^*) - \log \tilde{p}_{\text{fix}}(s_b|\mu_m, N^*)\right] , \tag{8}$$

where $p_{\text{fix}}(s|\mu, N)$ denotes the fixation probability of a first-order mutation from Eq. (5) and N^* is defined by $v(\mu_m, N^*) = v$ (SI Section 5.2). When $I(\delta\mu) \lesssim 1$, the fixation probability of the modifier initially grows as $\tilde{p}_{\text{fix}}(\delta\mu) \approx \exp\left[(x_c/s_b) \cdot I(\delta\mu)\right]$. This shows that the relevant fitness scale for second-order selection is set by s_b. For $s > s_b$, even small increases in the net mutation rate ($\int_{s_b}^{\infty} \delta\mu(s)\, ds \ll U_b$) can generate large changes in the fixation probability (Fig. 4B). In contrast, fitness benefits slightly smaller than s_b require many multiples of U_b to have the same effect. Since s_b and U_b both emerge from the competition between linked mutations, the location of this transition can vary with the population size and the shape of $\mu(s)$ (Fig. 4C).

For more strongly beneficial modifiers [$\log \tilde{p}_{\text{fix}}(\delta\mu) \gtrsim x_c/s_b$], the integrand in Eq. (8), is often concentrated around a characteristic fitness benefit $s_b'(\mu_m, N)$, which depends on the perturbation $\delta\mu(s)$ as well as the wildtype parameters s_b and U_b. In this case, we find that the fate of the modifier can be predicted from our single-s theory in Eq. (6),

$$\log \tilde{p}_{\text{fix}}(\delta\mu) \approx \log \tilde{p}_{\text{fix}}(U_b, s_b \to U_b', s_b') , \tag{9}$$

where $U_b' \propto \delta\mu(s_b')$ is the total mutation rate for mutations close to s_b' (Fig. 4B-E; SI Section 5). This effective selection coefficient coincides with the mutations most likely to fix in a successful modifier lineage (SI Section 5), providing a mechanistic interpretation for the effective parameters.

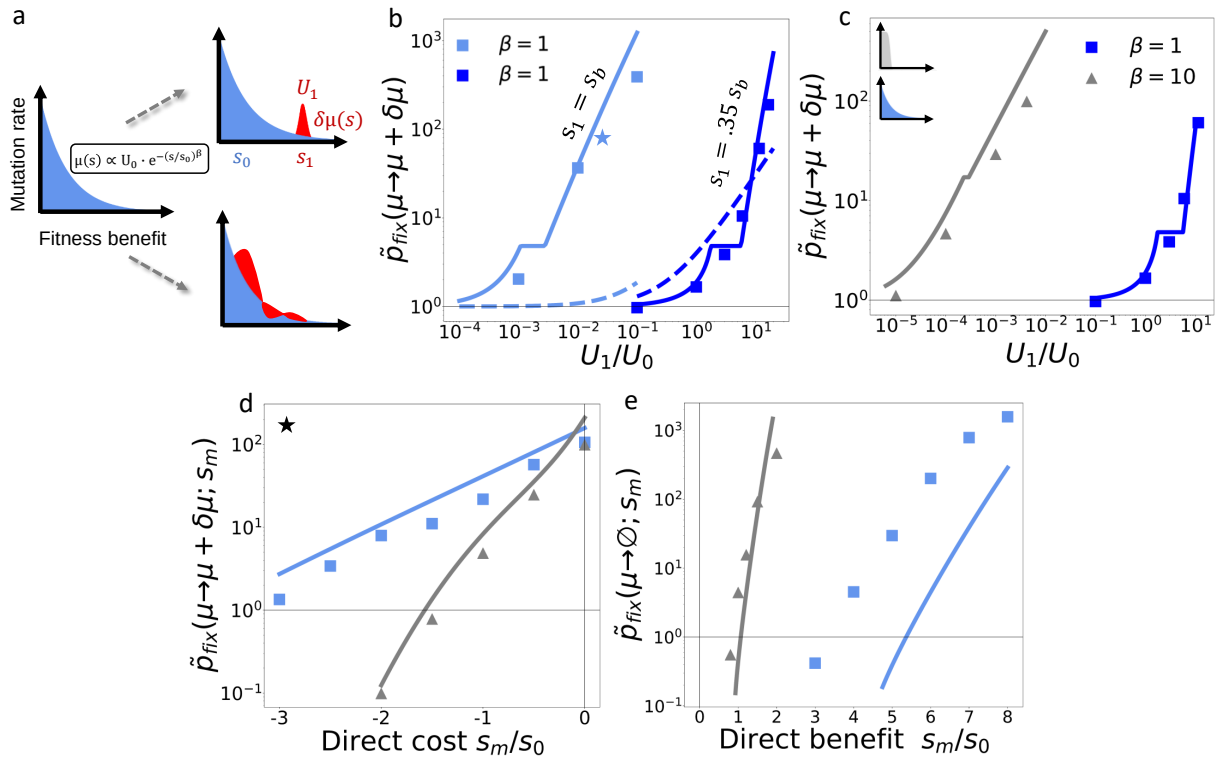

Fig 4. Second-order selection with a continuous distribution of fitness effects. (a) Examples of evolvability modifiers for a continuous distribution of fitness effects. Top right panel shows a small local addition $\delta\mu(s) \approx U_1\delta(s - s_1)$ (red), while bottom right panel shows a $\delta\mu(s)$ with a mixture of positive and negative changes. **(b, c)** Scaled fixation probability of the modifier in panel a (top right) for an exponential $\mu(s)$ ($\beta = 1$), as well as a distribution with a shorter tail ($\beta = 10$). Blue and gray curves in the inset of panel c illustrate differences in the shapes of these background distributions. In both panels, symbols denote the results of forward-time simulations for $s_0 = 10^{-2}$, $U_0 = 10^{-5}$, and $N = 10^8$. Solid lines denote our theoretical predictions (SI Section 8.2), while dashed lines indicate the null expectation in the absence of clonal interference. **(d)** Fixation probability of a modifier with a direct cost. Base parameters are indicated by the stars in panel c. **(e)** Fixation probability of an evolutionary dead-end with a short-term fitness benefit. Symbols denote the results of forward-time simulations for $s_0 = 10^{-2}$, $U_0 = 10^{-5}$, and $N = 10^8$, while solid lines denote our theoretical predictions.

The equivalence principle in Eq. (9) shows that second-order selection can generally be understood as a combination of the mutation-rate and selection-strength axes in Fig. 2. However, the relevant parameters will not coincide with the mean and height of $\mu(s)$. Instead, due to the exponential weighting of mutations with $s \gtrsim s_b$, otherwise subtle additions to $\mu(s)$ can be strongly favored by natural selection, even when they have a negligible impact on the overall mean or mutation rate (Fig. 4B-D). Conversely, large reductions in these global parameters will be invisible to natural selection unless they also deplete mutations near s_b. This sensitivity to local changes could help explain previous experimental observations in *E. coli* (Fig. 1A), where the potentiation of just a few beneficial genes was sufficient to overcome a large short-term fitness cost (8).

Discussion

Our results provide a framework for understanding how "first-order" selection for fitness interacts with "second-order" selection for evolvability in rapidly adapting populations. We have shown that when beneficial mutations are common, the competition between linked mutations can dramatically enhance selection for subtle differences in evolvability, leading to large deviations from the linear scaling predicted by classical evolutionary models (27). These results suggest that second-order selection could play a previously unappreciated role in driving the success of genetic variants in large microbial populations, from laboratory evolution experiments (8, 10, 13, 30, 49) to natural systems like cancer (14, 15), influenza (31), or SARS-CoV-2 (34, 50). This could have important consequences for evolutionary forecasting (32, 51–53), since it implies that the fitness effects of such variants might fail to explain their long-term evolutionary success.

Our theory indicates that it could be difficult to detect these evolvability differences using traditional metrics like the rate of adaptation (8, 54, 55) or the substitution rate (49), since the associated changes in these observables are not always large (SI Section 4.1.7). Previously documented cases like Fig. 1A might therefore only represent a fraction of the selectable variation in evolvability. Future efforts could instead focus on mapping the aggregate changes to the distribution of fitness effects (Fig. 1B), e.g. using barcoded lineage tracking (9, 56–58) or mutation trap experiments (59). However, we have also shown that the important changes in this distribution will often occur in its high-fitness tail, and are poorly captured by existing heuristics like the mean mutational effect (27, 60). They can also depend on external factors like the population size and the overall mutation rate. This suggests that the evolvability benefits of a mutation should not be viewed as an intrinsic property of the cell, but rather a collective effect that can vary across populations or within the same population over time. Our theory provides a framework for predicting where these evolutionarily important differences will occur.

While we have primarily focused on the supply of beneficial mutations, our results can also be extended to account for changes in the supply of strongly deleterious mutations ($|s| \gg v/x_c$; SI Section 6), which behave like an effective direct cost, $s_m^{\text{eff}} \approx \int_{-\infty}^{-v/x_c} \delta\mu(s)ds$ (Fig. S2). This mapping helps us understand how rapidly adapting populations balance the trade-offs (or synergies) between robustness and evolvability (61–63). It implies that rapidly adapting populations will be more willing to sacrifice short-term robustness for long-term gains in evolvability (Fig. 3A), but also that sufficiently large robustness gains can still be favored even if they eliminate the opportunities for future adaptation (Fig. 3B).

We have also assumed that the evolvability benefits of the modifier lineage remain fixed as it competes for dominance in the population. In reality, epistatic interactions could cause these benefits to attenuate – or even reverse – as the modifier acquires further mutations. Our heuristic analysis suggests that our current results will continue to hold as long as $\mu(s)$ and $\mu_m(s)$ remain constant over a typical fixation time (roughly $\sim v/x_c$ generations, or $\sim x_c/s_b$ additional mutations). This timescale is often modest in practice, allowing our minimal model to capture a broader range of epistatic scenarios than its idealized nature might originally suggest (Figs. S3 and S4; SI Section 7). Further extensions of this framework to allow for more rapidly varying distributions of fitness effects ("macroscopic epistasis"; 41, Fig. S1A) could be useful for understanding how large populations navigate rugged fitness landscapes (64).

Finally, while we have focused on the asexual dynamics common in laboratory experiments (8) and somatic evolution (14, 15), natural microbial populations often exhibit some degree of recombination (65). Widespread recombination will alter our predictions by decoupling the modifier locus from the future mutations that it produces (23, 66). Previous work suggests that some of our results may still apply on short genomic distances that remain tightly linked over the characteristic fixation time ($l \sim \frac{1}{r} \cdot \frac{v}{x_c}$) (67–69). However, modifiers can also benefit from transient linkage to mutations outside these asexual linkage blocks (70), and the fitness effects of the mutations could change when decoupled from the background that produced them (64). Understanding the interplay between these forces will be critical for understanding how second-order selection acts in other contexts.

References

1. Matic I, Radman M, Taddei F, Picard B, Doit C, Bingen E, et al. Highly Variable Mutation Rates in Commensal and Pathogenic Escherichia coli. Science. 1997;277:1833–1834. doi:10.1126/science.277.5333.1833.

2. LeClerc JE, Li B, Payne WL, Cebula TA. High Mutation Frequencies Among Escherichia coli and Salmonella Pathogens. Science. 1996;274:1208–1211. doi:10.1126/science.274.5290.1208.

3. Loeb LA. Human cancers express mutator phenotypes: origin, consequences and targeting. Nature Reviews Cancer. 2011;11:450–457. doi:10.1038/nrc3063.

4. Sane M, Diwan GD, Bhat BA, Wahl LM, Agashe D. Shifts in mutation spectra enhance access to beneficial mutations. Proceedings of the National Academy of Sciences. 2023;120:e2207355120. doi:10.1073/pnas.2207355120.

5. Sniegowski PD, Gerrish PJ, Lenski RE. Evolution of high mutation rates in experimental populations of E. coli. Nature. 1997;387:703–705. doi:10.1038/42701.

6. Couce A, Guelfo JR, Blázquez J. Mutational Spectrum Drives the Rise of Mutator Bacteria. PLOS Genetics. 2013;9:e1003167–.

7. Freddolino PL, Goodarzi H, Tavazoie S. Fitness Landscape Transformation through a Single Amino Acid Change in the Rho Terminator. PLOS Genetics. 2012;8:e1002744–.

8. Woods RJ, Barrick JE, Cooper TF, Shrestha U, Kauth MR, Lenski RE. Second-Order Selection for Evolvability in a Large Escherichia coli Population. Science. 2011;331:1433–1436. doi:10.1126/science.1198914.

9. Aggeli D, Li Y, Sherlock G. Changes in the distribution of fitness effects and adaptive mutational spectra following a single first step towards adaptation. Nature Communications. 2021;12:5193. doi:10.1038/s41467-021-25440-7.

10. Johnson MS, Gopalakrishnan S, Goyal J, Dillingham ME, Bakerlee CW, Humphrey PT, et al. Phenotypic and molecular evolution across 10,000 generations in laboratory budding yeast populations. eLife. 2021;10:e63910. doi:10.7554/eLife.63910.

11. Blount ZD, Borland CZ, Lenski RE. Historical contingency and the evolution of a key innovation in an experimental population of Escherichia coli. Proceedings of the National Academy of Sciences. 2008;105:7899–7906. doi:10.1073/pnas.0803151105.

12. Palmer ME, Moudgil A, Feldman MW. Long-term evolution is surprisingly predictable in lattice proteins. Journal of The Royal Society Interface. 2013;10:20130026. doi:10.1098/rsif.2013.0026.

13. Barrick JE, Kauth MR, Strelioff CC, Lenski RE. Escherichia coli rpoB Mutants Have Increased Evolvability in Proportion to Their Fitness Defects. Molecular Biology and Evolution. 2010;27:1338–1347. doi:10.1093/molbev/msq024.

14. Karlsson K, Przybilla MJ, Kotler E, Khan A, Xu H, Karagyozova K, et al. Deterministic evolution and stringent selection during preneoplasia. Nature. 2023;618:383–393. doi:10.1038/s41586-023-06102-8.

15. Colom B, Herms A, Hall MWJ, Dentro SC, King C, Sood RK, et al. Mutant clones in normal epithelium outcompete and eliminate emerging tumours. Nature. 2021;598:510–514. doi:10.1038/s41586-021-03965-7.

16. Fabre MA, de Almeida JG, Fiorillo E, Mitchell E, Damaskou A, Rak J, et al. The longitudinal dynamics and natural history of clonal haematopoiesis. Nature. 2022;606:335–342. doi:10.1038/s41586-022-04785-z.

17. Gifford DR, Furió V, Papkou A, Vogwill T, Oliver A, MacLean RC. Identifying and exploiting genes that potentiate the evolution of antibiotic resistance. Nature Ecology I& Evolution. 2018;2:1033–1039. doi:10.1038/s41559-018-0547-x.

18. Vogwill T, Kojadinovic M, MacLean RC. Epistasis between antibiotic resistance mutations and genetic background shape the fitness effect of resistance across species of Pseudomonas. Proceedings of the Royal Society B: Biological Sciences. 2016;283:20160151. doi:10.1098/rspb.2016.0151.

19. Cummins EA, Snaith AE, McNally A, Hall RJ. The role of potentiating mutations in the evolution of pandemic Escherichia coli clones. European Journal of Clinical Microbiology I& Infectious Diseases. 2021;doi:10.1007/s10096-021-04359-3.

20. Altenberg L, Liberman U, Feldman MW. Unified reduction principle for the evolution of mutation, migration, and recombination. Proceedings of the National Academy of Sciences. 2017;114:E2392–E2400. doi:10.1073/pnas.1619655114.

21. Wagner GP. Feedback selection and the evolution of modifiers. Acta Biotheoretica. 1981;30:79–102. doi:10.1007/BF00047674.

22. Payne JL, Wagner A. The causes of evolvability and their evolution. Nature Reviews Genetics. 2019;20:24–38. doi:10.1038/s41576-018-0069-z.

23. Pigliucci M. Is evolvability evolvable? Nature Reviews Genetics. 2008;9:75–82. doi:10.1038/nrg2278.

24. Good BH, Desai MM. Evolution of Mutation Rates in Rapidly Adapting Asexual Populations. Genetics. 2016;204:1249–1266. doi:10.1534/genetics.116.193565.

25. van Gestel J, Bareia T, Tenennbaum B, Co AD, Guler P, Aframian N, et al. Short-range quorum sensing controls horizontal gene transfer at micron scale in bacterial communities. Nature Communications. 2021;12:2324. doi:10.1038/s41467-021-22649-4.

26. Raynes Y, Wylie CS, Sniegowski PD, Weinreich DM. Sign of selection on mutation rate modifiers depends on population size. Proceedings of the National Academy of Sciences. 2018;115:3422–3427. doi:10.1073/pnas.1715996115.

27. Tuffaha MZ, Varakunan S, Castellano D, Gutenkunst RN, Wahl LM. Shifts in mutation bias promote mutators by altering the distribution of fitness effects. The American Naturalist. 2023;doi:10.1086/726010.

28. Lang GI, Rice DP, Hickman MJ, Sodergren E, Weinstock GM, Botstein D, et al. Pervasive genetic hitchhiking and clonal interference in forty evolving yeast populations. Nature. 2013;500:571–574. doi:10.1038/nature12344.

29. Levy SF, Blundell JR, Venkataram S, Petrov DA, Fisher DS, Sherlock G. Quantitative evolutionary dynamics using high-resolution lineage tracking. Nature. 2015;519:181–186. doi:10.1038/nature14279.

30. Good BH, McDonald MJ, Barrick JE, Lenski RE, Desai MM. The dynamics of molecular evolution over 60,000 generations. Nature. 2017;551:45–50. doi:10.1038/nature24287.

31. Strelkowa N, Lässig M. Clonal Interference in the Evolution of Influenza. Genetics. 2012;192:671–682. doi:10.1534/genetics.112.143396.

32. Zanini F, Brodin J, Thebo L, Lanz C, Bratt G, Albert J, et al. Population genomics of intrapatient HIV-1 evolution. eLife. 2015;4:e11282. doi:10.7554/eLife.11282.

33. Rochman ND, Wolf YI, Faure G, Mutz P, Zhang F, Koonin EV. Ongoing global and regional adaptive evolution of SARS-CoV-2. Proceedings of the National Academy of Sciences. 2021;118:e2104241118. doi:10.1073/pnas.2104241118.

34. Kemp SA, Collier DA, Datir RP, Ferreira IATM, Gayed S, Jahun A, et al. SARS-CoV-2 evolution during treatment of chronic infection. Nature. 2021;592:277–282. doi:10.1038/s41586-021-03291-y.

35. Barroso-Batista J, Sousa A, Lourenço M, Bergman ML, Sobral D, Demengeot J, et al. The First Steps of Adaptation of Escherichia coli to the Gut Are Dominated by Soft Sweeps. PLOS Genetics. 2014;10:e1004182–.

36. Zhao S, Lieberman TD, Poyet M, Kauffman KM, Gibbons SM, Groussin M, et al. Adaptive Evolution within Gut Microbiomes of Healthy People. Cell Host and Microbe. 2019;25:656–667.e8. doi:10.1016/j.chom.2019.03.007.

37. Lieberman TD, Flett KB, Yelin I, Martin TR, McAdam AJ, Priebe GP, et al. Genetic variation of a bacterial pathogen within individuals with cystic fibrosis provides a record of selective pressures. Nature Genetics. 2014;46:82–87. doi:10.1038/ng.2848.

38. Neher RA. Genetic Draft, Selective Interference, and Population Genetics of Rapid Adaptation. Annual Review of Ecology, Evolution, and Systematics. 2013;44:195–215. doi:10.1146/annurev-ecolsys-110512-135920.

39. Good BH, Rouzine IM, Balick DJ, Hallatschek O, Desai MM. Distribution of fixed beneficial mutations and the rate of adaptation in asexual populations. Proceedings of the National Academy of Sciences. 2012;109:4950–4955. doi:10.1073/pnas.1119910109.

40. Fisher DS. Asexual evolution waves: fluctuations and universality. Journal of Statistical Mechanics: Theory and Experiment. 2013;2013:P01011. doi:10.1088/1742-5468/2013/01/P01011.

41. Good BH, Desai MM. The Impact of Macroscopic Epistasis on Long-Term Evolutionary Dynamics. Genetics. 2015;199:177–190. doi:10.1534/genetics.114.172460.

42. Lenski RE, Rose MR, Simpson SC, Tadler SC. Long-Term Experimental Evolution in Escherichia coli. I. Adaptation and Divergence During 2,000 Generations. The American Naturalist. 1991;138:1315–1341. doi:10.1086/285289.

43. Neher RA, Shraiman BI, Fisher DS. Rate of Adaptation in Large Sexual Populations. Genetics. 2010;184:467–481. doi:10.1534/genetics.109.109009.

44. Birky CW, Walsh JB. Effects of Linkage on Rates of Molecular Evolution. Proceedings of the National Academy of Sciences of the United States of America. 1988;85:6414–6418.

45. Keightley PD, Otto SP. Interference among deleterious mutations favours sex and recombination in finite populations. Nature. 2006;443:89–92. doi:10.1038/nature05049.

46. Good BH, Desai MM. Deleterious Passengers in Adapting Populations. Genetics. 2014;198:1183–1208. doi:10.1534/genetics.114.170233.

47. Desai MM, Fisher DS. Beneficial Mutation–Selection Balance and the Effect of Linkage on Positive Selection. Genetics. 2007;176:1759–1798. doi:10.1534/genetics.106.067678.

48. Hegreness M, Shoresh N, Hartl D, Kishony R. An Equivalence Principle for the Incorporation of Favorable Mutations in Asexual Populations. Science. 2006;311:1615–1617. doi:10.1126/science.1122469.

49. Ba ANN, Cvijović I, Echenique JIR, Lawrence KR, Rego-Costa A, Liu X, et al. High-resolution lineage tracking reveals travelling wave of adaptation in laboratory yeast. Nature. 2019;575:494–499. doi:10.1038/s41586-019-1749-3.

50. Karthikeyan S, Levy JI, Hoff PD, Humphrey G, Birmingham A, Jepsen K, et al. Wastewater sequencing reveals early cryptic SARS-CoV-2 variant transmission. Nature. 2022;609:101–108. doi:10.1038/s41586-022-05049-6.

51. Luksza M, Lässig M. A predictive fitness model for influenza. Nature. 2014;507:57–61. doi:10.1038/nature13087.

52. Lee JM, Huddleston J, Doud MB, Hooper KA, Wu NC, Bedford T, et al. Deep mutational scanning of hemagglutinin helps predict evolutionary fates of human H3N2 influenza variants. Proceedings of the National Academy of Sciences. 2018;115:E8276–E8285. doi:10.1073/pnas.1806133115.

53. Luksza M, Sethna ZM, Rojas LA, Lihm J, Bravi B, Elhanati Y, et al. Neoantigen quality predicts immunoediting in survivors of pancreatic cancer. Nature. 2022;606:389–395. doi:10.1038/s41586-022-04735-9.

54. Kryazhimskiy S, Rice DP, Jerison ER, Desai MM. Global epistasis makes adaptation predictable despite sequence-level stochasticity. Science. 2014;344:1519–1522. doi:10.1126/science.1250939.

55. Wiser MJ, Ribeck N, Lenski RE. Long-Term Dynamics of Adaptation in Asexual Populations. Science. 2013;342:1364–1367. doi:10.1126/science.1243357.

56. Ascensao JA, Wetmore KM, Good BH, Arkin AP, Hallatschek O. Quantifying the local adaptive landscape of a nascent bacterial community. Nature Communications. 2023;14:248. doi:10.1038/s41467-022-35677-5.

57. Couce A, Magnan M, Lenski RE, Tenaillon O. Predictability shifts from local to global rules during bacterial adaptation. bioRxiv. 2023; p. 2022.05.17.492360. doi:10.1101/2022.05.17.492360.

58. Limdi A, Owen SV, Herren C, Lenski RE, Baym M. Parallel changes in gene essentiality over 50,000 generations of evolution. bioRxiv. 2022; p. 2022.05.17.492023. doi:10.1101/2022.05.17.492023.

59. Frenkel EM, Good BH, Desai MM. The Fates of Mutant Lineages and the Distribution of Fitness Effects of Beneficial Mutations in Laboratory Budding Yeast Populations. Genetics. 2014;196:1217–1226. doi:10.1534/genetics.113.160069.

60. Wagner A. Evolvability-enhancing mutations in the fitness landscapes of an RNA and a protein. Nature Communications. 2023;14:3624. doi:10.1038/s41467-023-39321-8.

61. Draghi JA, Parsons TL, Wagner GP, Plotkin JB. Mutational robustness can facilitate adaptation. Nature. 2010;463:353–355. doi:10.1038/nature08694.

62. Johnson MS, Martsul A, Kryazhimskiy S, Desai MM. Higher-fitness yeast genotypes are less robust to deleterious mutations. Science. 2019;366:490–493. doi:10.1126/science.aay4199.

63. Zheng J, Guo N, Wagner A. Selection enhances protein evolvability by increasing mutational robustness and foldability. Science. 2020;370:eabb5962. doi:10.1126/science.abb5962.

64. Agarwala A, Fisher DS. Adaptive walks on high-dimensional fitness landscapes and seascapes with distance-dependent statistics. Theoretical Population Biology. 2019;130:13–49. doi:https://doi.org/10.1016/j.tpb.2019.09.011.

65. Hanage WP. Not So Simple After All: Bacteria, Their Population Genetics, and Recombination. Cold Spring Harbor Perspectives in Biology. 2016;8. doi:10.1101/cshperspect.a018069.

66. Gardner A, Kalinka AT. Recombination and the evolution of mutational robustness. Journal of Theoretical Biology. 2006;241(4):707–715. doi:https://doi.org/10.1016/j.jtbi.2006.01.011.

67. Weissman DB, Hallatschek O. The Rate of Adaptation in Large Sexual Populations with Linear Chromosomes. Genetics. 2014;196:1167–1183. doi:10.1534/genetics.113.160705.

68. Neher RA, Kessinger TA, Shraiman BI. Coalescence and genetic diversity in sexual populations under selection. Proceedings of the National Academy of Sciences. 2013;110:15836–15841. doi:10.1073/pnas.1309697110.

69. Good BH, Walczak AM, Neher RA, Desai MM. Genetic Diversity in the Interference Selection Limit. PLOS Genetics. 2014;10:e1004222–.

70. Dawson KJ. Evolutionarily Stable Mutation Rates. Journal of Theoretical Biology. 1998;194:143–157. doi:https://doi.org/10.1006/jtbi.1998.0752.

71. Kondrashov FA, Kondrashov AS. Multidimensional epistasis and the disadvantage of sex. Proceedings of the National Academy of Sciences. 2001;98(21):12089–12092. doi:10.1073/pnas.211214298.

72. Johnson T. The approach to mutation–selection balance in an infinite asexual population, and the evolution of mutation rates. Proceedings of the Royal Society of London Series B: Biological Sciences. 1999;266:2389–2397. doi:10.1098/rspb.1999.0936.

73. Desai MM, Fisher DS. The Balance Between Mutators and Nonmutators in Asexual Populations. Genetics. 2011;188:997–1014. doi:10.1534/genetics.111.128116.

74. James A, Jain K. Fixation probability of rare nonmutator and evolution of mutation rates. Ecology and Evolution. 2016;6:755–764. doi:https://doi.org/10.1002/ece3.1932.

75. Kondrashov DA, Kondrashov FA. Topological features of rugged fitness landscapes in sequence space. Trends in Genetics. 2015;31:24–33. doi:https://doi.org/10.1016/j.tig.2014.09.009.

76. Leigh EG. The evolution of mutation rates. Genetics. 1973;73:Suppl 73:1–18.

77. Giles LE. Natural Selection and Mutability. The American Naturalist. 1970;104:301–305. doi:10.1086/282663.

78. Painter PR. Mutator genes and selecton for the mutaton rate in bacteria. Genetics. 1975;79:649–660. doi:10.1093/genetics/79.4.649.

79. Gillespie JH. Mutaton modification in a random environment. Evolution. 1981;35:468–476. doi:10.1111/j.1558-5646.1981.tb04910.x.

80. Ishii K, Matsuda H, Iwasa Y, Sasaki A. Evolutionarily stable mutation rate in a periodically changing environment. Genetics. 1989;121:163–174. doi:10.1093/genetics/121.1.163.

81. Kessler DA, Levine H. Mutator Dynamics on a Smooth Evolutionary Landscape. Physical Review Letters. 1998;80:2012–2015. doi:10.1103/PhysRevLett.80.2012.

82. Johnson T. Beneficial Mutations, Hitchhiking and the Evolution of Mutation Rates in Sexual Populations. Genetics. 1999;151:1621–1631. doi:10.1093/genetics/151.4.1621.

83. Tanaka MM, Bergstrom CT, Levin BR. The Evolution of Mutator Genes in Bacterial Populations: The Roles of Environmental Change and Timing. Genetics. 2003;164:843–854. doi:10.1093/genetics/164.3.843.

84. Andre JB, Godelle B. The Evolution of Mutation Rate in Finite Asexual Populations. Genetics. 2006;172:611–626. doi:10.1534/genetics.105.046680.

85. Wylie CS, Ghim CM, Kessler D, Levine H. The Fixation Probability of Rare Mutators in Finite Asexual Populations. Genetics. 2009;181:1595–1612. doi:10.1534/genetics.108.094532.

86. Gardiner C. Handbook of Stochastic Methods. New York: Springer; 1985.

87. Gerrish PJ, Lenski RE. The fate of competing beneficial mutations in an asexual population. Genetica. 1998;102:127–144. doi:10.1023/A:1017067816551.

88. Good BH, Desai MM. Fluctuations in fitness distributions and the effects of weak linked selection on sequence evolution. Theoretical population biology. 2013;85:86–102.

89. Reddy G, Desai MM. Global epistasis emerges from a generic model of a complex trait. eLife. 2021;10:e64740. doi:10.7554/eLife.64740.

90. Boffi NM, Guo Y, Rycroft CH, Amir A. How microscopic epistasis and clonal interference shape the fitness trajectory in a spin glass model of microbial long-term evolution. Elife. 2023;doi:10.7554/elife.87895.2.

91. Weissman DB, Desai MM, Fisher DS, Feldman MW. The rate at which asexual populations cross fitness valleys. Theoretical Population Biology. 2009;75:286–300. doi:https://doi.org/10.1016/j.tpb.2009.02.006.

92. Ochs IE, Desai MM. The competition between simple and complex evolutionary trajectories in asexual populations. BMC Evolutionary Biology. 2015;15:55. doi:10.1186/s12862-015-0334-0.

93. Merrikh CN, Merrikh H. Gene inversion potentiates bacterial evolvability and virulence. Nature Communications. 2018;9:4662. doi:10.1038/s41467-018-07110-3.

94. Horton JS, Flanagan LM, Jackson RW, Priest NK, Taylor TB. A mutational hotspot that determines highly repeatable evolution can be built and broken by silent genetic changes. Nature Communications. 2021;12:6092. doi:10.1038/s41467-021-26286-9.

95. Bloom JD, Gong LI, Baltimore D. Permissive Secondary Mutations Enable the Evolution of Influenza Oseltamivir Resistance. Science. 2010;328:1272–1275. doi:10.1126/science.1187816.

96. Javanmardi K, Segall-Shapiro TH, Chou CW, Boutz DR, Olsen RJ, Xie X, et al. Antibody escape and cryptic cross-domain stabilization in the SARS-CoV-2 Omicron spike protein. Cell Host and Microbe. 2022;30:1242–1254.e6. doi:https://doi.org/10.1016/j.chom.2022.07.016.

97. Rodrigues JV, Bershtein S, Li A, Lozovsky ER, Hartl DL, Shakhnovich EI. Biophysical principles predict fitness landscapes of drug resistance. Proceedings of the National Academy of Sciences. 2016;113:E1470–E1478. doi:10.1073/pnas.1601441113.

98. Li Y, Roberts ND, Wala JA, Shapira O, Schumacher SE, Kumar K, et al. Patterns of somatic structural variation in human cancer genomes. Nature. 2020;578:112–121. doi:10.1038/s41586-019-1913-9.

99. Venkataram S, Dunn B, Li Y, Agarwala A, Chang J, Ebel ER, et al. Development of a Comprehensive Genotype-to-Fitness Map of Adaptation-Driving Mutations in Yeast. Cell. 2016;166:1585–1596.e22. doi:https://doi.org/10.1016/j.cell.2016.08.002.

100. Tsimring LS, Levine H, Kessler DA. RNA virus evolution via a fitness-space model. Physical review letters. 1996;76(23):4440.

101. Rouzine IM, Wakeley J, Coffin JM. The solitary wave of asexual evolution. Proceedings of the National Academy of Sciences. 2003;100:587–592. doi:10.1073/pnas.242719299.

102. Hallatschek O. The noisy edge of traveling waves. Proceedings of the National Academy of Sciences. 2011;108:1783–1787. doi:10.1073/pnas.1013529108.

103. Melissa MJ, Good BH, Fisher DS, Desai MM. Population genetics of polymorphism and divergence in rapidly evolving populations. Genetics. 2022;221:iyac053. doi:10.1093/genetics/iyac053.

104. Neher RA, Shraiman BI. Genetic Draft and Quasi-Neutrality in Large Facultatively Sexual Populations. Genetics. 2011;188:975–996. doi:10.1534/genetics.111.128876.

105. Schiffels S, Szöllősi GJ, Mustonen V, Lässig M. Emergent Neutrality in Adaptive Asexual Evolution. Genetics. 2011;189:1361–1375. doi:10.1534/genetics.111.132027.

106. Sung W, Ackerman MS, Miller SF, Doak TG, Lynch M. Drift-barrier hypothesis and mutation-rate evolution. Proceedings of the National Academy of Sciences. 2012;109:18488–18492. doi:10.1073/pnas.1216223109.

107. Virtanen P, Gommers R, Oliphant TE, Haberland M, Reddy T, Cournapeau D, et al. SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. Nature Methods. 2020;17:261–272. doi:10.1038/s41592-019-0686-2.

Acknowledgments

We thank Daniel Wong for useful discussions, and Sophie Walton, Olivia Ghosh, and Zhiru Liu for comments and feedback on the manuscript. This work was supported in part by the Alfred P. Sloan Foundation (FG-2021-15708) and a Terman Fellowship from Stanford University. B.H.G. is a Chan Zuckerberg Biohub – San Francisco Investigator.

Author contributions: Conceptualization: J.T.F and B.H.G.; theory and methods development: J.T.F and B.H.G..; analysis: J.T.F and B.H.G.; writing: J.T.F. and B.H.G.

Competing interests: None declared.

Data and materials availability: Source code for forward-time simulations, numerical calculations, and figure generation are available at Github (`https://github.com/bgoodlab/evolution_of_evolvability`).

Supplementary Information

Contents

Page

1 Comparison to deterministic modifier theory **7**

2 Evolvability modifiers in the successive mutations regime **8**

3 Evolvability modifiers in the clonal interference regime **11**
3.1 Macroscopic epistasis expansion . . . 11
3.2 Fitness wave formalism . . . 15
3.3 Asymptotic solution in the "sharp shoulder" regime . . . 17

4 Solution for a simple model of the distribution of fitness effects **20**
4.1 Multiple mutations regime . . . 21
4.1.1 Location of the interference threshold . . . 21
4.1.2 Extending the shoulder solution to lower fitness values . . . 22
4.1.3 First-order mutations and the rate of adaptation . . . 23
4.1.4 Modifiers without direct costs or benefits . . . 26
4.1.5 Modifiers with direct costs or benefits . . . 28
4.1.6 Fixation of a dead-end modifier . . . 31
4.1.7 Relation between the fixation probability and the long-term rate of adaptation . . . 34
4.2 Quasi-sweep regime . . . 34
4.2.1 Location of the interference threshold . . . 34
4.2.2 Extending the shoulder solution to lower fitness values . . . 35
4.2.3 Fixation probabilities of modifiers . . . 37

5 Extension to continuous distributions of fitness effects **40**
5.1 Location of the interference threshold for the wildtype population . . . 41
5.2 Perturbative regime . . . 43
5.2.1 Location of the interference threshold . . . 43
5.2.2 Fixation probabilities of modifiers . . . 45
5.3 Modifier-dominated regime (multiple mutations) . . . 45
5.3.1 Location of the interference threshold . . . 46
5.3.2 Extending the shoulder solution to lower fitness values. . . . 46
5.3.3 Fixation probabilities of modifiers . . . 51
5.4 Modifier-dominated regime (quasi-sweeps) . . . 52
5.4.1 Location of the interference threshold . . . 52
5.4.2 Extending the shoulder solution to lower fitness values . . . 52

6 Incorporating deleterious mutations **54**

7 Extensions to more general fitness landscapes **56**
7.1 Weak Macroscopic Epistasis 56
7.2 Transient differences in evolvability 56
7.3 Global diminishing returns epistasis 57

8 Forward-Time Simulations and Numerical Methods **59**
8.1 Forward-time simulations 59
8.2 Theoretical predictions 60
8.3 Empirical example from Ref. (8) 62

List of Figures

S1 Examples of epistatic fitness landscapes that satisfy the minimal model in Fig. 1 3
S2 Deleterious mutations and second-order selection for robustness 4
S3 Relaxing the assumption that modifiers permanently change the mutation spectrum 5
S4 Selection for evolvability in the presence of diminishing returns epistasis 6

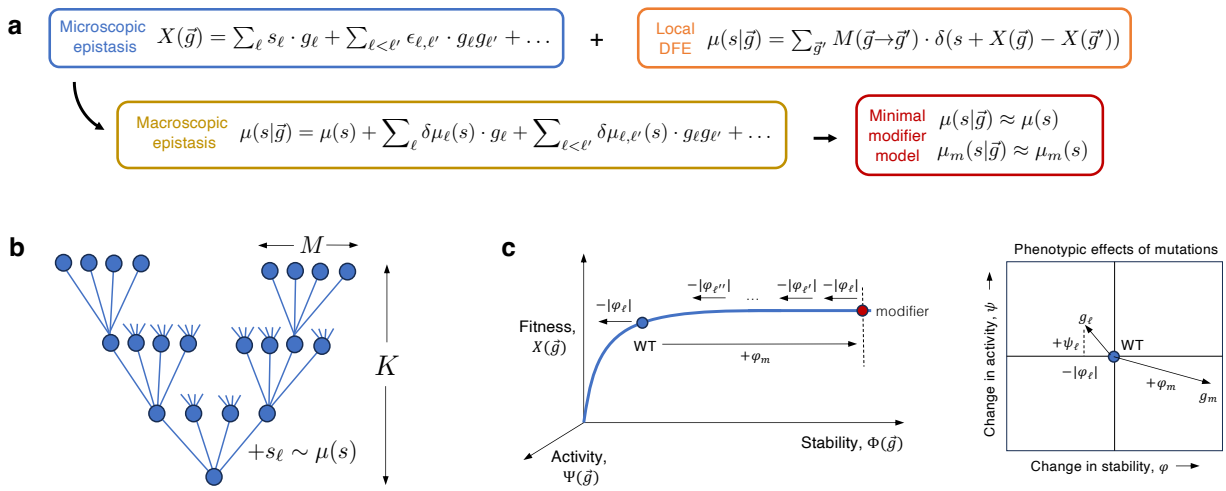

Fig S1. Examples of epistatic fitness landscapes that satisfy the minimal model in Fig. 1. (a) The evolvability modifier in Fig. 1B can be viewed as the lowest order term in a general macroscopic epistasis expansion (left, SI Section 3.1). Different fitness landscapes can produce the same macroscopic behavior. (b,c) Examples of highly epistatic fitness landscapes that satisfy the simple model above. (b) A "maximally epistatic" landscape of branching uphill paths, which generalizes the model in Ref. (71). Each step $k = 1, \ldots, K$ of a given path can access $M \ll L$ beneficial mutations; all other genotypes have fitness zero. (c) A fitness landscape formed by a non-linear combination of two global phenotypes, e.g., stability, $\Phi(\vec{g})$, and activity, $\Psi(\vec{g})$. Individual mutations can affect both traits simultaneously (right). Stabilizing mutations can act like modifier alleles by potentiating the fitness benefits of mutations that would destabilize the protein on their own (left). In particular, a strongly stabilizing mutation can allow $K \approx \phi_m/|\phi_\ell|$ new mutations to accumulate before their effects on stability become important. See SI Section 3.1 for more details.

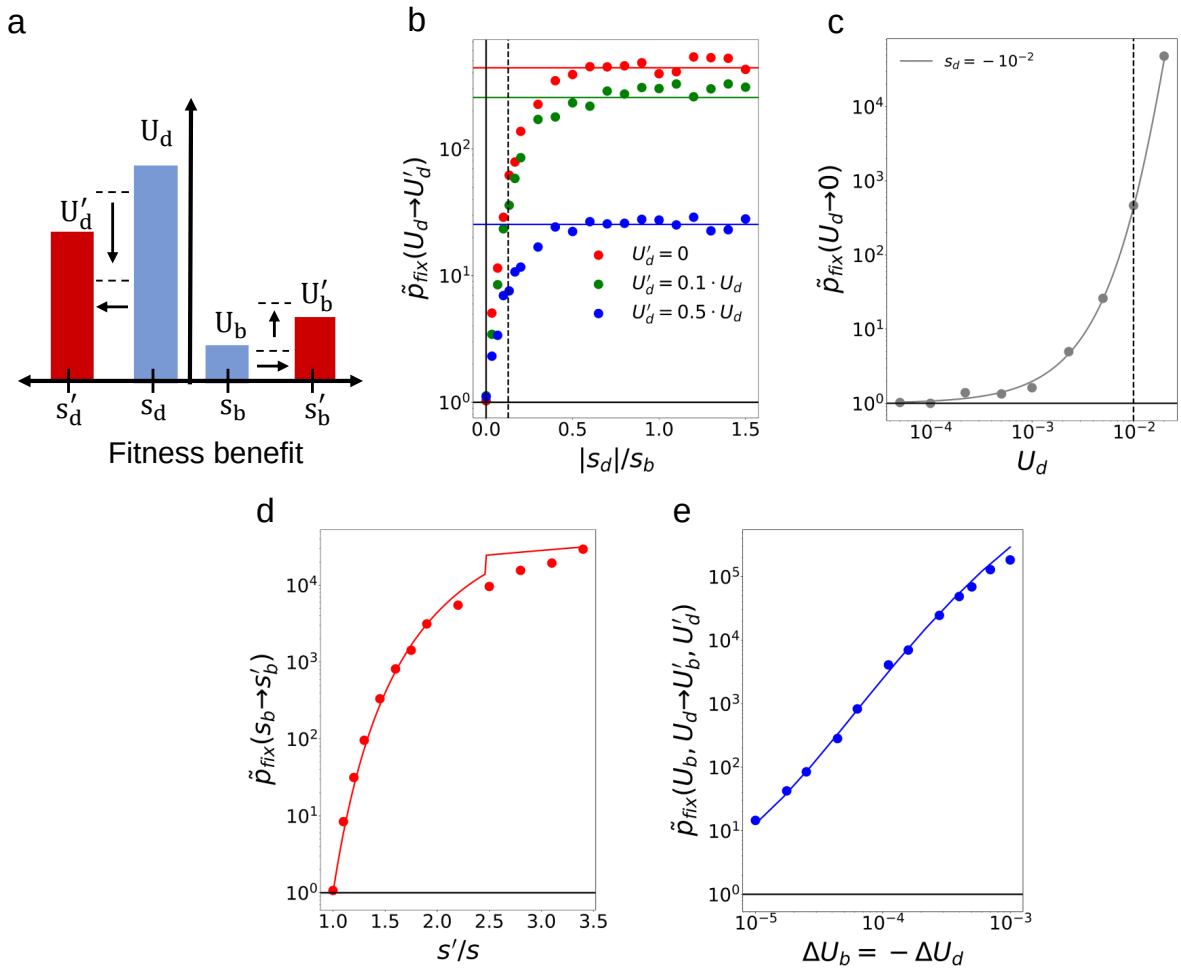

Fig S2. Deleterious mutations and second-order selection for robustness. (a) A generalization of the simplified model in Fig. 2A that incorporates deleterious mutations with a characteristic cost s_d and rate U_d. A modifier mutation shifts these parameters to a new combination, $s_b, U_b, s_d, U_d \to s'_d, U'_b, s'_d, U'_d$. **(b, c)** Fixation probability of a robustness-enhancing modifier with $U'_d < U_d$ (and all other parameters are held fixed). Symbols denote the results of forward time simulations for $N = 10^8$, $s_b = 10^{-2}$, and $U_b = 10^{-5}$, while solid lines denote our theoretical predictions in SI Section 6. Panel (b) shows that the purgeable mutations approximation holds across a broad range of fitness costs, with the dashed line marking the predicted transition to quasi-neutrality ($|s_d| \approx v/x_c$). Panel (c) shows that selection for increased robustness is relatively weak unless $U_d \gtrsim s_b$ (dashed line). **(d)** Fixation probability of a modifier that imposes a tradeoff between robustness and evolvability by increasing the strength of selection on beneficial and deleterious mutations simultaneously. Results are shown for $s_b = |s_d| = s$ and $U_d = 10^{-2}$, with the remaining parameters the same as panel b. Since $U_d \gg U_b$, this example shows that strong selection for evolvability can occur for modifiers reduce the average fitness effect of mutations ($\Delta\bar{s} \propto U_b \Delta s - U_d \Delta s < 0$). **(e)** Fixation probability of modifier that enhances robustness and evolvability at the same, by shifting mutations from deleterious to beneficial ($U_d - U'_d = U'_b - U_b$). Symbols denote results of forward-time simulations with $s_d = -10^{-2}$ and $U_d = 10^{-2}$, with the remaining parameters the same as panel (b). Lines denote our theoretical predictions in the absence of deleterious mutations ($U_d = U'_d = 0$). This example shows thatenhancements in evolvability are weighted more strongly than comparable increases in robustness, even when nearly all new mutations are deleterious ($U_b \ll U_d$).

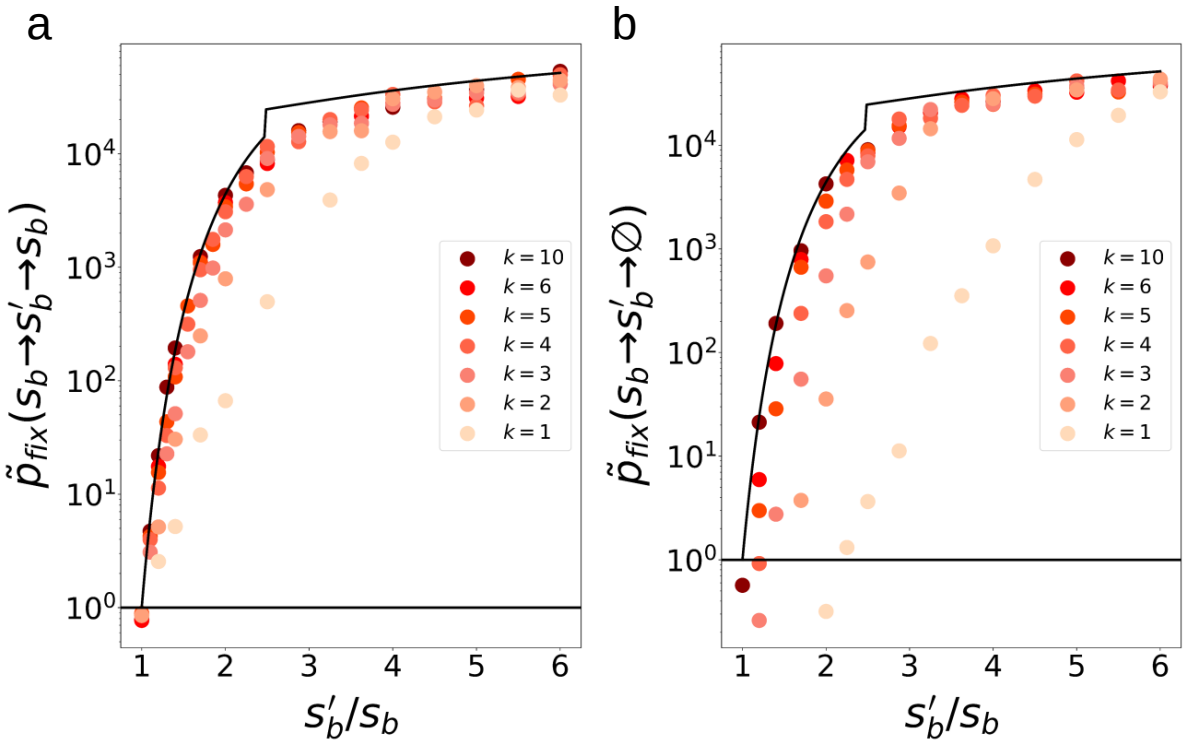

Fig S3. Relaxing the assumption that modifiers permanently change the mutation spectrum. (a) A generalization of the model in Fig. 2B, where the modifier reverts back to the wildtype distribution of fitness effects after acquiring K additional mutations (SI Section 7.2). Symbols denote the results of forward-time simulations for $N = 10^8$, $s_b = 10^{-2}$, and $U_b = 10^{-5}$, while the line denotes our theoretical predictions for the minimal modifier model in Fig. 1B (i.e. $K = \infty$). Consistent with our calculations in SI Section 7.2, these results show that the $K \approx \infty$ approximation remains highly accurate even for moderate values of K (e.g. 2-3), and for as little as a single mutation ($K = 1$) in the quasi-sweeps regime.**(b)** An alternative version of the model in panel a, where the modifier reverts to an evolutionary dead-end ($\mu_m = \emptyset$) after K mutations; all other parameters are the same as panel a. Even in this extreme case, our minimal modifier model (solid line) remains highly accurate for moderate values of K, and as little as $K = 1$ in the quasi-sweeps regime. This demonstrates that large populations can only "see" across the fitness landscape for $\sim x_{cm}/s'_b$ additional mutations (SI Section 7.2).

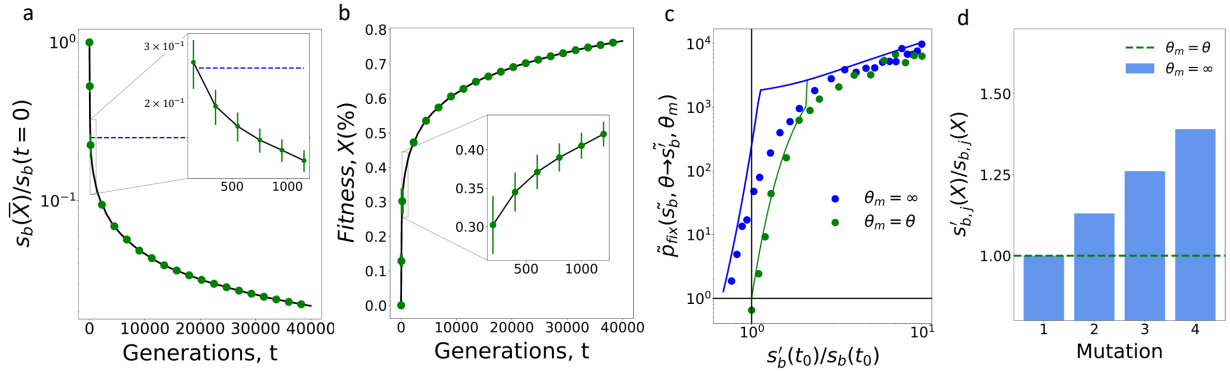

Fig S4. Selection for evolvability in the presence of diminishing returns epistasis. (a, b) A simple model of global diminishing returns epistasis motivated by the empirical example in Ref. (55) (SI Section 7.3). The fitness effects of new mutations shrink as the population adapts (panel a), leading to a decelerating rate of adaptation over time (panel b). Points denote the results of forward-time simulations for the distribution of fitness effects $\mu(s|\vec{g}) = U_b \cdot \delta(s - \tilde{s}_b \cdot e^{-X(\vec{g})/\theta})$, with $U_b = 10^{-5}$, $\tilde{s}_b = 10^{-1}$, $\theta = 0.2$, and $N = 10^7$; points are connected by solid lines to aid visualization. **(c, d)** The fixation probability of an evolvability modifier that arises at the beginning of the inset in panel a, where the fitness trajectory is still decelerating. Green symbols in (c) show a selection-strength modifier with the same diminishing returns schedule as the background population ($\theta_m \approx \theta$), while the blue symbols show an alternate example where the modifier avoids future diminishing returns once it arises ($\theta_m \approx \infty$). The green line illustrates the predictions from the "adiabatic" approximation in SI Section 7.3, demonstrating that the permanent modifier model [$\mu_m(s|\vec{g}) \approx \mu_m(s)$] provides a good approximation when the local selection strengths are properly renormalized. The blue line shows the predictions from our heuristic analysis in SI Section 7.3, which accounts for the additional benefits that accrue for the modifier lineage when $\theta_m \gg \theta$ (panel d). This example illustrates that the evolvability advantages that accrue from large differences in diminishing returns epistasis can drive modest deviations from our existing theory when θ grows close to x_c.

1 Comparison to deterministic modifier theory

In this section, we rederive a key result from classical modifier theory known as the "mean fitness principle" (20). This theory predicts that in an infinitely large asexual population, natural selection will favor modifiers that increase the long-term mean fitness of the population. Our derivation closely follows the one given in Ref. (21), as well as related work on mutators (72–74). We reproduce it here for completeness, using the same notation employed in our more general analysis below.

To establish this result, we note that in the absence of genetic drift ($N = \infty$), the deterministic dynamics of a well-mixed asexual population can be written in the general form,

$$\frac{\partial f(\vec{g})}{\partial t} = \left[X(\vec{g}) - \overline{X}(t)\right] f(\vec{g}) + \sum_{\vec{g}'} M(\vec{g}' \to \vec{g}) f(\vec{g}') - \sum_{\vec{g}} M(\vec{g} \to \vec{g}') f(\vec{g}) \,, \tag{S1}$$

where $f(\vec{g}, t)$ is the frequency of genotype $\vec{g}$, $X(\vec{g})$ is the (log) fitness of genotype $\vec{g}$, $\overline{X}(t) = \sum_{\vec{g}} X(\vec{g}) f(\vec{g}, t)$ is the mean fitness of the population, and $M(\vec{g} \to \vec{g}')$ is the mutation rate from genotype $\vec{g}$ to $\vec{g}'$. The fate of a general evolvability modifier allele can be analyzed by introducing an analogous set of equations,

$$\frac{\partial f_m(\vec{g})}{\partial t} = \left[X_m(\vec{g}) - \overline{X}(t)\right] f_m(\vec{g}) + \sum_{\vec{g}'} M_m(\vec{g}' \to \vec{g}) f_m(\vec{g}') - \sum_{\vec{g}} M_m(\vec{g} \to \vec{g}') f_m(\vec{g}) \,, \tag{S2}$$

where $f_m(\vec{g}, t)$ is the frequency of individuals with the mutant allele at the modifier locus and genotype $\vec{g}$ elsewhere, $X_m(\vec{g})$ is genotype-to-fitness map for the modifier, $M_m(\vec{g} \to \vec{g}')$ is the corresponding set of mutation rates, and the mean fitness in Eqs. (S1) and (S2) now sums over both the mutant and wildtype lineages,

$$\overline{X}(t) = \sum_{\vec{g}} X(\vec{g}) f(\vec{g}, t) + \sum_{\vec{g}} X_m(\vec{g}) f_m(\vec{g}, t) \,. \tag{S3}$$

In principle, this model allows for arbitrary changes to arbitrary non-linear fitness landscapes, encapsulating all possible forms of epistasis (75).

The solutions to this coupled system of equations can be written in the general form

$$f_m(\vec{g}, t) = f_m(t) h_m(\vec{g}, t) \,, \quad f(\vec{g}, t) = [1 - f_m(t)] h_w(\vec{g}, t) \,, \tag{S4a}$$

where $f_m(t)$, $h_m(t)$, and $h_w(t)$ satisfy the related set of equations,

$$\frac{\partial h_w(\vec{g})}{\partial t} = \left[X(\vec{g}) - \overline{X}_w(t)\right] h_w(\vec{g}) + \sum_{\vec{g}'} M(\vec{g}' \to \vec{g}) h_w(\vec{g}') - \sum_{\vec{g}} M(\vec{g} \to \vec{g}') h_w(\vec{g}) \,, \tag{S4b}$$

$$\frac{\partial h_m(\vec{g})}{\partial t} = \left[X_m(\vec{g}) - \overline{X}_m(t)\right] h_m(\vec{g}) + \sum_{\vec{g}'} M_m(\vec{g}' \to \vec{g}) h_m(\vec{g}') - \sum_{\vec{g}} M_m(\vec{g} \to \vec{g}') h_m(\vec{g}) \,, \tag{S4c}$$

and

$$\frac{\partial f_m}{\partial t} = \left[\overline{X}_m(t) - \overline{X}_w(t)\right] f_m(1 - f_m) \,, \tag{S4d}$$

with

$$\overline{X}_w(t) \equiv \sum_{\vec{g}} X(\vec{g}) h_w(\vec{g}, t) \,, \quad \overline{X}_m(t) \equiv \sum_{\vec{g}} X_m(\vec{g}) h_m(\vec{g}, t) \,. \tag{S4e}$$

In this notation, $h_m(\vec{g},t)$ represents the re-normalized genotype distribution within the modifier lineage, $h_w(\vec{g},t)$ denotes the corresponding distribution within the wildtype, $\overline{X}_m(t)$ and $\overline{X}_w(t)$ represent the mean fitnesses of each lineage, and $f_m(t)$ denotes the total frequency of the modifier.

This change-of-variables shows that the total frequency of the modifier lineage only depends on the relative values of the mean fitnesses, $\overline{X}_m(t)$ and $\overline{X}_w(t)$, yielding the time-dependent solution,

$$\log\left[\frac{f_m(t)}{1-f_m(t)}\right] = \log\left[\frac{f_m(0)}{1-f_m(t)}\right] + \int_0^t \left[\overline{X}_m(t') - \overline{X}_w(t')\right] dt' \,. \tag{S5}$$

This shows that at long times ($t \to \infty$) the lineage that will dominate the population is the one with the higher value of $\int_0^\infty \overline{X}_i(t')dt'$. These mean fitnesses can be predicted from the dynamics of the intra-lineage frequencies, $h_m(\vec{g},t)$ and $h_w(\vec{g},t)$, which decouple from each other and from the total size of $f_m(t)$. In most cases of interest, the mean fitness of each lineage will approach an equilibrium value, $\overline{X}_i(t) \to \overline{X}_i$, which may be a complicated function of $X_i(\vec{g})$ and $M_i(\vec{g} \to \vec{g}')$, but is otherwise independent of the surrounding population. Equation (S5) then shows that the modifier will take over the population if and only if it increases the equilibrium mean fitness. This generalizes the "mean fitness principle" derived in previous work (20).

However, this deterministic calculation neglects two key factors that are relevant for any large but finite population. First, the mean-field dynamics in Eqs. (S1) and (S2) neglect the random occurrence of new mutations and the stochastic fluctuations they experience while rare. These fluctuations can dramatically influence the dynamics of the mean fitness – particularly when multiple beneficial mutations are available (47). In addition, the deterministic calculation neglects the possibility that the mutant or wildtype lineage may fix before their long-term benefits in Eq. (S5) are fully realized. As we will see below, both of these effects will become extremely important in the adapting populations that we analyze in this work. Interestingly, we will see that in some cases, it will be possible to account for these effects in an approximate manner by inserting a upper limit on the integral in Eq. (S5) (SI Section 4.1.7), providing a conceptual link between classical modifier theory and the more complex scenarios studied in this work. Identifying such cases and their appropriate time horizons is the goal of the next several sections.

2 Evolvability modifiers in the successive mutations regime

Another useful limit occurs in small populations, where the production of beneficial mutations is sufficiently rare that adaptation proceeds via a sequence of discrete selective sweeps. In this regime, the fate of a given modifier will strongly depend on its ability to generate the next beneficial mutation. We can formalize this idea by defining the local distribution of fitness effects (DFE),

$$\mu(s|\vec{g}) = \sum_{\vec{g}'} M(\vec{g} \to \vec{g}') \cdot \delta(s - X(\vec{g}') + X(\vec{g}))\,, \tag{S6}$$

which tabulates the fitness effects of all the mutations that can be accessed from a given genotype $\vec{g}$. Modifier individuals will have their own corresponding set of DFEs,

$$\mu_m(s|\vec{g}) = \sum_{\vec{g}'} M_m(\vec{g} \to \vec{g}') \cdot \delta(s - X_m(\vec{g}') + X_m(\vec{g}))\,, \tag{S7}$$

along with a direct cost or benefit

$$s_m(\vec{g}) = X_m(\vec{g}) - X(\vec{g})\,. \tag{S8}$$

The distributions in Eqs. (S6) and (S7) can be computed for any epistatic fitness landscape, and will constitute the key input parameters in our analysis below.

In the absence of any modifier mutations, the evolutionary dynamics of the successive sweeps regime are well described by previous work (47). The wildtype population produces new beneficial mutations at a total rate $NU_b = N \cdot \int_0^\infty \mu(s|\vec{g})ds$ per generation. Most of these mutations will drift to extinction before they reach appreciable frequencies. However, with probability $p_{\text{est}} \approx 2s$, a lucky mutant will fluctuate to a sufficiently large frequency ($f \approx 1/2Ns$) that it begins to grow deterministically [$\partial_t f \approx sf(1-f)$], and will sweep through the population on a timescale of order $T_{\text{fix}} = \frac{2}{s}\log(Ns)$. The population will produce these successful mutations at a total rate $\lambda = \int_0^\infty N\mu(s|\vec{g}) \cdot 2s = 2NU_b\bar{s}$ per generation, where $\bar{s} \equiv \int s\mu(s|\vec{g})\, ds / \int \mu(s|\vec{g})ds$ is the average beneficial fitness effect that is accessible to the current genotype. This implies that the typical waiting time for the next sweep event is $T_c \approx 1/2NU_b\bar{s}$ generations. The condition that successive sweeps will not interfere with each requires that $T_{\text{fix}} \ll T_c$, which requires that $NU_b \log[N\bar{s}] \ll 1$ (47). This limit is also known as the "strong selection, weak mutation regime" (SSWM).

A general evolvability modifier ($\mu \to \mu_m$) will produce a change in the beneficial mutation rate ($U_b \to U_b'$) as well as the typical fitness benefit ($\bar{s} \to \bar{s}'$). (It may also change the spectrum of deleterious mutations, which we neglect for the time being; see SI Section 6). The mutator version of this model was previously analyzed in Ref. (24), as well as a number of related studies (27, 73, 76–85). We reproduce these calculations below, while generalizing them to allow for changes in the average fitness benefits of mutations (similar to Ref. 27) as well as a broader range of direct costs and benefits. These classical results will provide a useful baseline for understanding the impact of larger populations, which we analyze in SI Section 3 below.

In the successive mutations regime, a newly arising modifier lineage will initially compete with the wildtype population according to the single-locus dynamics

$$\frac{\partial f_m}{\partial t} = s_m f_m(1-f_m) + \sqrt{\frac{f_m(1-f_m)}{N}}\eta(t)\,, \tag{S9}$$

where $\eta(t)$ is a Browian noise term (86) and s_m is the direct cost or benefit of the modifier. The next selective sweep will now be generated by a pair of competing Poisson processes with rates

$$\lambda_m(t) = 2NU_b'\bar{s}' \cdot f_m(t)\,, \tag{S10a}$$

$$\lambda_w(t) = 2NU_b\bar{s} \cdot [1 - f_m(t)]\ , \tag{S10b}$$

which correspond to the mutant and wildtype lineages, respectively. Since the mutation rate and fitness benefit both contribute linearly to $\lambda_i(t)$, this process is formally equivalent to the mutator scenario analyzed in (24), with $r = U_b'\bar{s}'/U_b\bar{s}$ replacing U_b'/U_b. This allows us to conclude that in absence of a direct cost ($s_m = 0$), the fixation probability of the modifier scales as

$$\tilde{p}_{\text{fix}}(\mu \to \mu_m) \approx \begin{cases} U_b'\bar{s}'/U_b\bar{s} & \text{if } \frac{U_b'\bar{s}'}{U_b\bar{s}} \ll NU_b \cdot N\bar{s}, \\ \sqrt{2\left(\frac{U_b'\bar{s}'}{U_b\bar{s}}\right)NU_b \cdot N\bar{s}} & \text{if } \frac{U_b'\bar{s}'}{U_b\bar{s}} \gg NU_b \cdot N\bar{s}, \end{cases} \tag{S11}$$

provided that $T_c \ll N$ (or $NU_b \cdot N\bar{s} \gg 1$). When $NU_b \cdot N\bar{s} \lesssim 1$, the modifier will either fix or go extinct neutrally ($p_{\text{fix}} \approx 1/N$) before the next sweep occurs. This shows that second-order selection is more efficient in larger populations, which is reminiscent of our results in Fig. 2. In this case, however, the benefits of second-order selection are capped by the ratio $U_b'\bar{s}'/U_b\bar{s}$, which implies that very large changes in U_b' or $\bar{s}'$ are required to produce an appreciable fixation probability.

A similar application of Ref. (24) shows that the fixation probability of a modifier with a direct cost ($s_m < 0$) scales as

$$\tilde{p}_{\text{fix}}\left(\mu \to \mu_m, s_m\right) \approx \frac{\frac{\int_{|s_m|}^{\infty}(s-s_m)\mu_m(s)ds}{U_b\bar{s}}}{1+\frac{|s_m|}{2NU_b\bar{s}}}, \tag{S12}$$

provided that $U_b'\bar{s}'/U_b\bar{s} \ll NU_b \cdot N\bar{s}$. The integral in the numerator implies that modifiers with direct costs larger than $\sim\bar{s}'$ will have dramatically reduced fixation probabilities, since future beneficial mutations in these backgrounds will still be less fit than the current wildtype population. However, since $NU_b \ll 1$, the fixation probability of the modifier will be significantly reduced by the denominator term well before these "wasted opportunities" start to become relevant. This illustrates how even small direct costs – much smaller than a single driver mutation – can overwhelm the evolvability benefits of mutations in the successive sweeps regime.

The fixation probability of a modifier with a direct benefit ($s_m > 0$) can be computed using a similar procedure. When $s_m \ll 2NU_b\bar{s} \ll \bar{s}$, the next sweep will typically occur before the modifier lineage establishes, so an analogous version of Eq. (S12) still holds:

$$\tilde{p}_{\text{fix}}\left(\mu \to \mu_m; s_m\right) \approx \frac{\frac{U_b'\bar{s}'}{U_b\bar{s}}}{1-\frac{s_m}{2NU_b\bar{s}}}, \qquad (s_m \ll 2NU_b\bar{s} \ll \bar{s}) \tag{S13}$$

For stronger fitness benefits ($s_m \gg 2NU_b\bar{s} \gg 1/N$), the modifier will have the opportunity establish and grow deterministically before the next sweep occurs:

$$f_m(t) \approx \begin{cases} \frac{\frac{1}{2Ns_m}e^{s_m t}}{1+\frac{1}{2Ns_m}(e^{s_m t}-1)} & \text{w/ prob } 2s_b, \\ 0 & \text{else.} \end{cases} \tag{S14}$$

In this regime, the fixation process is similar to the "first-step" clonal interference analysis in Ref. (87). Provided that $s_m \ll \bar{s}$, the additional fitness benefit of the modifier will have a negligible impact the establishment probability of the next sweep in either genetic background, so that

$$\lambda_m(t) = 2NU_b'(\bar{s}' + s_m\left[1-f_m(t)\right] \cdot f_m(t) \approx 2NU_b'\bar{s}' \cdot f_m(t), \tag{S15a}$$

$$\lambda_w(t) = 2N\int_{s_m}^{\infty}\left[s - s_m f_m(t)\right]\mu(s)\,ds \cdot \left[1-f_m(t)\right] \approx 2NU_b\bar{s}\left[1-f_m(t)\right]. \tag{S15b}$$

The fixation probability is determined by these competing Poisson processes, so that

$$\begin{aligned} \tilde{p}_{\text{fix}}\left(\mu \to \mu_m; s_m\right) &= N\left\langle \int_0^{\infty} dt\lambda_m(t)e^{-\int_0^{\infty}dt'[\lambda_m(t')+\lambda_w(t')]}\right\rangle, \\ &\approx 2NU_b'\bar{s}'\int_0^{\infty} dt\; e^{s_m t - 2NU_b\bar{s}t - \left[2N\left(\frac{U_b'\bar{s}'-U_b\bar{s}}{s_m}\right)+1\right]\cdot\log\left[1+\frac{1}{2Ns_m}(e^{s_m t}-1)\right]}, \\ &\approx 2Ns_m e^{-\frac{2NU_b\bar{s}}{s_m}\log[2Ns_m]}, \end{aligned} \tag{S16}$$

which will be valid in the limit that $2NU_b\bar{s} \ll s_m \ll \bar{s}$. This result has a simple heuristic interpretation as the establishment probability of the modifier multiplied by the probability the wildtype population does not generate a sweep before the modifier fixes on its own (87). Since this fixation probability is independent of $\mu_m(s)$, it implies that even small direct benefits — much smaller than the size of a typical driver mutation —

can override the effects of second order selection, even if they reduce the long-term rate of adaptation to zero. Together with the direct cost results above, these calculations emphasize that natural selection can be extremely sensitive to the short-term costs or benefits of a mutation in the successive sweeps regime, in contrast to what we observe in larger populations like Fig. 1A.

3 Evolvability modifiers in the clonal interference regime

In larger populations ($NU_b \gtrsim 1$), the assumption of discrete selective sweeps will start to break down. Multiple beneficial lineages will segregate in the population at the same time, and will interfere with each other as they compete for dominance in the the population (47). In this *clonal interference regime*, the fate of a given mutation will sensitively depend on the genetic background that it arises on, and the future mutations that its descendants produce before they fix or are driven to extinction. This requires a stochastic generalization of the full multi-locus dynamics in Eqs. (S1) and (S2),

$$\begin{aligned}\frac{\partial f(\vec{g})}{\partial t} = &\underbrace{\left[X(\vec{g}) - \overline{X}(t)\right] f(\vec{g})}_{\text{selection}} + \underbrace{\sum\nolimits_{\vec{g}'} M(\vec{g}' \to \vec{g}) f(\vec{g}') - \sum\nolimits_{\vec{g}} M(\vec{g} \to \vec{g}') f(\vec{g})}_{\text{mutation}} \\ &+ \underbrace{\sqrt{\frac{f(\vec{g})}{N}}\eta(\vec{g},t) - f(\vec{g}) \sum\nolimits_{\vec{g}'} \sqrt{\frac{f(\vec{g}')}{N}}\eta(\vec{g}',t) - \sum\nolimits_{\vec{g}'} \sqrt{\frac{f_m(\vec{g}')}{N}}\eta_m(\vec{g}',t)}_{\text{genetic drift}},\end{aligned} \tag{S17a}$$

and

$$\begin{aligned}\frac{\partial f_m(\vec{g})}{\partial t} = &\underbrace{\left[X_m(\vec{g}) - \overline{X}(t)\right] f_m(\vec{g})}_{\text{selection}} + \underbrace{\sum\nolimits_{\vec{g}'} M_m(\vec{g}' \to \vec{g}) f_m(\vec{g}') - \sum\nolimits_{\vec{g}} M_m(\vec{g} \to \vec{g}') f_m(\vec{g})}_{\text{mutation}} \\ &+ \underbrace{\sqrt{\frac{f(\vec{g})}{N}}\eta(\vec{g},t) - f(\vec{g}) \sum\nolimits_{\vec{g}'} \sqrt{\frac{f(\vec{g}')}{N}}\eta(\vec{g}',t) - \sum\nolimits_{\vec{g}'} \sqrt{\frac{f_m(\vec{g}')}{N}}\eta_m(\vec{g}',t)}_{\text{genetic drift}},\end{aligned} \tag{S17b}$$

where $\eta(\vec{g},t)$ and $\eta_m(\vec{g},t)$ are uncorrelated Brownian noise terms (88). While this model is straightforward to write down, there are no known solutions for arbitrary choices of $X(\vec{g})$ and $X_m(\vec{g})$. Further progress requires us to make specific assumptions about the shapes of these fitness landscapes, as well as the corresponding mutation kernels $M(\vec{g} \to \vec{g}')$ and $M_m(\vec{g} \to \vec{g}')$. We introduce one particularly convenient parameterization in the next section, which is motivated by the concept of *macroscopic epistasis* explored in previous work (41, 89).

3.1 Macroscopic epistasis expansion

There are many possible fitness landscapes one could consider. These are often parameterized as a power series involving different combinations of loci,

$$X(\vec{g}) = \sum_{\ell=1}^{L} s_\ell g_\ell + \sum_{\ell<\ell'} \epsilon_{\ell,\ell'} g_\ell g_{\ell'} + \sum_{\ell<\ell'<\ell''} \epsilon_{\ell,\ell',\ell''} g_\ell g_{\ell'} g_{\ell''} + \dots \tag{S18}$$

where $g_\ell = 1$ if there is a mutation at site ℓ and 0 otherwise (75). This representation can be viewed as a Taylor expansion around a perfectly smooth fitness landscape, where all the $\{\epsilon_{\ell,\ldots,\ell'}\}$ coefficients vanish. Nonzero values of the $\epsilon_{\ell,\ldots,\ell'}$ coefficients correspond to epistatic interactions between loci; following previous work (41), we will refer to these interactions as *microscopic epistasis*, since they can in principle vary across all possible combinations of sites.

Modifier alleles can be expressed in this framework by designating an arbitrary site as the modifier locus (e.g. $\ell = m$), and recalculating the landscape for $g_m = 1$ (modifiers) and $g_m = 0$ (wildtype) separately. This notation makes it clear that a modifier that changes the fitness effects of other mutations must necessarily involve some microscopic epistasis, corresponding to terms like $\epsilon_{m,\ell}$, $\epsilon_{m,\ell,\ell'}$, and so on.

The space of epistatic fitness landscapes is enormous, and their impact on the evolutionary dynamics of large populations is not well understood (64). Some studies have attempted to navigate this complexity by truncating Eq. (S18) after the pairwise terms (90); others have focused on smaller landscapes containing just a handful of interacting loci (91, 92). In our analysis below, we will show that it will be useful to consider an alternative limit of Eq. (S18), in which the local DFEs defined in Eqs. (S6) and (S7) are approximately constant for different genotypes:

$$\mu(s|\vec{g}) \approx \mu(s)\,, \quad \mu_m(s|\vec{g}) \approx \mu_m(s)\,. \tag{S19}$$

We can view this approximation as the lowest order term in an alternative expansion of the fitness landscape,

$$\mu(s|\vec{g}) \approx \mu(s) + \sum_\ell \delta\mu_\ell(s)\cdot g_\ell + \sum_{\ell<\ell'} \delta\mu_{\ell,\ell'}(s)\cdot g_\ell g_{\ell'} + \ldots\,, \tag{S20}$$

which works directly in the DFE basis. This genotype dependence of the DFE is sometimes known as *macroscopic epistasis* (41), since it aggregates over a large number of microscopic interactions in Eq. (S18). The approximation in Eq. (S19) can therefore be viewed as the simplest possible model that incorporates some amount of macroscopic epistasis, with a non-zero $\delta\mu_\ell(s)$ term at the modifier locus ($\ell = m$) and all other $\delta\mu_{\ell,\ldots,\ell'}$ terms vanishing.

We note that such macroscopic epistasis can arise even in the absence of microscopic epistasis ($\epsilon_{\ell,\ldots,\ell'} = 0$) if the modifier alters the mutation rates at other loci. For example, mutator alleles are often modeled as a simple change of scale,

$$M_m(\vec{g} \to \vec{g}') = r \cdot M_m(\vec{g} \to \vec{g}')\,, \tag{S21}$$

leading to a proportional change in the DFE,

$$\mu_m(s|\vec{g}) = r \cdot \mu(s|\vec{g})\,, \tag{S22}$$

which has been the starting point for many previous studies (24, 26, 85). In practice, mutator strains are typically biased toward specific types of mutations, so that the simple proportional model in Eq. (S21) does not necessarily apply (4, 6, 27). Other mutagenesis mechanisms that target specific genomic regions (93, 94) lead to conceptually similar complications. Both require us to consider changes in the shape of the DFE in addition to its overall scale (4, 6, 27).

In addition to mutator alleles, Eq. (S19) can also emerge from epistatic interactions between loci. The simplest example is a pairwise epistasis model, with nonzero values for the coefficients involving $\ell = m$, and zeros everywhere else. This is sufficient to recover the limit in Eq. (S19), but it is not the only possibility: any other landscape that gives rise to the same overall distributions in Eq. (S19) will exhibit similar dynamics, even if the individual fitness effects $[s_\ell(\vec{g}) \equiv X(\vec{g}+\ell) - \vec{X}(\vec{g})]$ are undergoing more complicated rearrangements under the hood.

Moreover, our results will not require Eq. (S19) to hold across the entire fitness landscape , but only within a smaller region that is explored before the modifier either fixes or goes extinct. We determine the size of this local neighborhood in SI Section 7 and find that it is often modest, corresponding to just a handful of mutational steps for many empirically relevant parameter values (Fig. S3). We will also show that the assumption in Eq. (S19) is most sensitive to a narrow range of beneficial fitness effects, so that substantial deviations in other parts of the DFE can still have a negligible impact on the modifier lineage (SI Section 7). At present, it is difficult to enumerate all of the microscopic landscapes that are consistent with a given DFE function, $\mu(s|\vec{g})$. However, we can still identify several examples that satisfy Eq. (S19) – at least in the approximate sense required – that go beyond the mutator allele and pairwise epistasis examples above:

Branching epistatic landscapes. At a formal level, we can consider a "maximally epistatic" landscape of branching uphill paths of length K, where each step k of a given path can access $M \ll L$ other beneficial mutations (Fig. S1A). If the uphill paths do not contain any loops, the fitness function can be expressed in the form

$$X(\vec{g}) = \sum_{k=1}^{K} \sum_{(\ell_1,\ldots,\ell_k)\in\mathcal{L}_k} \left(\sum_{j=1}^{k} s_{\ell_j}\right)\left(\prod_{j=1}^{k} g_{\ell_j}\right) \prod_{\ell\notin\{\ell_j\}_{j=1}^{k}} (1-g_\ell)\,, \tag{S23}$$

where $\mathcal{L}_k$ denotes the set of all uphill sub-paths of length k, and the s_ℓ coefficients are independently drawn from the target DFE $\mu(s)$. This construction ensures that the DFEs calculated from Eq. (S23) will satisfy $\mu(s|\vec{g}) \approx \mu(s)$ when M is large, even though the landscape contains large amounts of microscopic epistasis. We can extend this construction to include a modifier allele by writing

$$X(\vec{g}) = \sum_{k=1}^{K} \sum_{(\ell_1,\ldots,\ell_k)\in\mathcal{L}_k} \left(\sum_{j=1}^{k} s_{\ell_j}(1-g_m) + s'_{\ell_j} g_m\right)\left(\prod_{j=1}^{k} g_{\ell_j}\right) \prod_{\ell\notin\{\ell_j\}_{j=1}^{k}} (1-g_\ell)\,, \tag{S24}$$

where the s'_ℓ coefficients are independently drawn from the modifier DFE $\mu_m(s)$. This microscopic landscape satisfies the macroscopic epistasis approximation in Eq. (S19), but includes many non-pairwise interactions by construction.

Non-linear global phenotypes (e.g. stability-activity tradeoffs). The approximation in Eq. (S19) can also emerge in more concrete settings, when the fitness of the organism is a non-linear function of multiple global phenotypes. A prototypical example is the stability-activity tradeoff that is often observed in viral evolution (95, 96) and other protein-binding problems (97), where stabilizing mutations can potentiate the fitness benefits of mutations that would destabilize the protein on their own. The essential ingredients of this behavior can be captured in a simple model containing two global phenotypes: (i) activity, denoted by $\Psi(\vec{g})$, which contributes additively to the total fitness, and (ii) stability, denoted by $\Phi(\vec{g})$, which has a Boltzmann-like contribution,

$$X(\vec{g}) = -\log\left(1+e^{-\Phi(\vec{g})}\right) + \Psi(\vec{g}) + \text{const}\,, \tag{S25a}$$

We will assume that both phenotypes can be expressed as additive functions of the genotype,

$$\Phi(\vec{g}) = \phi_0 + \sum_{\ell}^{L} \phi_\ell \cdot g_\ell\,, \quad \Psi(\vec{g}) = \sum_{\ell}^{L} \psi_\ell \cdot g_\ell\,, \tag{S25b}$$

where ϕ_0 denotes the stability of the reference strain. In this model, mutations that only increase the activity of the protein are always beneficial ($s_\ell \approx \psi_\ell$). However, mutations that increase activity while also decreasing stability can be either costly or beneficial depending on the stability of the background that they occur on.

Suppose that there are a large number (M) of such mutations with $|\phi_\ell| - \phi_0 \gg 1$, which implies that they will be deleterious on the wildtype background. In this scenario, any mutation that strongly increases stability will function like an evolvability modifier, by allowing these previously deleterious mutations to occur. In particular, if the stability enhancement ϕ_m is much larger than $|\phi_\ell| - \phi_0$, then the fitness benefits of the unleashed mutations will be approximately constant ($s_\ell \approx \psi_\ell$) until roughly $K \approx \phi_m/(|\phi_\ell| - \phi_0)$ such variants have accumulated. If $M \gg K \gg 1$, this example will satisfy the approximation in Eq. (S19), while relying on the higher-order epistatic interactions in Eq. (S25).

While this particular example was motivated by protein stability, similar behavior can occur for other combinations of phenotypes, as long as they include the appropriate non-linearities. For example, a simple model of stabilizing selection involving a one nearly optimized phenotype and another non-optimized one can be expressed as

$$X(\vec{g}) = -|\Phi(\vec{g})| + \Psi(\vec{g}) \,, \tag{S26a}$$

where

$$\Phi(\vec{g}) = \sum_{\ell}^{L} \phi_\ell \cdot g_\ell \,, \quad \Psi(\vec{g}) = \sum_{\ell}^{L} \psi_\ell \cdot g_\ell \,. \tag{S26b}$$

A mutation that increases $\Psi(\vec{g})$ while displacing $\Phi(\vec{g})$ from its optimal value would enable $K \approx \phi_m/\phi_\ell$ previously deleterious mutations to accumulate — each providing a fitness benefit $s_\ell \approx \phi_\ell$ — before the phenotypic optimum is reattained. If K is sufficiently large, this evolvability modifier would also satisfy the approximation in Eq. (S19), while involving a distinct form of non-pairwise epistasis.

Chromosomal duplications / aneuploidy. The approximation in Eq. (S19) can also apply to scenarios that are difficult to capture with a traditional fitness landscape, because they involve changes to the structure of the genome itself. Classical examples include chromosomal duplications and other copy-number changes, which are frequently observed in cancer evolution (14, 98) and laboratory evolution experiments in eukaryotes (99). These copy number variants involve changes in ploidy — in addition to changes in target size — so that dominance effects start to become important. For example, a whole-genome duplication of a haploid genome would lead to a modified DFE of the form

$$\mu_m(s) = \sum_{\ell=1}^{L} 2\mu_\ell \cdot \delta(s - s_\ell h_\ell) \,, \tag{S27}$$

where h_ℓ denotes the dominance coefficient of the mutation at site ℓ. The successive sweeps picture in SI Section 2 predicts that the fixation probability of this copy-number variant is $\propto 2\overline{sh}/s$, which is neutral in purely semi-dominant case ($\overline{sh} = s/2 \implies \tilde{p}_{\text{fix}} = 1$) and only moderately beneficial for complete dominance ($\overline{sh} = s/2 \implies \tilde{p}_{\text{fix}} \approx 2$). In contrast, our results below suggest that in larger populations, these variants can be both strongly favored or disfavored by second-order selection depending on the joint distribution of (s_ℓ, h_ℓ).

Together, these examples illustrate that the limiting behavior in Eq. (S19) — while still capturing just a subset of all possible fitness landscapes — can nevertheless approximate many biologically relevant scenarios where

second-order selection is thought to play an important role. We will therefore use this model as a starting point for all of our mathematical derivations below. We will revisit this assumption in SI Section 7, where we discuss extensions to other possible fitness landscapes.

3.2 Fitness wave formalism

An advantage of the model in Eq. (S19) is that it allows us to exploit existing "fitness wave" methods for modeling clonal interference (24, 39, 40, 46, 47, 69, 100–103). This previous literature has shown that the multi-locus dynamics in Eq. (S17) can often simplified by considering a coarse-grained picture, which groups together individuals with the same overall fitness:

$$f(X,t) = \sum_{\vec{g}} f(\vec{g},t) \cdot \delta(X - X(\vec{g}))\,. \tag{S28}$$

We can extend this idea to our present context, defining a corresponding fitness distribution for the modifier lineage as well:

$$f_m(X,t) = \sum_{\vec{g}} f_m(\vec{g},t)\delta(X - X_m(\vec{g}))\,. \tag{S29}$$

In the special case that $\mu(s|\vec{g}) = \mu(s)$ and $\mu_m(s|\vec{g}) = \mu_m(s)$ for all genotypes $\vec{g}$, Eq. (S17) can be written in the coarse-grained form,

$$\begin{aligned}\frac{\partial f(X)}{\partial t} = &\underbrace{\left[X - \overline{X}(t)\right] f(X)}_{\text{selection}} + \int ds\ \underbrace{\mu(s)\left[f(X-s) - f(X)\right]}_{\text{mutation}} \\ &+ \underbrace{\int dX'\left[\delta(X-X') - f(X)\right]\sqrt{\frac{f(X')}{N}}\eta(X') - f(X)\int dX'\sqrt{\frac{f_m(X')}{N}}\eta_m(X')}_{\text{genetic drift}}\,,\end{aligned} \tag{S30a}$$

$$\begin{aligned}\frac{\partial f_m(X)}{\partial t} = &\underbrace{\left[X - \overline{X}(t)\right] f_m(X)}_{\text{selection}} + \int ds\ \underbrace{\mu_m(s)\left[f_m(X-s) - f_m(X)\right]}_{\text{mutation}} \\ &+ \underbrace{\int dX'\left[\delta(X-X') - f_m(X)\right]\sqrt{\frac{f_m(X')}{N}}\eta_m(X') - f_m(X)\int dX'\sqrt{\frac{f(X')}{N}}\eta(X')}_{\text{genetic drift}}\,,\end{aligned} \tag{S30b}$$

where $\overline{X}(t) = \int X f(X,t)dX + \int X f_m(X,t)dX$ is the mean fitness of the population, and the $\eta_i(X)$ are uncorrelated Brownian noise terms (88). This allows us to generalize the notion of a mutator allele to handle more general differences in evolvability, while still bypassing the enormous complexity of the underlying fitness landscape.

In the absence of the modifier ($f_m = 0$), the wildtype distribution $f(X,t)$ approaches a traveling wave form that increases in fitness at an average rate $\langle \partial_t \overline{X}(t)\rangle \equiv v(\mu, N)$ (39, 47). Previous work has shown that the typical profile $f(x)$ is well-approximated by the deterministic equation,

$$-v\partial_x f(x) = x \cdot f(x) + \int \mu(s)\left[f(x-s) - f(x)\right] ds\,, \tag{S31}$$

which is the expected value of Eq. (S30a) when $\overline{X}(t) \approx vt$ (39, 40).

We assume that the modifier mutation will arise in this steady-state population, on a genetic background drawn from $f(x)$. Its descendants will found a second fitness wave, $f_m(X,t)$, which competes with the wildtype population $f(X,t)$ as they continue to acquire additional mutations according to the dynamics in Eq. (S17) (Fig. 1).

Branching process approximation. The key approximation we will make in this work is that the fate of the modifier will often be determined while it is still at a low frequency in the population. Previous studies have shown that this is a good approximation for first-order mutations – including both neutral and deleterious mutations – in the clonal interference regime (39, 40, 43, 46, 103, 104). In this work, we use a combination of self-consistency arguments and comparisons with simulations to show that this same approximation holds for a broad range of modifier alleles as well. The main exceptions will occur for modifiers that dramatically reduce evolvability (e.g. the dead-end modifiers in Fig. 3D); we treat this case separately in SI Section 4.1.6.

When the frequency of the modifier is small, the mean fitness of the population can be approximated by the wildtype value $\overline{X}(t) \approx vt$, and the higher-order contributions in the drift term in Eq. (S30b) can be neglected. The dynamics of the modifier then reduce to the linear branching process,

$$\frac{f_m(X)}{\partial t} = [X - vt]\, f_m(X) + \int ds\, \mu_m(s)[f_m(X-s) - f_m(X)] + \sqrt{\frac{f_m(X)}{N}}\eta_m(X)\,, \tag{S32}$$

subject to the initial condition $f_m(X,t) = \delta(X - x - s_m)$. The non-extinction probability of this process satisfies the standard branching process recursion,

$$\begin{aligned} 0 = &\underbrace{x \cdot w(x|\mu \to \mu_m)}_{\text{selection}} + \int \underbrace{\mu_m(s)\,[w(x+s|\mu \to \mu_m) - w(x|\mu \to \mu_m)]}_{\text{mutation}}\, ds \\ &- \underbrace{v(\mu, N) \cdot \partial_x w(x|\mu \to \mu_m)}_{\text{competition with wildtype}} - \underbrace{\frac{1}{2} \cdot w(x|\mu \to \mu_m)^2}_{\text{genetic drift while rare}}\,, \end{aligned} \tag{S33}$$

which is a straightforward generalization of the mutator version in Ref. (24). If the fate of the modifier is determined while it is rare, then this non-extinction probability must also coincide with the long-term probability of fixation. This implies that the overall fixation probability of the modifier can be obtained from Eq. (S33) by averaging over the random genetic background,

$$p_{\text{fix}}(\mu \to \mu_m; s_m) = \int f(x) w(x + s_m|\mu \to \mu_m) dx\,. \tag{S34}$$

In the limit that $\mu_m(s) \to \mu(s)$, Eq. (S34) reduces to the fixation probability of a first-order mutation which has been studied in previous work (39, 40, 46, 103). In particular, this previous work has shown that a self-consistency condition can be obtained from the fact that the overall fixation probability of a neutral mutation must always be equal to $1/N$ (this follows directly from the normalizability of the full dynamics in Eq. S30). Combining this fact with Eq. (S34) yields the self-consistency condition,

$$p_{\text{fix}}(\mu \to \mu; 0) = \int dx f(x) w(x|\mu \to \mu) = \frac{1}{N}\,. \tag{S35}$$

Together with Eqs. (S31) and (S33), this completely determines the wildtype rate of adaptation $v(\mu, N)$ as a function of $\mu(s)$ and N (39, 40, 46, 103).

When $\mu_m(s) \neq \mu(s)$, the conditional fixation probability in Eq. (S33) will differ from that of a first-order mutation because there are different mutation spectra in the mean fitness and mutation terms. This is because the modifier lineage primarily competes against the wildtype population [whose mean fitness is controlled by $\mu(s)$], but acquires additional mutations from its own distribution, $\mu_m(s)$. We note, however, that since the competition with the wildtype population is completely mediated by $v(\mu, N)$ (a monotonic function of N), the conditional fixation probability of the modifier lineage can always be mapped to the conditional fixation probability of a first-order mutation in a population that is fixed for the modifier allele, but with a different population size N^*. In other words,

$$w(x|\mu \to \mu_m, N) = w(x|\mu_m \to \mu_m, N^*)\,, \tag{S36}$$

where N^* is defined by

$$v(\mu_m, N^*) = v(\mu, N)\,. \tag{S37}$$

This implies that the space of solutions for $w(x|\mu \to \mu_m)$ will have the same general form as the "first-order" $w(x)$ function that has been studied in previous work (39, 40, 46, 102, 103). It also implies that the dynamics of non-extinction in Eq. (S32), which were previously described in Ref. (40), will be qualitatively similar as well. However, we will see below that actually using this result for evolvability modifiers will often require solutions to Eq. (S33) that go beyond the parameter regimes that have been examined in these earlier studies. We have therefore developed a new approach for deriving approximate analytical solutions for $w(x|\mu \to \mu_m)$ that will apply across this broader range of parameters. We outline the general approach in SI Section 3.3 below, and apply it to different classes of distributions of fitness effects in SI Sections 4 and 5.

For notational convenience, we will suppress the explicit dependence on $\mu(s)$ and $\mu_m(s)$ in the following sections, writing $w(x) \equiv w(x|\mu \to \mu)$ for the conditional fixation probability of a first-order mutation and $w_m \equiv w(x|\mu \to \mu_m)$ for the conditional fixation probability of the modifier lineage. We will also let $v \equiv v(\mu)$ denote the rate of adaptation in the wildtype population and $v_m \equiv v(\mu_m)$ denote the long-term rate of adaptation that is achieved if the modifier takes over.

3.3 Asymptotic solution in the "sharp shoulder" regime

There are many parameter regimes of clonal interference that one can consider (40, 103). Here, we will primarily focus on a regime that is relevant for a broad range of naturally and experimentally evolving populations, which roughly corresponds to the case where the fitness benefits of a typical "driver mutation" (s_b) are much larger than the total rate at which they occur (U_b) (40, 46). We will define these conditions more precisely below.

Relative fitness distribution. Previous work has shown that in our regime of interest, the solution to Eq. (S31) can be approximated by a truncated Gaussian distribution,

$$f(x) \approx \begin{cases} \frac{1}{\sqrt{2\pi v}} e^{-\frac{x^2}{2v}} & x \leq x_c\,, \\ 0 & x > x_c\,, \end{cases} \tag{S38}$$

where x_c denotes the location of the fittest individuals that are likely to exist within the population (39, 40, 46). This solution is valid when $x_c - s_b \gg \sqrt{v}$ and $s_b \gg \sqrt{v}$, which constitutes the formal definition of the regime that we consider in this work. The first condition ($x_c - s_b \gg \sqrt{v}$) implies that the parents of the most fit individuals are substantially more fit than the majority of the population (i.e. clonal interference is common).

The second condition ($s_b \gg \sqrt{v}$) implies that the majority of the population is concentrated near the mean fitness. This is also known as the "moderate speeds" regime (40), since the second condition can be rewritten as $v \ll s_b^2$. We will assume that these conditions hold for the wildtype population in which the modifier arises.

Lineage fixation probability. Previous work (39, 46) has shown that in our parameter regime of interest, the solutions to Eq. (S33) can be decomposed into a high-fitness region, where the mutation term is sub-dominant:

$$0 \approx \underbrace{x \cdot w_m(x)}_{\text{selection}} - \underbrace{v \cdot \partial_x w_m(x)}_{\text{competition w/ wt}} - \underbrace{\frac{1}{2} \cdot w_m(x)^2}_{\text{genetic drift}}, \tag{S39}$$

and a linearized region, where the mutation term is important but the drift term can be neglected:

$$0 \approx \underbrace{x \cdot w_m(x)}_{\text{selection}} + \int ds \; \underbrace{\mu_m(s)[w_m(x+s) - w_m(x)]}_{\text{mutation}} - \underbrace{v \cdot \partial_x w_m(x)}_{\text{competition w/ wt}} . \tag{S40}$$

There is also a narrow region in the middle where both approximations are valid, so that asymptotic matching can be used to obtain a full global solution. This asymptotic decomposition will continue to be valid for our modifier case as well.

In the high fitness region, the solution to Eq. (S39) is given by the *shoulder solution*,

$$w_m(x) \approx \frac{2x_{cm} e^{(x^2 - x_{cm}^2)/2v}}{1 + (x_{cm}/x)\, e^{(x^2 - x_{cm}^2)/2v}}, \tag{S41}$$

where x_{cm} is a constant of integration that will be determined self-consistently below. When $x_{cm} \gg \sqrt{v}$, this shoulder solution develops a narrow boundary layer around $x = x_{cm} \pm O(v/x_{cm})$, such that the fixation probability can be well-approximated by the piecewise form,

$$w_m(x) \approx \begin{cases} 2x & \text{if } x - x_{cm} \gtrsim O(v/x_{cm}), \\ 2x_{cm} e^{\frac{x^2 - x_{cm}^2}{2v}} & \text{if } x_{cm} - x \lesssim O(v/x_{cm}). \end{cases} \tag{S42}$$

This piecewise function has a simple interpretation in terms of the dominant balances in Eq. (S39) (39). The linear scaling above x_{cm} emerges from a balance between the selection and genetic drift terms; this implies that lineages with $x \gtrsim x_{cm}$ will fix provided that they survive genetic drift. Conversely, the exponential scaling at lower x emerges from a balance between the selection and mean fitness terms; this implies that lineages with $x \lesssim x_{cm}$ are strongly influenced by competition with the surrounding population (i.e. clonal interference). In the special case where $\mu_m(s) = \mu(s)$, previous work has shown that x_{cm} approximately coincides with the nose of the fitness distribution in Eq. (S38) (39, 40, 46). This makes some intuitive sense: lineages with relative fitness $\gtrsim x_c$ do not experience clonal interference, and will fix if they survive genetic drift; this suggests that the steady-state fitness distribution should not contain such individuals, since they should already have fixed. We will therefore also refer to x_c as the *interference threshold* for the wildtype population. When $\mu_m(s) \neq \mu(s)$ the interference threshold of the modifier (x_{cm}) will generally differ from the location of the nose (x_c) — this will be critically important for our analysis in SI Sections 4 and 5 below.

The mutation terms must eventually become important for smaller values of x, since the shoulder solution in Eq. (S42) starts to increase for $x < 0$. To avoid this unphysical behavior, and ensure $w_m(x)$ decreases as

$x \to \infty$, the shoulder solution must eventually map on to the correct branch of the solution to Eq. (S33). Previous work has used saddle point methods to identify the relevant solutions to Eq. (S40) (40, 103), while other studies have utilized thresholding approximations that set $w_m(x) \approx 0$ below a critical fitness value (39, 46). Here we take a slightly different approach, by recasting Eq. (S33) as an integral equation.

Multiplying both sides of Eq. (S33) by $e^{\frac{-(x-U_0)^2}{2v}}$ and integrating from $-\infty$ to x, we can rewrite Eq. (S33) in the recursive form,

$$\begin{aligned} w_m(x) = e^{\frac{(x-U_0)^2}{2v}} \int_0^\infty ds \int_{-\infty}^{x+s} \frac{\mu_m(s)}{v} \cdot e^{\frac{-(y-s-U_0)^2}{2v}} w_m(y)\, dy \\ - e^{\frac{(x-U_0)^2}{2v}} \int_{-\infty}^{x} e^{\frac{-(y-U_0)^2}{2v}} \cdot \frac{w_m(y)^2}{2v}\, dy \,, \end{aligned} \tag{S43}$$

where $U_0 \equiv \int \mu_m(s)\, ds$ is the total mutation rate. Since $w_m(x)$ rapidly declines below the interference threshold, the contributions from the second term will become negligible when $x \lesssim x_{cm} - O(v/x_{cm})$, and Eq. (S43) will reduce to the simpler form

$$w_m(x) = e^{\frac{(x-U_0)^2}{2v}} \int_0^\infty ds \int_{-\infty}^{x+s} \frac{\mu_m(s)}{v} \cdot e^{\frac{-(y-s-U_0)^2}{2v}} \cdot w_m(y)\, dy \,. \tag{S44}$$

This recursive formula has a simple intuitive interpretation illustrated by the schematic in Fig. 1. A modifier lineage founded at a relative fitness $x < x_{cm}$ will start as a single clone, whose average size will evolve as

$$\langle n(t) \rangle = \exp\left[xt - \frac{vt^2}{2} - U_0 t \right] . \tag{S45}$$

These growth dynamics account for the steady increase in the wildtype mean fitness, as well as the outflow of individuals that acquire further mutations as the clone is growing. The total production rate of new mutations is $\langle n(t) \rangle \cdot \mu(s) ds$, and each of these events will found a new modifier lineage with relative fitness $x + s - vt$; the original lineage will fix if one of these descendant lineages is ultimately successful, yielding the recursive formula,

$$w_m(x) = \int_0^\infty dt \int_0^\infty ds\, \mu_m(s) \cdot e^{xt - \frac{vt^2}{2} - U_0 t} \cdot w_m(x + s - vt) \,. \tag{S46}$$

The linearity of this expression implies that successful clones are highly unlikely to give rise to multiple successful mutants, which is consistent with the assumption that clonal interference is very strong ($w_m(x) \ll 2x$) when $x < x_{cm}$. Equation (S44) can be recovered from Eq. (S46) by changing variables from t to $y \equiv x - v \cdot t + s$, which represents the time-dependent landing fitness of the clone's mutant offspring.

Note that while the recursion in Eq. (S44) is also a solution to the linearized version of Eq. (S33), our present derivation shows that the interpretation is slightly different here. In particular, since the upper limit of the y-integral in Eq. (S44) is larger than x, the mutation term can in principle depend on the behavior of $w_m(x)$ for $x > x_{cm}$, where the effects of the nonlinear $w_m(x)^2$ term start to become important. We can account for this non-locality using the shoulder solution in Eq. (S42), by rewriting Eq. (S44) as a piecewise integral,

$$w_m(x) = e^{\frac{(x-U_0)^2}{2v}} \int_0^\infty ds \int_{-\infty}^{x+s} \frac{\mu_m(s)}{v} \cdot e^{\frac{-(y-s-U_0)^2}{2v}} \cdot \left[\frac{2x_{cm} e^{\frac{y^2 - x_{cm}^2}{2v}} \theta(y - x_{\min})}{1 + \frac{x_{cm}}{y} e^{\frac{y^2 - x_{cm}^2}{2v}}} + w_m(y) \theta(x_{\min} - y) \right] dy \,, \tag{S47}$$

where $\theta(\cdot)$ is the unit step function, and $x_{\min}$ represents the point a which the shoulder solution starts to break down. When $x_{\min} \lesssim x_{cm} - O(v/x_{cm})$, we can use this expression to determine x_{cm}, by noting that Eq. (S47) must also match the shoulder solution in Eq. (S42) in the overlap region where both approximations our valid $[x_{\min} \lesssim x \lesssim x_c - O(x_c)]$.

While this integral formulation is similar to the transform methods (40, 103) that have previously been used to analyze $w_m(x)$, it offers practical advantages that will become important for our analysis of modifier mutations below. In particular, we will see that in our parameter regime of interest, we can use Eq. (S47) to analytically extend the shoulder solution to progressively lower fitness values. These analytical approximations will be critically important for treating modifier mutations with large direct costs (Fig. 3).

In the following sections, we use this framework to derive explicit solutions for $w_m(x)$ different choices of $\mu_m(s)$. We begin by considering a simple model, where the driver mutations all share the same characteristic fitness benefit (SI Section 4). This will allow us to verify the self-consistency conditions assumed above, and to obtain explicit predictions for the overall fixation probabilities of different modifier mutations. We then extend these calculations to continuous distributions of fitness effects in SI Section 5, and show that more general distributions can often be understood using the simple model in SI Section 4.

4 Solution for a simple model of the distribution of fitness effects

To make analytical progress, we first consider a simple scenario in which the wildtype population produces mutations with a single fitness benefit s_b at a total rate U_b. The distribution of fitness effects can then be written as a point mass,

$$\mu(s) = U_b \cdot \delta(s - s_b)\,, \tag{S48}$$

where $\delta(z)$ is the Dirac delta function. Previous work has shown that in our parameter regime of interest, the rate of adaptation (v) and interference threshold (x_c) follow the approximate scaling,

$$v \approx \frac{2s_b^2 \log(Ns_b)}{\log^2(s_b/U_b)}\,, \qquad x_c \approx \frac{2s_b \log(Ns_b)}{\log(s_b/U_b)}\,, \tag{S49}$$

which are valid in the limit that $x_c \gg s_b$ (39, 40, 47). In terms of these parameters, the conditions we assumed for our solution in SI Section 3.3 ($x_c \gg \sqrt{v}$ and $s_b \gg \sqrt{v}$) become

$$1 \ll \log(Ns_b) \ll \log^2(s_b/U_b)\,, \tag{S50}$$

which require that $Ns_b \gg 1$ and $s_b \gg U_b$.

For distributions of fitness effects in the same class as Eq. (S48), the most general evolvability modifier is one that produces mutations with a different benefit s_b' and different rate U_b', so that

$$\mu_m(s) = U_b' \cdot \delta(s - s_b')\,. \tag{S51}$$

In the analysis below, we will assume that U_b' and s_b' are chosen such that $U_b' \ll s_b'$, $x_{cm} \gg \sqrt{v}$, and $s_b' \gg \sqrt{v}$. In this regime, the outflow of mutations can be neglected in Eq. (S47). We will also assume that the fold change in the mutation rate may be large $[\log(U_b'/U_b) \gg 1]$, but the fold change in the selection strength will always be comparatively modest $[\log(s_b'/s_b) \lesssim \log(10)]$. This will be sufficient to derive all of the results in the main text.

Substituting Eq. (S51) into Eq. (S47) yields an integral relation for $w_m(x)$,

$$w_m(x) = 2x_{cm} e^{\frac{x^2}{2v} - \frac{x_{cm}^2}{2v}} \times \frac{U_b'}{v} \int_{-\infty}^{x+s_b'} \left(\frac{e^{\frac{ys_b'}{v} - \frac{s_b'^2}{2v}} \theta(y - x_{\min})}{1 + \frac{x_{cm}}{y} e^{\frac{y^2}{2v} - \frac{x_{cm}^2}{2v}}} + \frac{w_m(y)\theta(x_{\min} - y)}{2x_{cm} e^{\frac{(y-s_b')^2}{2v} - \frac{x_{cm}^2}{2v}}} \right) dy \,, \tag{S52}$$

where we have neglected the outflow due to new mutations since $U_b' \ll s_b' \lesssim x$. The interference threshold is determined by matching Eq. (S52) to the shoulder solution when $x \lesssim x_{cm} - O(v/x_{cm})$, which yields a related integral,

$$1 \approx \frac{U_b}{v} \int_{-\infty}^{x_{cm}+s_b'} \left(\frac{e^{\frac{ys_b'}{v} - \frac{s_b'^2}{2v}} \theta(y - x_{\min})}{1 + \frac{x_{cm}}{y} e^{\frac{y^2}{2v} - \frac{x_{cm}^2}{2v}}} + \frac{w_m(y)\theta(x_{\min} - y)}{2x_{cm} e^{\frac{(y-s_b')^2}{2v} - \frac{x_{cm}^2}{2v}}} \right) dy \,. \tag{S53}$$

The integrals in Eqs. (S52) and (S53) will sensitively depend on the relative values of x_{cm} and s_b'. When $x_{cm} - s_b' \gg \sqrt{v}$, the integrand in Eq. (S53) will be maximized for relative fitnesses near x_{cm}. We will refer to this limit as the *multiple mutations regime*, since it implies that successful lineages must always possess anomalously high relative fitnesses ($x_{cm} - s_b' \gg \sqrt{v}$). In the opposite extreme ($s_b' - x_{cm} \gg \sqrt{v}$), the integrand in Eq. (S53) will be maximized for relative fitnesses near s_b', where the shoulder solution has already transitioned to the Haldane limit [$w(x) \approx 2x$]. We will refer to this limit as the *quasi-sweep regime*, since it implies that lineages will be guaranteed to fix if they produce a single mutation that survives genetic drift. Our assumptions imply that the wildtype population will always fall in the multiple mutations regime, but the modifier lineage may differ depending on the relative values of (s_b', U_b') and (s_b, U_b). We consider each case separately below.

4.1 Multiple mutations regime

4.1.1 Location of the interference threshold

When $x_{cm} - s_b' \gg \sqrt{v}$, the integral in Eq. (S52) will be dominated by fitnesses close to x_{cm}. In particular, the contributions from $x < x_{cm}$ will be dominated by fitnesses within $O(v/s_b')$ of x_{cm}, while the contributions from $x > x_{cm}$ will be dominated by fitness within $O(v/(x_{cm} - s_b'))$ of x_{cm}. The dominant contribution will therefore depend on the relative magnitudes of s_b' and $x_{cm} - s_b'$. When $x_{cm} - s_b' \gg s_b' \gg \sqrt{v}$, the dominant contribution will come from the exponential region of Eq. (S42), so that the auxilliary condition becomes

$$1 \approx \left(\frac{U_b'}{s_b'} \right) \exp\left[\frac{x_{cm} s_b'}{v} - \frac{s_b'^2}{v} \right] . \tag{S54}$$

Note that this derivation implicitly assumes that the shoulder solution extends at least $O(v/s_b')$ below x_{cm}; we will validate this assumption below. In the opposite case where $s_b' \gg x_{cm} - s_b' \gg \sqrt{v}$, the dominant contribution to the integral in Eq. (S53) will come from the linear region of Eq. (S42), so that the auxilliary condition becomes

$$1 \approx \left(\frac{U_b'}{x_{cm} - s_b'} \right) \exp\left[\frac{x_{cm} s_b'}{v} - \frac{s_b'^2}{2v} \right] . \tag{S55}$$

For convenience, we will summarize these two equations with the common expression,

$$1 \approx \left(\frac{U_b'}{s_b'} \right) \exp\left[\frac{x_{cm} s_b'}{v} - \frac{s_b'^2}{2v} \right] \left(\frac{x_{cm}}{x_{cm} - s_b'} \right) , \tag{S56}$$

which reduces to the correct scaling in the corresponding limits. Since we have assumed that the wildtype population always lies in the multiple mutations regime ($x_c - s_b \gg \sqrt{v}$), the background interference threshold must satisfy an analogous condition,

$$1 \approx \left(\frac{U_b}{s_b}\right) \exp\left[\frac{x_c s_b}{v} - \frac{s_b^2}{2v}\right] \left(\frac{x_c}{x_c - s_b}\right) , \tag{S57}$$

which matches the expression derived in Ref. (46). These expressions allow us to solve for x_{cm} and x_c as a function of (s'_b, U'_b), (s_b, U_b), and v. In particular, by dividing Eq. (S56) by Eq. (S57), we can solve for x_{cm} as a function of the fold changes in U_b and s_b:

$$x_{cm} \approx \frac{s_b}{s'_b}\left[x_c + \frac{s_b}{2}\left(\frac{s'^2_b}{s_b^2} - 1\right) - \frac{v}{s_b}\log\left(\frac{U'_b}{U_b}\right)\right] , \tag{S58}$$

where we have neglected logarithmic corrections in x_{cm}/x_c and s'_b/s_b. This shows that the $x_{cm} - s'_b \gg \sqrt{v}$ assumption will be valid provided that

$$\frac{s'_b}{s_b} \lesssim \sqrt{\frac{2x_c - s_b}{s_b} - \frac{2v}{s_b^2}\log\left(\frac{U'_b}{U_b}\right)} . \tag{S59}$$

In particular, Eq. (S59) shows that the multiple mutations regime will apply for pure mutation rate modifiers as large as $\log(U'_b/U_b) \lesssim (x_c - s_b)s_b/v$, while pure selection strength modifiers require the more stringent condition, $s'_b/s_b \lesssim \sqrt{(2x_c - s_b)/s_b}$. Violations of these conditions are considered in SI Section 4.2 below.

4.1.2 Extending the shoulder solution to lower fitness values

With the location of x_{cm} fixed by Eq. (S56), we will now use the integral recursion in Eq. (S52) to extend the shoulder solution for $w_m(x)$ to progressively lower fitness values. When $x + s'_b \gtrsim x_{cm} + O(v/(x_{cm} - s'_b))$, the dominant contribution to Eq. (S53) will also be contained within the region of integration in Eq. (S52). We can therefore substitute Eq. (S53) to obtain

$$w_m(x) \approx 2x_{cm} e^{\frac{x^2 - x_{cm}^2}{2v}} . \tag{S60}$$

This derivation makes it clear that the shoulder solution will continue to be valid for fitnesses as low as

$$x_{\min}(\mu_m) \equiv x_{cm} - s'_b , \tag{S61}$$

which will serve as our definition of $x_{\min}$ in Eqs. (S53) and (S52) above. This validates our assumption that the mutation term in Eq. (S33) is negligible for $x \gtrsim x_{\min} + O(v/(x_{cm} - s'_b))$. Moreover, since $x_{cm} - x_{\min} = s'_b \gg v/s'_b$, it also validates our assumption that the location of x_{cm} in Eq. (S56) is completely determined by regions where the shoulder solution is valid ($x > x_{\min}$). For fitnesses below $x_{\min}$, the finite upper limit in Eq. (S52) will start to become important. However, as long as $x + s'_b \gtrsim x_{\min} + O(v/s'_b)$, the integral will continue to be dominated by regions where the shoulder solution is valid. When $x \lesssim x_{\min} - O(v/x_{cm})$, the upper limit of integration will fall within the exponential region of the shoulder solution, yielding

$$w_m(x) \approx 2x_{cm} e^{\frac{x^2 - x_{cm}^2}{2v}} \cdot \frac{U'_b}{v} \int_{x_{\min}}^{x+s'_b} e^{\frac{y s'_b}{v} - \frac{s'^2_b}{2v}}\, dy \approx 2x_{cm} e^{\frac{x^2 - x_{cm}^2}{2v}} \cdot \frac{U'_b}{s'_b} e^{\frac{x s'_b}{v} + \frac{s'^2_b}{2v}} . \tag{S62}$$

After dividing this expression by the auxiliary condition for x_{cm} in Eq. (S56), we obtain

$$w_m(x) \approx 2x_{cm} e^{\frac{x^2 - x_{cm}^2}{2v}} \cdot \left(\frac{x_{cm} - s_b'}{x_{cm}}\right) e^{-\frac{(x_{cm} - s_b' - x)s_b'}{v}} , \tag{S63}$$

which will be valid for relative fitnesses in the range

$$x_{cm} - s_b' - O(v/x_{cm}) \gtrsim x \gtrsim x_{cm} - 2s_b' + O(v/s_b') . \tag{S64}$$

By comparing Eq. (S60) and Eq. (S63), we can see that $w_m(x)$ declines by a factor of $\sim e^{s_b'^2/v}$ across this region. Since we have assumed that $s_b' \gg \sqrt{v}$, this provides a natural justification for the thresholding approximation employed in Ref. (46), which assumed that $w(x)$ was negligible for fitnesses below $x_{\min}$. Equation (S63) constitutes a more quantitative version of this approximation, which will be useful for the analysis below.

We can continue these calculations to recursively extend $w_m(x)$ to progressively lower fitness values. When $x \lesssim x_{cm} - 2s_b' + O(v/x_{cm})$, the shoulder solution will not contribute at all to Eq. (S52), but the region in Eq. (S63) will dominate, yielding

$$\begin{aligned} w_m(x) &\approx 2x_{cm} e^{\frac{x^2 - x_{cm}^2}{2v}} \left(\frac{x_{cm} - s_b'}{x_{cm}}\right) \frac{U_b'}{v} \int^{x+s_b'} e^{\frac{2ys_b'}{v} - \frac{s_b'^2}{2v} - \frac{(x_{cm} - s_b')s_b'}{v}} \, dy , \\ &\approx 2x_{cm} e^{\frac{x^2 - x_{cm}^2}{2v}} \cdot \frac{1}{2} \left(\frac{x_{cm} - s_b'}{x_{cm}}\right)^2 e^{-\frac{2s_b'}{v}\left(x_{cm} - \frac{3s_b'}{2} - x\right)} , \end{aligned} \tag{S65}$$

for $x \gtrsim x_{cm} - 3s_b' + O(v/s_b')$. More generally, for relative fitnesses in the range

$$x_{cm} - ns_b' - O(v/s_b') \gtrsim x \gtrsim x_{cm} - (n+1)s_b' + O(v/s_b') , \tag{S66}$$

one can continue this argument to show that

$$w_m(x) \approx 2x_{cm} e^{\frac{x^2 - x_{cm}^2}{2v}} \cdot \frac{1}{n!} \left(\frac{x_{cm} - s_b'}{x_{cm}}\right)^n e^{-\frac{ns_b'(x_{cm} - ns_b' - x)}{v} - \frac{n(n-1)s_b^2}{2v}} . \tag{S67}$$

This shows that each additional "step" reduces the original shoulder solution by factor of $\sim e^{n(n-1)s_b'^2/2v} \gg 1$. For fitness values that are many multiples of s_b' below x_{cm}, we can substitute $n \approx (x_{cm} - x)/s_b'$ to show that the leading order contribution to $w_m(x)$ scales as

$$\log w_m(x) \sim \text{const} - \frac{(x_{cm} - x)(x_{cm} - s_b'/2)}{v} , \tag{S68}$$

which obeys the required boundary condition that $w_m(x) \to 0$ as $x \to -\infty$. Thus, Eq. (S67) provides an asymptotic solution for $w_m(x)$ that is valid across the full range of relative fitnesses below x_{cm}.

4.1.3 First-order mutations and the rate of adaptation

When $\mu_m(s) \to \mu(s)$, our solution for $w_m(x)$ can be used to derive the fixation probabilities of first-order mutations, which have been studied in previous work (39, 40, 46). We reproduce these results here for completeness, using the new expressions for $w(x)$ that we have derived above. This will allow us to make comparisons to the second-order selection pressures analyzed in SI Section 4.1.4 below. (Readers who are familiar with this material may skip directly to SI Section 4.1.4.)

For a completely neutral mutation ($s_m = 0$), we can substitute our solution for $f(x)$ in Eq. (S38) into the self-consistency condition in Eq. (S35) to rewrite it in the convenient form:

$$\frac{1}{N} = \int f(x)w(x)\,dx = \frac{2x_c e^{-\frac{x_c^2}{2v}}}{\sqrt{2\pi v}} \int_{-\infty}^{x_c} \frac{w(x)}{2x_c e^{\frac{x^2-x_c^2}{2v}}}\,dx\,, \tag{S69}$$

which depends on the ratio between the true value of $w(x)$ and the exponential shoulder solution in Eq. (S42). Substituting our solution for $w(x)$ in Eq. (S67) then yields

$$\frac{1}{N} = \frac{2x_c e^{-\frac{x_c^2}{2v}}}{\sqrt{2\pi v}} \left(\int_{x_c - s_b}^{x_c} dx + \int_{x_c - 2s_b}^{x_c - s_b} e^{\frac{-s_b(x_c - s_b - x)}{v}}\,dx + \ldots \right)\,, \tag{S70}$$

where we have only included the $n = 0$ and $n = 1$ regions from Eq. (S67); all of the other terms are smaller by additional factors of $\exp\left(n(n-1)s_b^2/2v\right)$, and will therefore provide a negligible contribution when $s_b \gg \sqrt{v}$. Note that the first term in Eq. (S70) contributes equally over the whole range of $x \in (x_c - s_b, s_b)$, while the second term is dominated by fitnesses within $O(v/s_b)$ of $x_c - s_b$. Since $s_b \gg v/s_b$, the first term dominates, yielding

$$1 \approx \frac{2Nx_c s_b}{\sqrt{2\pi v}} e^{-\frac{x_c^2}{2v}}\,, \tag{S71}$$

which matches the condition previously derived in Ref. (46). When combined with the auxiliary condition for x_c in Eq. (S57), this allows us to solve for v and x_c as a function of the underlying parameters N, s_b, and U_b. An iterative solution of Eqs. (S71) and (S57) assuming that $x_c \gg s_b \gg \sqrt{v}$ yields the asymptotic expressions for v and x_c in Eq. (S49). The fixation probabilities of non-neutral mutations can be calculated using a similar procedure. We first use the self-consistency condition in Eq. (S71) to rewrite the scaled fixation probability in a similar form as Eq. (S69),

$$\tilde{p}_{\text{fix}}(s) = \int Nf(x-s)w(x)\,dx \approx \int_{-\infty}^{x_c+s} e^{\frac{xs}{v} - \frac{s^2}{2v}} \cdot \frac{w(x)}{2x_c e^{\frac{x^2-x_c^2}{2v}}} \cdot \frac{dx}{s_b}\,. \tag{S72}$$

This differs from the previous integral in Eq. (S69) by the presence of the exponential factor $e^{xs/v}$ and the shifting of the upper limit of integration from x_c to $x_c + s$. The behavior of this integral will therefore depend on the sign of s.

Beneficial mutations. When $s > 0$, the $e^{xs/v}$ term will enhance the contributions from higher fitness values, so the $n \geq 1$ terms in Eq. (S67) will continue to be negligible. Moreover, since the upper limit of integration is now larger than x_c, there will also be a new contribution from the Haldane region of the shoulder solution, so that

$$\begin{aligned}
\tilde{p}_{\text{fix}}(s) &\approx e^{\frac{x_c s}{v} - \frac{s^2}{2v}} \int_{x_c - s_b}^{x_c} e^{-\frac{(x_c - x)s}{v}} \frac{dx}{s_b} + \int_{x_c}^{x_c+s} \frac{2xe^{\frac{-(x-s)^2}{2v}}}{2x_c s_b e^{-\frac{x_c^2}{2v}}}\,, \\
&\approx e^{\frac{x_c s}{v} - \frac{s^2}{2v}} \cdot \frac{1 - e^{-\frac{s_b s}{v}}}{\frac{s_b s}{v}} + 2Ns \cdot \Phi\left(\frac{s - x_c}{\sqrt{v}}\right)\,,
\end{aligned} \tag{S73}$$

where $\Phi(z) \equiv \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{z} e^{-u^2/2} du$ is the Gaussian cumulative distribution function. The leading order scaling is given by

$$\log \tilde{p}_{\text{fix}}(s) \approx \begin{cases} \frac{x_c s}{v} & \text{if } 0 < s \ll x_c, \\ \log(2Ns) & \text{if } s \gg x_c, \end{cases} \tag{S74}$$

which transitions between a regime of strong clonal interference for $s \ll x_c$ and the Haldane limit when $s \gg x_c$ (39).

The fixation probabilities in Eq. (S74) have a simple heuristic interpretation. To be a successful, a moderately beneficial mutation ($s \ll x_c$) will typically arise on a background near $x = x_c - s$. This mutation will found a new lineage which experiences the effects of clonal interference and must generate multiple more fit descendants to take over and fix. Over this fixation time, $T_{sw} = x_c/v$, the beneficial mutation will provide an exponential advantage to its founding lineage. On the other hand, successful mutations that are extremely beneficial ($s \gg x_c$) will typically land above the interference threshold ($x > x_c - s$) and fix if they survive genetic drift.

Deleterious mutations. The fixation probabilities of deleterious mutations ($s < 0$) can be calculated using a similar procedure. In this case, however, the $e^{-x|s|/v}$ term becomes increasingly large at lower fitness values, and will need to be cut off by one of the $n \geq 1$ terms in Eq. (S67). To build intuition, let us first consider the case where $|s| < s_b$, so that

$$\tilde{p}_{\text{fix}}(s) \approx e^{-\frac{s^2}{2v}} \left[\int_{x_c - s_b}^{x_c - |s|} e^{-\frac{x|s|}{v}} \frac{dx}{s_b} + \left(\frac{x_c - s_b}{x_c} \right) \int_{x_c - 2s_b}^{x_c - s_b} e^{\frac{-(x_c - s_b - x)s_b}{v} - \frac{x|s|}{v}} \frac{dx}{s_b} + \dots \right] . \tag{S75}$$

The contribution from the $n = 0$ term is now peaked at its lower limit of integration ($x \approx x_c - s_b$). However, when $|s|$ is smaller than s_b, the contribution from the $n = 1$ term is still peaked at its upper limit of integration ($x \approx x_c - s_b$), which implies that the total contribution to the fixation probability will also be peaked at $x \approx x_c - s_b$. [The contributions from the $n \geq 2$ terms will be smaller by additional factors of $\exp(n(n-1)s_b^2/2v)$, and will therefore be negligible in the limit that $s_b \gg \sqrt{v}$.] Evaluating the two integrals then yields

$$\tilde{p}_{\text{fix}}(s) \approx e^{-\frac{(x_c - s_b)|s|}{v} - \frac{|s|^2}{2v}} \left(\frac{1 - e^{-\frac{(s_b - |s|)|s|}{v}}}{\frac{s_b|s|}{v}} + \left(\frac{x_c - s_b}{x_c} \right) \frac{1 - e^{-\frac{s_b(s_b - |s|)}{v}}}{\frac{s_b(s_b - |s|)}{v}} \right) , \tag{S76}$$

which is valid for fitness costs $s \in (-s_b + O(v/s_b), 0)$. When $|s| \ll s_b$, this matches the fixation probability derived using the thresholding approximation in Ref. (46), but it starts to deviate from Ref. (46) for $|s| \sim O(s_b)$. We expect that the expression in Eq. (S76) will be more accurate in this case, since it better captures the behavior of $w(x)$ below $x \approx x_c - s_b$.

The fixation probabilities of more strongly deleterious mutations can be calculated in a similar manner. In this case, it will be convenient to write

$$|s| = ks_b + \Delta , \tag{S77}$$

where k is the largest integer such that $\Delta > 0$. The upper limit of integration at $x_c - s \equiv x_c - ks_b - \Delta$ implies that the first term with nonzero contribution to p_{fix} will be the $n = k$ term, so that

$$\begin{aligned} \tilde{p}_{\text{fix}}(s) \approx \frac{e^{-\frac{s^2}{2v}}}{k!} \left(1 - \frac{s_b}{x_c}\right)^k e^{-\frac{k(k-1)s_b^2}{2v}} \Bigg[& \int_{x_c - (k+1)s_b}^{x_c - ks_b - \Delta} e^{-\frac{x(ks_b + \Delta)}{v} - \frac{ks_b(x_c - ks_b - x)}{v}} \frac{dx}{s_b} \\ & + \frac{e^{-\frac{ks_b^2}{v}}}{k+1} \left(\frac{x_c - s_b}{x_c} \right) \int_{x_c - (k+2)s_b}^{x_c - (k+1)s_b} e^{-\frac{x(ks_b + \Delta)}{v} - \frac{(k+1)s_b(x_c - (k+1)s_b - x)}{v}} \frac{dx}{s_b} + \dots \Bigg] . \end{aligned} \tag{S78}$$

As above, the contribution from the $n = k$ term is peaked at the lower limit of integration ($x \approx x_c - (k+1)s_b$), while the contribution from the $n = k+1$ term is peaked at its upper limit ($x \approx x_c - (k+1)s_b$). The terms with

$n \geq k+2$ are smaller by exponential factors of $O(s_b^2/v)$, and will therefore provide a negligible contribution to the fixation probability when $s_b \gg \sqrt{v}$. Evaluating the two integrals then yields,

$$\begin{aligned}\tilde{p}_{\text{fix}}(s) &\approx \frac{e^{-\frac{s^2}{2v}}}{k!}\left(1-\frac{s_b}{x_c}\right)^k e^{-\frac{k(k-1)s_b^2}{2v}}\left[e^{-\frac{(x_c-(k+1)s_b)\Delta}{v}-\frac{ks_b(x_c-ks_b)}{v}}\left(\frac{1-e^{-\frac{\Delta(s_b-\Delta)}{v}}}{\frac{s_b\Delta}{v}}\right)\right.\\ &\left.+\frac{e^{-\frac{ks_b^2}{v}}}{k+1}\left(\frac{x_c-s_b}{x_c}\right)e^{-\frac{(x_c-(k+1)s_b)(ks_b+\Delta)}{v}}\left(\frac{1-e^{-\frac{s_b(s_b-\Delta)}{v}}}{\frac{s_b(s_b-\Delta)}{v}}\right)\right],\\ &\approx e^{-\frac{\left(x_c-\frac{s_b}{2}\right)|s|}{v}}\frac{e^{-\frac{\Delta^2}{2v}+\frac{s_b\Delta}{2v}}}{k!}\left(1-\frac{s_b}{x_c}\right)^k\left[\left(\frac{1-e^{-\frac{\Delta(s_b-\Delta)}{v}}}{\frac{s_b\Delta}{v}}\right)+\frac{\left(\frac{x_c-s_b}{x_c}\right)}{k+1}\left(\frac{1-e^{-\frac{s_b(s_b-\Delta)}{v}}}{\frac{s_b(s_b-\Delta)}{v}}\right)\right],\end{aligned} \tag{S79}$$

whose leading order behavior simplifies to

$$\log\tilde{p}_{\text{fix}}(s) \approx \begin{cases}-\frac{(x_c-s_b)|s|}{v} & \text{if } s<0 \text{ and } |s| \ll s_b,\\ -\frac{(x_c-\frac{s_b}{2})|s|}{v} & \text{if } s<0 \text{ and } |s| \gg s_b.\end{cases} \tag{S80}$$

For the purposes of numerical evaluation, it is useful to employ a modified version of Eq. (S79),

$$\begin{aligned}\tilde{p}_{\text{fix}}(s) &\approx e^{-\frac{\left(x_c-\frac{s_b}{2}\right)|s|}{v}}\frac{e^{-\frac{\Delta^2}{2v}+\frac{s_b\Delta}{2v}}}{k!}\left(1-\frac{s_b}{x_c}\right)^k\\ &\times\left[\left(\frac{1-e^{-\frac{\Delta(s_b-\Delta)}{v}}}{\frac{s_b\Delta}{v}}\right)\left(1-e^{-\frac{s_b(s_b-\Delta)}{v}}\right)+\frac{\left(\frac{x_c-s_b}{x_c}\right)}{k+1}\left(\frac{1-e^{-\frac{s_b(s_b-\Delta)}{v}}}{\frac{s_b(s_b-\Delta)}{v}}\right)\left(1-e^{-\frac{\Delta s_b}{v}}\right)\right],\end{aligned} \tag{S81}$$

which has the same asymptotic limit when $v \ll s_b^2$, but enforces strict continuity at $\Delta = s_b$.

4.1.4 Modifiers without direct costs or benefits

We now are in a position to calculate the fixation probabilities of modifier mutations ($\mu_m(s) \neq \mu(s)$). In the absence of a direct cost or benefit ($s_m = 0$), this will only be a slight generalization of the neutral fixation probability calculation in SI Section 4.1.3. In the case of a modifier mutation, Eq. (S69) becomes

$$\tilde{p}_{\text{fix}}(s_b, U_b \to s_b', U_b') = \int Nf(x)w_m(x) \approx \frac{x_{cm}s_b'}{x_c s_b}e^{\frac{x_c^2-x_{cm}^2}{2v}}\int_{-\infty}^{x_c}\frac{w_m(x)}{2x_{cm}e^{\frac{x^2-x_{cm}^2}{2v}}}\frac{dx}{s_b'}, \tag{S82}$$

where the primary difference is the presence of the prefactor term, and the fact that the upper limit of integration is now equal to $x_c \neq x_{cm}$. The behavior of this integral will sensitively depend on the relative magnitudes of x_{cm} and x_c.

Positively selected modifiers. If the modifier interference threshold is less than that of the wildtype ($x_{cm} < x_c$), then the integral in Eq. (S82) will contain both the $n \geq 0$ terms as well as a portion of the Haldane

region,

$$\tilde{p}_{\text{fix}}(s_b, U_b \to s_b', U_b') \approx \frac{x_{cm}s_b'}{x_c s_b} e^{\frac{x_c^2 - x_{cm}^2}{2v}} \left[\int_{-\infty}^{x_{cm}} \frac{w_m(x)}{2x_{cm} e^{\frac{x^2 - x_{cm}^2}{2v}}} \frac{dx}{s_b'} + \int_{x_{cm}}^{x_c} \frac{2x}{2x_{cm} e^{\frac{x^2 - x_{cm}^2}{2v}}} \frac{dx}{s_b'} \right],$$
$$\approx \frac{x_{cm}s_b'}{x_c s_b} e^{\frac{x_c^2 - x_{cm}^2}{2v}} \left[1 + \frac{v}{x_{cm}s_b'} \left(1 - e^{\frac{x_{cm}^2 - x_c^2}{2v}} \right) \right]. \tag{S83}$$

The second term is negligible in our regime of interest where $x_{cm}s_b'/v \gg 1$, so the fixation probability reduces to

$$\tilde{p}_{\text{fix}}(s_b, U_b \to s_b', U_b') \approx \frac{x_{cm}s_b'}{x_c s_b} e^{\frac{x_c^2 - x_{cm}^2}{2v}}. \tag{S84}$$

Since $x_{cm} < x_c$, Eq. (S84) implies that $\tilde{p}_{\text{fix}} > 1$ (i.e., the modifier is favored by natural selection).

Negatively selected modifiers. In the opposite case, where the interference threshold of the modifier is greater than that of the wildtype ($x_{cm} > x_c$), the upper limit of the integral in Eq. (S82) will occur somewhere in the interference region of $w_m(x)$. The precise behavior will depend on how far x_c extends into this region. If $x_c > x_{cm} - s_b' + O(v/s_b')$, then the dominant contribution will still come from the $n = 0$ region of Eq. (S67), so that

$$\tilde{p}_{\text{fix}}(s_b, U_b \to s_b', U_b') \approx \frac{x_{cm}s_b'}{x_c s_b} e^{\frac{x_c^2 - x_{cm}^2}{2v}} \left(1 - \frac{x_{cm} - x_c}{s_b'} \right). \tag{S85}$$

Since $x_{cm} > x_c$, this implies that $p_{\text{fix}}/p_0 < 1$ (i.e., the modifier is disfavored by natural selection). More generally, if

$$x_c = x_{cm} - ks_b' - \Delta, \tag{S86}$$

for some $\Delta \in (0, s_b')$, then the primary contribution will come from the $n = k$ region of Eq. (S67). This yields

$$\tilde{p}_{\text{fix}}(s_b, U_b \to s_b', U_b') \approx \frac{x_{cm}s_b'}{x_c s_b} e^{\frac{x_c^2 - x_{cm}^2}{2v}} \frac{\left(\frac{x_{cm} - s_b'}{x_{cm}}\right)^k}{k!} e^{-\frac{k(k-1)s_b'^2}{2v}} \int_{x_{cm}-(k+1)s_b'}^{x_{cm}-ks_b'-\Delta} e^{-\frac{ks_b'(x_{cm}-ks_b'-x)}{v}} \frac{dx}{s_b'},$$
$$\approx \frac{x_{cm}s_b'}{x_c s_b} e^{\frac{x_c^2 - x_{cm}^2}{2v}} \frac{\left(\frac{x_{cm} - s_b'}{x_{cm}}\right)^k}{k!} e^{-\frac{k(k-1)s_b'^2}{2v}} \frac{v}{ks_b'^2} e^{-\frac{ks_b'\Delta}{v}}. \tag{S87}$$

Predictions for specific values of s_b' and U_b'. We can express the fixation probability of a modifier in terms of the underlying parameters, s_b', U_b', and N by substituting Eq. (S58) into Eq. (S84) to obtain,

$$\log \tilde{p}_{\text{fix}}(s_b, U_b \to s_b', U_b') \approx \left(1 - \frac{s_b^2}{s_b'^2}\right)\left(\frac{x_c^2}{2v} - \frac{x_c s_b}{2v}\right) + \frac{x_c}{s_b}\frac{s_b^2}{s_b'^2}\log\left(\frac{U_b'}{U_b}\right) - \frac{v}{2s_b'^2}\log^2\left(\frac{U_b'}{U_b}\right)$$
$$+ \frac{1}{2}\left(1 - \frac{s_b^2}{s_b'^2}\right)\log\left(\frac{U_b'}{U_b}\right), \tag{S88}$$

where we have assumed that $x_{cm} - s'_b \gg \sqrt{v}$. In the case of a selection strength modifier ($U'_b = U_b$), the leading order contributions simplify to

$$\log \tilde{p}_{\text{fix}}(s_b \to s'_b) \approx \left(1 - \frac{s_b^2}{s_b'^2}\right)\left(\frac{x_c^2}{2v} - \frac{x_c s_b}{2v}\right) \approx \left(1 - \frac{s_b^2}{s_b'^2}\right)\left(\log\sqrt{(Ns_b)\,(NU_b)}\right) , \tag{S89}$$

where we have substituted $x_c^2/2v \approx \log(Ns_b)$ and $x_c s_b/v \approx \log(s_b/U_b)$ on the right hand side.

The fixation probability of a selection strength modifier has a simple heuristic explanation. For small changes in the selection coefficient ($s'_b - s_b \ll s'_b$), successful mutations arise in the high-fitness "nose" of $f(x)$ ($x \approx x_c \gg s_b$) and must acquire $\sim x_c/s_b$ additional mutations before they reach $O(1)$ frequencies. In each of these steps, a selection-strength modifier produces $\sim \exp\left[\frac{s_b}{v} \cdot j(s'_b - s_b)\right]$ more mutations than a wildtype individual with the same fitness, leading to the exponential scaling observed in Eq. (3).

Similarly, the fixation probability of a mutation rate modifier is given by setting $s'_b = s_b$ in Eq. (S88), which yields the leading order contribution

$$\log \tilde{p}_{\text{fix}}(U_b \to U'_b) \approx \frac{x_c}{s_b}\log\left(\frac{U'_b}{U_b}\right) - \frac{v}{2s_b^2}\log^2\left(\frac{U'_b}{U_b}\right) \approx \frac{x_c}{s_b}\log\left(\frac{U'_b}{U_b}\right) , \tag{S90}$$

where we have assumed that $x_c/s_b \approx 2\log(Ns_b)/\log(s_b/U_b)$ and $\log\left(U'_b/U_b\right) \ll x_c s_b/v \approx \log(s_b/U_b) \gg 1$. This matches the result previously derived in Ref. (24). The heuristic picture is similar to the one described in that work: a mutation rate modifier must also arise in the high-fitness nose of the fitness distribution ($x \approx x_c$) and accrue x_c/s_b additional mutations to fix. In each of these x_c/s_b steps before a coalescent event, the mutation rate modifier produces U'_b/U_b more offspring, leading to the observed scaling in Eq. (S90).

More generally, we see that the fixation probability of a joint selection strength and mutation rate modifier can be expressed in terms of the fixation probabilities of these stand-alone modifiers,

$$\log \tilde{p}_{\text{fix}}(U_b, s_b \to U'_b, s'_b) \approx \log \tilde{p}_{\text{fix}}(s_b \to s'_b) + \alpha \cdot \log \tilde{p}_{\text{fix}}(U_b \to U'_b) , \tag{S91}$$

where $\alpha = s_b^2/s_b'^2$. The presence of this factor α indicates that a change to the selection strength of mutations can modulate the effect of a linked mutation rate change. The effect of this interplay can be extremely important to determining the fate a modifier mutation, causing modifiers that would be disfavored by natural selection without this modulation to be favored. Interestingly, the modulation effect only works in this direction: modifiers that would be favored without the effect cannot actually be deleterious.

4.1.5 Modifiers with direct costs or benefits

Finally, we can extend these calculations to modifiers with direct costs or benefits by combining our results from SI Sections 4.1.3 and 4.1.4. Generalizing Eqs. (S72) or (S82) to this case yields

$$\begin{aligned}\tilde{p}_{\text{fix}}(\mu \to \mu_m; s_m) &= \int N f(x - s_m) w_m(x)\, dx \\ &\approx \frac{x_{cm} s'_b}{x_c s_b} e^{\frac{x_c^2 - x_{cm}^2}{2v} - \frac{s_m^2}{2v}} \int_{-\infty}^{x_c + s_m} e^{\frac{x s_m}{v}} \cdot \frac{w_m(x)}{2x_{cm} e^{\frac{x^2 - x_{cm}^2}{2v}}} \cdot \frac{dx}{s'_b} .\end{aligned} \tag{S92}$$

Our analysis above suggests that the two key considerations will be (i) the sign of s_m, and (ii) where $x_c + s_m$ falls relative to the internal scales of $w_m(x)$ (e.g. x_{cm}, $x_{cm} - s'_b$, etc.).

Modifiers with direct fitness benefits. For a modifier with a direct fitness benefit ($s_m > 0$), the exponential term will once again amplify the contributions from higher fitness values. The behavior of Eq. (S92) will therefore strongly depend on the location of the upper limit of integration. If $s_m - (x_{cm} - x_c) > O(v/x_{cm})$, there will be a new contribution from the Haldane region of $w_m(x)$, so that

$$p_{\text{fix}}(\mu \to \mu_m; s_m) \approx \frac{x_{cm} s_b'}{x_c s_b} e^{\frac{x_c^2 - x_{cm}^2}{2v} - \frac{s_m^2}{2v}} \int_{x_{cm} - s_b}^{x_{cm}} e^{\frac{x s_m}{v}} \cdot \frac{dx}{s_b'} + \int_{x_{cm}}^{x_{cm} + s_m - (x_{cm} - x_c)} \frac{2Nx e^{-\frac{(x - s_m)^2}{2v}}}{\sqrt{2\pi v}} dx \,. \tag{S93}$$

Since $x_c \gg \sqrt{v}$, the upper limit from the Haldane region can be taken to infinity, so that Eq. (S93) reduces to

$$\tilde{p}_{\text{fix}}(\mu \to \mu_m; s_m) \approx \frac{x_{cm} s_b'}{x_c s_b} e^{\frac{x_c^2 - x_{cm}^2}{2v}} \left(e^{\frac{x_{cm} s_m}{v} - \frac{s_m^2}{2v}} \cdot \frac{1 - e^{-\frac{s_b' s_m}{v}}}{\frac{s_b' s_m}{v}} \right) + 2N s_m \Phi\left(\frac{s_m - x_{cm}}{\sqrt{v}} \right) . \tag{S94}$$

In the opposite case, where $s_m \lesssim x_{cm} - x_c - O(v/x_{cm})$, the upper limit of integration in Eq. (S92) will fall below the interference threshold of $w_m(x)$. When $s_m > 0$, this will only occur when $x_{cm} > x_c$ (i.e., when the modifier would be disfavored on its own). Similar to the pure modifier case in Eqs. (S85) and (S87), the behavior of Eq. (S92) will depend on where $x_c + s_m$ falls relative to the internal scales of $w_m(x)$. If $x_{cm} \gtrsim x_c + s_m \gtrsim x_{cm} - s_b'$, then the dominant contribution will come from the $n = 0$ region of Eq. (S67), so that

$$\tilde{p}_{\text{fix}}(\mu \to \mu_m; s_m) \approx \frac{x_{cm} s_b'}{x_c s_b} e^{\frac{x_c^2 - x_{cm}^2}{2v}} \cdot e^{\frac{x_{cm} s_m}{v} - \frac{s_m^2}{2v}} \cdot \frac{e^{\frac{-(x_{cm} - x_c - s_m) s_m}{v}} \left(1 - e^{-\frac{(s_b' + x_c - x_{cm} - s_m) s_m}{v}} \right)}{\frac{s_b' s_m}{v}} . \tag{S95}$$

More generally, if we have

$$x_c + s_m = x_{cm} - k s_b' - \Delta \,, \tag{S96}$$

for some $\Delta \in (0, s_b')$, then the primary contribution will come from the $n = k$ region of Eq. (S67). This yields

$$\begin{aligned} \tilde{p}_{\text{fix}}(\mu \to \mu_m; s_m) &\approx \frac{x_{cm} s_b'}{x_c s_b} e^{\frac{x_c^2 - x_{cm}^2}{2v}} \frac{\left(\frac{x_{cm} - s_b'}{x_{cm}} \right)^k}{k!} e^{-\frac{k(k-1) s_b'^2}{2v}} \int_{x_{cm} - (k+1) s_b'}^{x_{cm} - k s_b' - \Delta} e^{\frac{x s_m}{v} - \frac{s_m^2}{2v} - \frac{k s_b' (x_{cm} - k s_b' - x)}{v}} \frac{dx}{s_b'} \,, \\ &\approx \frac{x_{cm} s_b'}{x_c s_b} e^{\frac{x_c^2 - x_{cm}^2}{2v} + \frac{x_{cm} s_m}{v} - \frac{s_m^2}{2v}} \frac{\left(\frac{x_{cm} - s_b'}{x_{cm}} \right)^k e^{-\frac{k(k-1) s_b'^2}{2v} - \frac{(k s_b' + \Delta) s_m}{v} - \frac{k s_b' \Delta}{v}} \left(1 - e^{-\frac{(s_m + k s_b')(s_b' - \Delta)}{v}} \right)}{k! \left(\frac{(s_m + k s_b') s_b'}{v} \right)} . \end{aligned} \tag{S97}$$

Modifiers with direct fitness costs. For a modifier with a direct fitness cost ($s_m < 0$), the exponential term will once again amplify contributions from lower fitness values, and must eventually be cut off by the $n = k + 1$ term in Eq. (S67), where k is defined by

$$|s_m| = k s_b' + \Delta \,, \tag{S98}$$

for some $\Delta \in (0, s_b')$. The major difference from the first-order selection scenario in Eq. (S81) is that the upper limit of integration includes an additional term, $x_c - x_{cm}$. If $x_{cm} < x_c$ (i.e. a positively selected modifier), then

the upper limit of integration is larger than in Eq. (S81). This can in principle lead to contributions from the $n < k$ terms in Eq. (S67). However, the $n < k$ terms will all be smaller by exponential factors of $O(s_b^2/v)$ so they will provide a negligible contribution when $s_b \gg \sqrt{v}$. The larger upper limit of integration will therefore only alter the $n = k$ integral in Eq. (S78), shifting it from $x_{cm} - ks'_b - \Delta \to \min\{x_c - ks'_b - \Delta, x_{cm} - ks'_b\}$. This implies that the modifier fixation probability will be given by a slight generalization of Eq. (S81),

$$\begin{aligned}\tilde{p}_{\text{fix}}(\mu \to \mu_m; s_m) \approx \frac{x_{cm}s'_b}{x_c s_b} e^{\frac{x_c^2 - x_{cm}^2}{2v}} e^{-\frac{\left(x_{cm} - \frac{s'_b}{2}\right)|s_m|}{v}} \frac{e^{-\frac{\Delta^2}{2v} + \frac{s'_b \Delta}{2v}}}{k!} \left(1 - \frac{s'_b}{x_{cm}}\right)^k \\ \times \left[\left(\frac{1 - e^{-\frac{\Delta \cdot \min(s'_b, s'_b - \Delta + x_c - x_{cm})}{v}}}{\frac{s'_b \Delta}{v}}\right) \cdot \left(1 - e^{-\frac{s'_b(s'_b - \Delta)}{v}}\right)\right. \\ \left. + \frac{\left(\frac{x_{cm} - s'_b}{x_{cm}}\right)}{k+1} \left(\frac{1 - e^{-\frac{s'_b(s'_b - \Delta)}{v}}}{\frac{s'_b(s'_b - \Delta)}{v}}\right)\left(1 - e^{-\frac{\Delta s'_b}{v}}\right)\right] ,\end{aligned} \tag{S99}$$

which will be valid for $x_{cm} < x_c$. The opposite scenario ($x_{cm} > x_c$) corresponds to a negatively selected modifier that has an additional a direct cost. We will not consider such mutations here, as they will have a negligible chance of fixing.

Leading order scaling. Combining these expressions, we see that the leading-order solution for the fixation probability of a modifier with a direct cost or benefit can be summarized as,

$$\log \tilde{p}_{\text{fix}}(\mu \to \mu_m; s_m) \approx \begin{cases} \frac{x_c^2}{2v} - \frac{x_{cm}^2}{2v} + \frac{(x_{cm} - s'_b)s_m}{v} & s'_b \gg |s_m|; s_m < 0 , \\ \frac{x_c^2}{2v} - \frac{x_{cm}^2}{2v} + \frac{(x_{cm} - s'_b/2)s_m}{v} & s'_b \ll |s_m|, s_m < 0 , \\ \frac{x_c^2}{2v} - \frac{x_{cm}^2}{2v} + \frac{x_{cm}s_m}{v} & x_c \gg |s_m|, s_m > 0 . \end{cases} \tag{S100}$$

Interestingly, we see that this fixation probability decomposes into contributions from first- and second-order selection,

$$\log \tilde{p}_{\text{fix}}(\mu \to \mu_m; s_m) \approx \log \tilde{p}_{\text{fix}}(U_b, s_b \to U'_b, s'_b) + \gamma \cdot \log \tilde{p}_{\text{fix}}(s_m) , \tag{S101}$$

where $\tilde{p}_{\text{fix}}(s_b, U_b \to s'_b, U'_b)$ is the fixation probability of a modifier without direct costs or benefits, $\tilde{p}_{\text{fix}}(s_m)$ is the fixation probability of a first order mutation, and the weighting factor γ satisfies $\gamma = \min(x_{cm}/x_c, \sqrt{2s_b/x_c})$. The presence of this factor γ demonstrates that second order selection modulates the effects of first order selection; the fixation probability of a deleterious or beneficial mutation arising in the background population scales with x_c/v, while the contribution of a direct cost or benefit scales with x_{cm}/v. This means that a modifier without direct costs or benefits that is favored by natural selection ($x_{cm} < x_c$) suppresses the contribution of a direct cost, while a modifier without direct costs or benefits that is disfavored by natural selection ($x_{cm} > x_c$) amplifies the contribution of a direct benefit. This effect can be extremely large, affecting the contribution of a direct cost or benefit and overall favorability of a modifier mutation by orders of magnitude.

For sufficiently large direct benefits ($|s_m| \gg x_c$), we expect that even the most deleterious modifiers must fix deterministically upon surviving genetic drift, irrespective of the background on which they arise ($p_{\text{fix}}/p_0 \approx 2Ns_m$). This expectation follows from the fact that a modifier with a large direct benefit will jump far out ahead of the 'traveling wave' and take over the population long before the background population is able to 'catch-up'. This is true even in the extreme case of a dead-end modifier, which would drive the rate

of adaptation to zero upon fixing. Our current theoretical approach, however, does not enable us to predict the fixation probability of a modifier with direct benefits $s_m \sim O(x_c)$. In particular, to derive Eq. (S33) we assumed that the fate of a modifier mutation is determined at small frequency while competing against the mean fitness set by the rate of adaptation of the background population. However, if the rate of adaptation does not change, our theory predicts the background population will always "catch up" and pass a deleterious modifier in fitness, no mater how large the direct benefit. This pathology can be seen in Eq. (S56), where $\lim_{s_b' \to 0} x_{cm} = \infty$.

This limitation prevents us from understanding a key aspect of the trade-off between short-term fitness and long-term evolvability. How large a direct benefit is sufficient to drive a deleterious modifier to be favored? In the next section, we focus on the extreme case of an evolutionary dead-end ($U_b' = 0$ or $s_b' = 0$), showing that even modest direct benefits can drive these modifiers to be favored by selection.

4.1.6 Fixation of a dead-end modifier

When a dead-end modifier arises with an extremely large direct benefit ($|s_m| \gg x_c$), it will sweep long before the background population is able to catch-up in fitness. On the other hand, if the direct benefit is not sufficiently large, the background population will surpass the modifier lineage in fitness. We can understand this cross-over by analyzing the dynamics of a modifier lineage arising with landing fitness, x, as it transitions from small to large frequency.

A newly established modifier clone with relative fitness x will initially start to grow deterministically as

$$f_m(t) \approx \frac{e^{xt - \frac{vt^2}{2}}}{2Nx}\,, \tag{S102}$$

where $1/2Nx$ corresponds to the size of the clone immediately after it establishes. This will be a good approximation as long as the modifier frequency remains small ($f_m \ll 1$), so that the mean fitness of the population is still primarily determined by the wildtype ($\partial_t \overline{X}(t) \approx v$). If $x \ll \sqrt{2v \log(N\sqrt{v})}$, then the modifier clone will remain small throughout its entire lifetime, and the wildtype population will eventually pass it by. However, for larger initial fitnesses, the modifier clone will eventually grow to $O(1)$ frequencies, and will start to exert its own effect on the population mean fitness. At this point, the dynamics in Eq. (S102) will start to break down.

The ultimate fate of a dead-end modifier will depend on how much fitness the wildtype population has gained during this time. To understand this process, let t^* denote the time required for the modifier clone to reach a reference frequency f_m^*. If $f_m^* \ll 1$, then t^* can be calculated from Eq. (S102):

$$\log f_m^* = -\log(2Nx) + xt^* - \frac{vt^{*2}}{2}\,. \tag{S103}$$

During this time, the wildtype population will have increased in fitness by a total amount $\Delta x = vt^*$, so that the current relative fitness of the modifier is $x - vt^*$. If the nose of the wildtype population has surpassed the modifier in this time ($x_c \gtrsim x - vt^*$), then regardless of how much the modifier grows in the short-term, it will be destined for extinction in the long-term, since it is unable to produce additional mutations.

On the other hand, if $x - vt^* > x_c$, then it is possible for the modifier to sweep through the rest of population extremely rapidly, and "freeze" the wildtype population in its place. If further adaptation of the wildtype can be neglected, then the modifier will transition from from frequency f_m^* to $1 - f_m^*$ according to the logistic dynamics,

$$\frac{\partial f_m}{\partial t} = (x - vt^*) f_m (1 - f_m)\,, \tag{S104}$$

which requires an additional time interval

$$\Delta t \approx \frac{2}{x - vt^*} \log\left(\frac{1}{f_m^*}\right) . \tag{S105}$$

Since we have assumed that $x - vt^* \geq x_c$, it is always possible to choose a reference frequency f_m^* that is $\ll 1$, but large enough that Eq. (S105) is much smaller than the time required for the wildtype population to acquire one additional mutation (s_b/v). For example, choosing $f_m^* = \exp(-\sqrt{x_c s_b/v}) \ll 1$ yields

$$\frac{v\Delta t}{s_b} \leq \frac{2v}{x_c s_b}\sqrt{\frac{x_c s_b}{v}} \ll 1 , \tag{S106}$$

since $x_c s_b/v \gg 1$. This shows that the wildtype population is effectively frozen in place while the modifier transitions from f_m^* to $1 - f_m^*$. At this point, the mean fitness of the population is dominated by the modifier ($\overline{X} \approx x - vt^*$). Since we have assumed that $x - vt^* > x_c$, this implies that even the fittest individuals in the wildtype population now have a negative relative fitness, so their further adaptation will be effectively halted. The modifier clone will therefore continue to sweep through the population, and will fix in a time of order $\sim(x - vt^*)^{-1}\log(N(x - vt^*)f_m^*)$.

Based on this reasoning, we conclude that a newly established modifier will fix if its initial fitness exceeds a critical threshold x^* defined by

$$x^* - vt^* \equiv x_c + \Delta x , \tag{S107}$$

where Δx is a small correction term [$\lesssim O(s_b)$] that accounts for potential ambiguities in the definition of the nose, as well as the possibility that a leapfrogged nose may produce one additional mutation via stochastic tunneling (91) before it goes extinct. It is also accounts for the fact that a small subclass of individuals in the background population will need to have fitness greater than x^* to out-compete the modifier after the mean fitness changes. The size of this subclass can be determined by requiring that the sum of the establishment probabilities of the individuals that have crossed x^* be equal to 1,

$$1 \approx \int_{x^*}^{x_c+vt^*} \frac{2N(x - x^*)}{\sqrt{2\pi v}} e^{-\frac{(x-vt^*)^2}{2v}} . \tag{S108}$$

Assuming $v/x_c \ll x^* - vt^* - x_c$ and $x_c - (x^* - vt^* - x_c) \gg \sqrt{v}$, we can evaluate this integral with a Laplace approximation to obtain the self-consistent solution,

$$\Delta x \equiv x^* - vt^* - x_c \approx -\frac{v}{x_c}\log\left(\frac{x_c s_b}{v}\cdot\frac{x_c^2}{v}\right) . \tag{S109}$$

This constitutes a small correction to x^* (as expected), but we will see that it provides an important contribution to the fixation probability below. To solve for x^*, we can substitute Eq. (S109) into Eq. (S103) to obtain the condition,

$$\frac{x_c^2}{2v}\left[\left(\frac{x^*}{x_c}\right)^2 - 2 - \frac{2\Delta x}{x_c}\right] = \log f_m^* - \log\left(\frac{s_b}{\sqrt{2\pi v}}\right) + \log\left(\frac{x^*}{x_c}\right) , \tag{S110}$$

where we have also used Eq. (S71) to substitute for N. In our regime of interest where $x_c \gg \sqrt{v}$, the terms on the left will be asymptotically larger than those on the right, and the dominant balance will be given by the terms in the square brackets, yielding

$$x^* \approx \sqrt{2}\cdot x_c\left(1 + \frac{\Delta x}{2x_c}\right) \approx \sqrt{2}\cdot x_c\left[1 - \frac{v}{2x_c^2}\log\left(\frac{x_c s_b}{v}\cdot\frac{x_c^2}{v}\right)\right] . \tag{S111}$$

This result implies that a modifier lineage with relative fitness greater than $x^* \approx \sqrt{2} \cdot x_c$ will be guaranteed to fix as long as it survives genetic drift while rare — regardless of its ability to produce additional mutations. This will always be consistent with the wildtype $w(x)$ function from the branching process approximation in SI Section 4.1.3, since $x^* > x_c$. However, it suggests that the interference threshold for the modifier lineage must be capped at a maximum value,

$$x_{cm} \approx \min\left\{x_{cm}^0, x^*\right\} , \tag{S112}$$

where x_{cm}^0 is our original solution for the interference threshold derived in SI Section 4.1.1 above. This constitutes a simple modification of the original branching process model in Eq. (1) that can capture some of these non-linear feedback effects.

In the extreme case of a dead-end modifier ($\mu \to \varnothing$), this ansatz leads to a conditional fixation probability of the form

$$w_m(x) \approx \begin{cases} 2x & \text{if } x \geq x^*, \\ 0 & \text{else .} \end{cases} \tag{S113}$$

Substituting this result into Eq. (S92) yields a prediction for the total fixation probability of a dead-end modifier with a direct benefit s_m:

$$\tilde{p}_{\text{fix}}(\mu \to \varnothing, s_m) \approx \begin{cases} \int_{x^*}^{x_c+s_m} \frac{x e^{-\frac{(x-s_m)^2}{2v}+\frac{x_c^2}{2v}}}{x_c s_b}\, dx & \text{if } s_m > x^* - x_c, \\ 0 & \text{else.} \end{cases} \tag{S114}$$

Evaluating the integral in the upper branch yields,

$$\int_{x^*}^{x_c+s_m} \frac{x e^{-\frac{(x-s_m)^2}{2v}+\frac{x_c^2}{2v}}}{x_c s_b}\, dx = \frac{v}{x_c s_b}\left[e^{\frac{x_c^2-(x^*-s_m)^2}{2v}} - 1\right] + 2Ns_m\left[\Phi\left(\frac{s_m - x^*}{\sqrt{v}}\right) - \Phi\left(\frac{-x_c}{\sqrt{v}}\right)\right] , \tag{S115}$$

where $\Phi(z) = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{z} e^{-u^2/2}\, du$ is the Gaussian cumulative distribution function. The leading-order scaling is given by

$$\tilde{p}_{\text{fix}}(\mu \to \varnothing, s_m) \approx \begin{cases} 2Ns_m & \text{if } s_m \gtrsim x^* + O(\sqrt{v}), \\ \frac{v x^*}{x_c s_b (x^*-s_m)} e^{\frac{x_c(s_m - x^* + x_c)}{v} - \frac{(s_m - x^* + x_c)^2}{2v}} & \text{if } x^* - O(\sqrt{v}) \gtrsim s_m \gtrsim x^* - x_c + O(\sqrt{v}), \\ 0 & \text{else.} \end{cases} \tag{S116}$$

These results suggests that modest direct benefits of size $s_m \gtrsim x^* - x_c \approx (\sqrt{2}-1)x_c$ are sufficient to cause an evolutionary dead-end to be favored by natural selection. While this critical size is larger than the advantage of a single mutation (s_b), it is smaller than the total fitness variation in the population (x_c), and is only weakly dependent on NU_b. This contrasts with the traditional linear scaling observed in models with small numbers of competing loci (92), highlighting the unique features that arise in genome-wide models of adaptation.

Together, this analysis shows that the fate of a dead-end modifier is still determined by chance events that occur while the modifier lineage is at low frequency (e.g. its random genetic background and whether it survives genetic drift while rare) even though its ability to take over critically relies on non-linear feedbacks that occur while it is common.

4.1.7 Relation between the fixation probability and the long-term rate of adaptation

It is useful to compare these results with the deterministic modifier theory in SI Section 1, which predicts that modifiers will be favored if they increase the long-term mean fitness of the population. While this result was originally derived for non-adapting populations at mutation-selection balance, a natural extension of this idea to adapting populations might suggest that natural selection would favor modifiers that increase the long-term rate of adaptation (i.e. $v_m > v$).

Our results above show that this simple heuristic clearly breaks down for modifiers with direct costs or benefits (Fig. 3). This discrepancy is most dramatic for the "dead-end" modifiers in SI Section 4.1.6, which can be strongly favored by selection even though they lower the rate of adaptation to zero. Our analysis in SI Section 4.1.6 showed that the origin of this effect could be traced to the early fixation of the modifier lineage before its long-term costs are fully realized. These fixation events are completely neglected by the deterministic theory in SI Section 1: beneficial variants can grow to arbitrarily large frequencies in the short-term, but their long-term fitness gains will always override their initial costs or benefits. This highlights how finite population sizes can be critically important, even when the population size is very large (47).

To understand this relationship in more detail, we can combine our results in Eqs. (S58), (S71), and (S100) to derive an approximate formula connecting p_{fix}, v_m, and s_m. In the limit that $x_c \gg s_b$ and $x_{cm} \gg s_b'$, we find that the leading-order contributions satisfy the approximate scaling,

$$\log \tilde{p}_{\text{fix}}(\mu \to \mu_m; s_m) \approx \frac{x_{cm}^2}{2v}\left(\frac{v_m}{v} - 1\right) + \frac{x_{cm} s_m}{v}\,, \tag{S117}$$

to leading order in the logarithm of $\tilde{p}_{\text{fix}}$. This result shows that in the absence of direct costs or benefits ($s_m = 0$), the sign of selection in our simple model is directly related to the modifier's effect on the long-term rate of adaptation ($v_m - v$). This is reminiscent of the mean fitness principle in SI Section 1. Equation (S117) allows us to generalize this result to modifiers with direct costs or benefits ($s_m \neq 0$). Interestingly, we find that Eq. (S117) has the same form as the deterministic prediction in Eq. (S5) if we take $\overline{X}_m \approx s_m + v_m t$ and $\overline{X}_w \approx vt$, as expected, but impose a finite integration time $T_{\max} = x_{cm}/v$. This time limit roughly coincides with the fixation time of a successful mutation. This allows us to recover a version of the mean fitness principle in SI Section 1: within our simple clonal interference model, natural selection will favor modifiers that produce a higher mean fitness within a typical fixation time.

We note, however, that Eq. (S117) is only an approximate formula, which captures the leading-order scaling of the logarithm of p_{fix} (rather than p_{fix} itself). The sub-leading contributions to the logarithm are often important in practice, since they can be large in an absolute sense ($\gtrsim 1$) and must be exponentiated to obtain p_{fix}. Thus, while Eq. (S117) can provide a useful rule of thumb, we use the full expressions in SI Section 4.1.5 when calculating our theoretical predictions in Figs. 2 and 3 in the main text.

4.2 Quasi-sweep regime

4.2.1 Location of the interference threshold

As we consider larger values of s_b', we will eventually reach a regime where a single established driver mutation will be sufficient to drive a modifier lineage to fixation ($s_b' - x_{cm} \gg \sqrt{v}$). To solve for the shape of $w_m(x)$ in this regime, we first revisit the auxilliary condition for x_{cm} in Eq. (S53). When $s_b' - x_{cm} \gg \sqrt{v}$, this integral is dominated by fitnesses within $O(\sqrt{v})$ of $y = s_b'$, where the shoulder solution has already transitioned to the linear Haldane limit. Performing a Gaussian Laplace approximation around this maximum

yields a new auxiliary condition for x_{cm},

$$1 \approx \frac{\sqrt{2\pi} U_b' s_b'}{x_{cm}\sqrt{v}} e^{\frac{x_{cm}^2}{2v}}, \tag{S118}$$

which is valid when $s_b' - x_{cm} \gg \sqrt{v}$ and $x_{cm} \gg \sqrt{v}$. Solving for x_{cm}, we obtain the leading order solution,

$$x_{cm} = \sqrt{2v \log\left(\frac{v}{U_b' s_b'}\right)}, \tag{S119}$$

which is valid when

$$\frac{U_b' s_b'}{U_b s_b} \ll \frac{v}{U_b s_b} \sim \frac{s_b}{U_b} \frac{\log(N s_b)}{\log^2(s_b/U_b)}. \tag{S120}$$

This is a broad regime when $U_b \ll s_b$, but violations of this condition can become important when we consider continuous distributions of fitness effects below (SI Section 5). It will also be useful to derive an alternative expression for x_{cm} by substituting the auxilliary condition for x_c in Eq. (S57) into Eq. (S118), yielding

$$\frac{x_{cm}^2}{2v} - \frac{x_c s_b}{v} + \frac{s_b^2}{2v} + \log\left(\frac{U_b'}{U_b} \cdot \frac{s_b'}{s_b}\right) + \log\left(\frac{\sqrt{2\pi} s_b^2}{x_{cm}\sqrt{v}}\right) \approx 0. \tag{S121}$$

The leading order solution is given by

$$x_{cm} \approx \sqrt{(2x_c - s_b)s_b - 2v \log(U_b'/U_b)}, \tag{S122}$$

where we have included the potential contribution from $\log(U_b'/U_b)$, but assumed that the corresponding $\log(s_b'/s_b)$ term is subdominant. The condition that $s_b' - x_{cm} \gg \sqrt{v}$ will be valid provided that

$$\frac{s_b'}{s_b} \gtrsim \sqrt{\frac{2x_c - s_b}{s_b} - \frac{2v}{s_b^2} \log\left(U_b'/U_b\right)}, \tag{S123}$$

which complements the analogous condition for the mulitple mutations regime in Eq. (S59). Thus, for a pure selection strength modifier, the quasi-sweep regime will occur for $s_b' \gg \sqrt{(2x_c - s_b)s_b}$, while a pure mutation rate modifier requires the more stringent condition, $\log\left(U_b'/U_b\right) \gtrsim (x_c - s_b)s_b/v$. This suggests that relatively modest changes to the selection strength will require the analysis described below.

4.2.2 Extending the shoulder solution to lower fitness values

Having fixed the location of x_{cm}, we can repeat the procedure in SI Section 4.1.2 to compute the shape of $w_m(x)$ for fitness values below x_{cm}. If $x + s_b' \gtrsim s_b + O(\sqrt{v})$, then the dominant contribution to Eq. (S53) will be contained within the region of integration, so that we can substitute the auxilliary condition to obtain

$$w_m(x) \approx 2x_{cm} e^{\frac{x^2 - x_{cm}^2}{2v}}. \tag{S124}$$

This once again coincides with the shoulder solution in Eq. (S42) in the region where $x < x_{cm} - O(v/x_{cm})$. This derivation shows that the shoulder solution will continue to be valid for fitnesses as low as

$$x_{\min}(\mu_m) = O(\sqrt{v}), \tag{S125}$$

which is significantly smaller than both s'_b and x_{cm}.

For fitnesses that are less than $-O(\sqrt{v})$, the integral in Eq. (S52) will start to be dominated by the upper limit of integration at $x + s'_b$. Provided that $x + s'_b - O(v/(s'_b - x_{cm})) \gtrsim x_{cm} + O(v/x_{cm})$, then the dominant contribution to the integral will still come from the Haldane region of the shoulder solution. An exponential Laplace approximation then yields

$$w_m(x) \approx \frac{2U'_b(s'_b - |x|)}{|x|} \approx \frac{2x_{cm}\sqrt{v}}{\sqrt{2\pi}|x|}\left(1 - \frac{|x|}{s'_b}\right)e^{-\frac{x_{cm}^2}{2v}}, \tag{S126}$$

where we have substituted the auxilliary condition in Eq. (S118). To capture the behavior in the intermediate region around $x = \pm O(\sqrt{v})$, we can turn to the full Gaussian integral,

$$\begin{aligned} w_m(x) &\approx 2x_{cm}e^{\frac{x^2 - x_{cm}^2}{2v}}\left[\frac{U'_b s'_b\sqrt{2\pi v}e^{\frac{x_{cm}^2}{2v}}}{v x_{cm}} \cdot \Phi\left(\frac{x}{\sqrt{v}}\right) - \frac{U'_b e^{\frac{x_{cm}^2 - x^2}{2v}}}{x_{cm}}\right], \\ &\approx 2x_{cm}e^{\frac{x^2 - x_{cm}^2}{2v}}\left[\Phi\left(\frac{x}{\sqrt{v}}\right) - \frac{\sqrt{v}}{s'_b\sqrt{2\pi}}e^{-\frac{x^2}{2v}}\right], \end{aligned} \tag{S127}$$

where $\Phi(z) = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{z} e^{-u^2/2}\,du$ is the Gaussian cumulative distribution function. This reduces to Eqs. (S124) and (S126) in the appropriate limits, but captures the intermediate region where $x = \pm O(\sqrt{v})$.

For fitness below $x_{cm} - s'_b - O(v/x_{cm})$, the upper limit of the integral in Eq. (S52) will start to fall within the exponential region of the shoulder solution in Eq. (S42). However, since the latter smoothly maps on to the Gaussian integral in Eq. (S127), we can use this solution to extend $w_m(x)$ down to $x \approx x_{cm} - 2s_b$:

$$\begin{aligned} w_m(x) &\approx 2x_{cm}e^{\frac{x^2 - x_{cm}^2}{2v}}\frac{U'_b}{v}\int_{-\infty}^{x+s'_b} e^{\frac{ys'_b}{v} - \frac{s_b'^2}{2v}}\left[\Phi\left(\frac{y}{\sqrt{v}}\right) - \frac{\sqrt{v}}{s'_b\sqrt{2\pi}}e^{-\frac{y^2}{2v}}\right]dy, \\ &\approx 2x_{cm}e^{\frac{x^2 - x_{cm}^2}{2v}}\left(\frac{U'_b}{s'_b}\right)\left[e^{\frac{xs'_b}{v} + \frac{s_b'^2}{2v}} \cdot \Phi\left(\frac{x + s'_b}{\sqrt{v}}\right) - \frac{2v}{|x|\sqrt{2\pi v}}e^{-\frac{x^2}{2v}}\right]. \end{aligned} \tag{S128}$$

Iterating this procedure again yields

$$\begin{aligned} w_m(x) &\approx 2x_{cm}e^{\frac{x^2 - x_{cm}^2}{2v}}\left(\frac{U'_b}{s'_b}\right)^2\frac{s'_b}{v}\left[\int_{-\infty}^{x+s'_b} e^{\frac{2ys'_b}{v}}\Phi\left(\frac{y + s'_b}{\sqrt{v}}\right)dy - \int_{-\infty}^{x+s'_b}\frac{2v}{y\sqrt{2\pi v}}e^{-\frac{(y - s'_b)^2}{2v}}\right]dy, \\ &\approx 2x_{cm}e^{\frac{x^2 - x_{cm}^2}{2v}}\left(\frac{U'_b}{s'_b}\right)^2\left[\frac{1}{2}e^{\frac{2xs'_b}{v} + \frac{2s_b'^2}{v}}\Phi\left(\frac{x + 2s'_b}{\sqrt{v}}\right) - \frac{ve^{-\frac{x^2}{2v}}}{2|x|\sqrt{2\pi v}} - \frac{2s'_b ve^{-\frac{x^2}{2v}}}{|x + s_b||x|\sqrt{2\pi v}}\right], \end{aligned} \tag{S129}$$

which will be valid for $x \in (x_{cm} - 3s'_b, x_{cm} - 2s'_b)$. More generally, for $x \in (x_{cm} - (n+1)s'_b, x_{cm} - ns'_b)$, one can show that the solution for $w_m(x)$ is given by:

$$\begin{aligned} w_m(x) \approx 2x_{cm}e^{\frac{x^2 - x_{cm}^2}{2v}}\left(\frac{U'_b}{s'_b}\right)^n\Bigg[&\frac{1}{n!}e^{\frac{2nxs'_b}{2v} + \frac{n^2 s_b'^2}{2v}}\Phi\left(\frac{x + ns'_b}{\sqrt{v}}\right) \\ &- \frac{ve^{-\frac{x^2}{2v}}}{s'_b\sqrt{2\pi v}}\left(\sum_{m=1}^{n}\frac{1}{m!}\prod_{j=0}^{n-m}\frac{s'_b}{|x + js'_b|} + \prod_{j=0}^{n-1}\frac{s'_b}{|x + js'_b|}\right)\Bigg], \end{aligned} \tag{S130}$$

which simplifies to,

$$w_m(x) \approx 2x_{cm} e^{\frac{x^2 - x_{cm}^2}{2v}} \left(\frac{U_b'}{s_b'}\right)^n \left[\frac{1}{n!} e^{\frac{2nxs_b'}{2v} + \frac{n^2 s_b'^2}{2v}} \Phi\left(\frac{x + ns_b'}{\sqrt{v}}\right) - \frac{v e^{-\frac{x^2}{2v}}}{s_b'\sqrt{2\pi v}} \left(\frac{-\Gamma\left(-\frac{x}{s_b'} - n\right)\left(n + 1 + \frac{x}{s_b'}\right)}{\Gamma\left(1 - \frac{x}{s_b'}\right)} - \frac{s_b'}{n!(x + ns_b')}\right)\right], \tag{S131}$$

and approaches the asymptotic limit,

$$w_m(x) \approx 2x_{cm} e^{\frac{-x_{cm}^2}{2v}} \left(\frac{U_b'}{s_b'}\right)^n \frac{v}{s_b'\sqrt{2\pi v}} \frac{1}{\Gamma(1 + \frac{|x|}{s_b'})} \frac{x + (n+1)s_b'}{|x + ns_b'|} \approx \frac{2s_b' \left(\frac{U_b'}{s_b'}\right)^{n+1}}{\Gamma(1 + \frac{|x|}{s_b'})} \frac{x + (n+1)s_b'}{|x + ns_b'|} \tag{S132}$$

when $x \lesssim -ns_b' - O(\sqrt{v})$. Each additional "step" reduces the original shoulder by a factor of $(U_b'/ns_b')^n$. For fitness values that are many multiples of s_b' below x_{cm}, we can substitute $n \approx (x_{cm} - x)/s_b'$ to show that the leading order contribution to $w_m(x)$ scales as

$$\log w_m(x) \sim \text{const} + (x_{cm} - x)/s_b' \log\left(U_b'/(x_{cm} - x)\right), \tag{S133}$$

which obeys the required boundary condition that $w_m(x) \to 0$ as $x \to \infty$. Thus Eq. (S131) provides an asymptotic solution for $w_m(x)$ that is valid across the full range of relative fitness below x_{cm} when $s_b' - x_{cm} \gg \sqrt{v}$.

4.2.3 Fixation probabilities of modifiers

Using the solution for $w_m(x)$ in Eq. (S131), we can repeat our calculations from SI Sections 4.1.4 and 4.1.5 to predict the fixation probabilities of modifier mutations in the quasi-sweep regime.

Modifiers without direct costs or benefits. In the absence of direct costs and benefits, the primary contribution to the fixation probability in Eq. (S69) will come from the $n = 0$ region of Eq. (S131). This yields

$$\tilde{p}_{\text{fix}}(s_b, U_b \to s_b', U_b') \approx \frac{x_{cm} s_b'}{x_c s_b} e^{\frac{x_c^2 - x_{cm}^2}{2v}} \int_0^{\min\{x_c, x_{cm}\}} \frac{dx}{s_b} \approx \frac{x_{cm} \min\{x_c, x_{cm}\}}{x_c s_b} e^{\frac{x_c^2 - x_{cm}^2}{2v}}. \tag{S134}$$

Using Eqs. (S71) and (S118), we can rewrite this expression in the compact form,

$$\frac{2NU_b' \min\{x_c, x_{cm}\} s_b'}{v}. \tag{S135}$$

This expression has a simple heuristic explanation. A modifier lineage founded at relative fitness x will produce a total of $NU_b' \int_0^\infty e^{xt - \frac{vt^2}{2}} dt$ offspring before it is purged from the population. The modifier lineage will produce the bulk of these offspring within $\sim 1/\sqrt{v}$ generations of time $t = x/v$, when it will have reached size $n(t = x/v) = e^{\frac{x^2}{2v}}$ and be near the mean fitness of the population. These offspring will fix provided they survive genetic drift, which occurs with probability $p_{\text{est}} \approx 2s_b'$. Multiplying these terms together, we see that the advantage of arising on a more fit background and hitchhiking to exponentially larger frequency is exactly balanced by the exponentially smaller probability of arising on one of these privileged backgrounds

$f(x) \sim e^{-\frac{x^2}{2v}}$. Consequently, a successful modifier is equally likely to have arisen on a background with relative fitness in the range $O(\sqrt{v}) < x < \min\{x_{cm}, x_c\}$.

By comparing Eq. (S90), Eq. (S89), and Eq. (S135) and Eq. (S88), we can see that there is a limit to the degree in which changes to the selection strength of mutations can modulate the effect of linked changes to mutation rate ($\log(U_b'/U_b) < (x_c - s_b)s_b/v)$): $\alpha = \max\{(s_b'/s_b)^2, s_b/x_c\}$. When $s_b' - x_{cm} \gg \sqrt{v}$, a mutation rate modification influences only a single subsequent mutation compared to multiple mutations when $x_{cm} - s_b' \gg \sqrt{v}$. We also note that Eq. (S135) implies that large changes to the mutation rate ($\log(U_b'/U_b) < (x_c - s_b)s_b/v) < (x_c - s_b)s_b/v) + s_b^2/2v$) can modulate the effect of a selection-strength modifier, limiting its ability to compound.

Modifiers with direct fitness benefits. For a modifier with a direct fitness benefit ($s_m > 0$), the fixation probability in Eq. (S92) will once again depend on how the upper limit of integration ($x_c + s_m$) compares to the interference threshold x_{cm}. If $s_m \gtrsim x_{cm} - x_c + O(v/x_{cm})$, there will again be a contribution from the Haldane region of $w_m(x)$, so that

$$\begin{aligned}\tilde{p}_{\text{fix}}(\mu \to \mu_m; s_m) &\approx \frac{x_{cm}^2}{x_c s_b} e^{\frac{x_c^2 - x_{cm}^2}{2v} - \frac{s_m^2}{2v}} \int_0^{x_{cm}} e^{\frac{x s_m}{v}} \cdot \frac{dx}{x_{cm}} + \int_{x_{cm}}^{x_{cm}+s_m-(x_{cm}-x_c)} \frac{2Nx e^{-\frac{(x-s_m)^2}{2v}}}{\sqrt{2\pi v}}\, dx\,, \\ &\approx \frac{x_{cm}^2}{x_c s_b} e^{\frac{x_c^2 - x_{cm}^2 - s_m^2}{2v}} \left(\frac{e^{\frac{x_{cm} s_m}{v}} - 1}{\frac{x_{cm} s_m}{v}} \right) + 2N s_m \cdot \Phi\left(\frac{s_m - x_{cm}}{\sqrt{v}} \right)\,, \end{aligned} \tag{S136}$$

where $\Phi(z) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{z} e^{-u^2/2} du$ is the Gaussian cumulative distribution function.

In the opposite case, where $s_m \lesssim x_{cm} - x_c - O(v/x_{cm})$, the upper limit of integration in Eq. (S92) will fall below the interference threshold of $w_m(x)$. Repeating the calculation in Eq. (S136) in this case yields

$$\tilde{p}_{\text{fix}}(\mu \to \mu_m; s_m) \approx \frac{x_{cm}^2}{x_c s_b} e^{\frac{x_c^2 - x_{cm}^2 - s_m^2}{2v}} \frac{e^{\frac{(x_c + s_m) s_m}{v}} - 1}{\frac{x_{cm} s_m}{v}}\,. \tag{S137}$$

Modifiers with direct fitness costs. For a modifier with a beneficial effect on evolvability ($x_c > x_{cm}$) but a direct fitness cost ($s_m < 0$), the derivation will be similar to Eq. (S99) above. The dominant contribution to the fixation probability in Eq. (S92) will again come from the $n = k$ term in Eq. (S131) (as well as the first portion of the $k + 1$ term), where k is defined by

$$|s_m| = k s_b' + \Delta\,, \tag{S138}$$

for some $\Delta \in (0, s'_b)$. The fixation probability of the modifier can therefore be approximated as

$$\tilde{p}_{\text{fix}}(\mu \to \mu_m; s_m) \approx \tilde{p}_{\text{fix}}(\mu \to \mu_m; 0) \cdot \left(\frac{U'_b}{s'_b}\right)^k \left\{ \frac{e^{-\frac{ks'_b\Delta}{v} - \frac{\Delta^2}{2v}}}{k!} \int_{x_{cm}-(k+1)s'_b}^{x_{cm}-ks'_b-\zeta} e^{-\frac{x\Delta}{v}} \Phi\left(\frac{x+ks'_b}{\sqrt{v}}\right) \frac{dx}{x_{cm}} \right.$$
$$+ \frac{v}{s'_b\sqrt{2\pi v}} \int_{x_{cm}-(k+1)s'_b}^{x_{cm}-ks'_b-\zeta} e^{-\frac{(x-s_m)^2}{2v}} \cdot \frac{(x+(k+1)s'_b)}{s'_b} \frac{\Gamma\left(-\frac{x}{s'_b} - k\right)}{\Gamma\left(1 - \frac{x}{s'_b}\right)} \frac{dx}{x_{cm}}$$
$$+ \frac{v}{s'_b\sqrt{2\pi v}} \int_{x_{cm}-(k+1)s'_b}^{x_{cm}-ks'_b-\zeta} e^{-\frac{(x-s_m)^2}{2v}} \cdot \frac{1}{k!} \frac{s'_b}{(x+ks'_b)} \frac{dx}{x_{cm}}$$
$$\left. + \frac{e^{-\frac{ks'_b\Delta}{v} - \frac{\Delta^2}{2v} + \frac{(2k+1)s'^2_b}{2v} - \frac{x^2_{cm}}{2v}}}{(k+1)!} \frac{x_{cm}\sqrt{v}}{s'^2_b\sqrt{2\pi}} \int_{x_{cm}-(k+2)s'_b}^{x_{cm}-(k+1)s'_b} e^{\frac{x(s'_b-\Delta)}{v}} \Phi\left(\frac{x+(k+1)s'_b}{\sqrt{v}}\right) \frac{dx}{x_{cm}} \right\}, \tag{S139}$$

where $\zeta \equiv \max\{0, \Delta - (x_c - x_{cm})\}$. When $k = 0$ and $\Delta \lesssim O(\sqrt{v})$, the fixation probability is dominated by the portion of the first term where $\Phi(z) \approx 1$. This yields the asymptotic approximation

$$\tilde{p}_{\text{fix}}(\mu \to \mu_m; s_m) \approx \tilde{p}_{\text{fix}}(\mu, \to \mu_m; 0) \cdot \frac{v}{x_{cm}\Delta}\left(1 - e^{-\frac{x_{cm}\Delta}{v}}\right), \quad (|s_m| \lesssim \sqrt{v}) \tag{S140}$$

On the other hand, when $O(\sqrt{v}) \lesssim \Delta \lesssim s'_b - x_{cm} - O(\sqrt{v})$, the fixation probability will be dominated by the sum of the first three terms. These contributions will be peaked around a narrow interval near $x^* = -ks'_b - \Delta$, where the sum of the first three terms transitions to the asymptotic expression in Eq. (S132):

$$\tilde{p}_{\text{fix}}(\mu \to \mu_m; s_m) \approx \tilde{p}_{\text{fix}}(\mu \to \mu_m; 0) \left(\frac{U'_b}{s'_b}\right)^k \frac{v}{s'_b\sqrt{2\pi v}} \int_{x_{cm}-(k+1)s'_b}^{x_{cm}-ks'_b-\zeta} \frac{e^{-\frac{(x-s_m)^2}{2v}}}{\Gamma(1+\frac{|x|}{s'_b})} \frac{x+(k+1)s'_b}{|x+ks'_b|} \frac{dx}{x_{cm}},$$
$$\approx \tilde{p}_{\text{fix}}(\mu \to \mu_m; 0) \left(\frac{U'_b}{s'_b}\right)^k \frac{v}{s'_b x_{cm}} \frac{1}{\Gamma(k+1+\frac{\Delta}{s'_b})} \frac{s'_b - \Delta}{\Delta}. \tag{S141}$$

which is valid for $O(\sqrt{v}) \lesssim \Delta \lesssim s'_b - x_{cm} - O(\sqrt{v})$. Finally, in the last region where when $s'_b - x_{cm} + O(\sqrt{v}) \lesssim \Delta \lesssim s'_b - O(s'_b)$, the fixation probability will be dominated by the sum of all four terms, but the first three

will be peaked near the lower boundary. This yields

$$\tilde{p}_{\text{fix}}(\mu \to \mu_m; s_m) \approx \tilde{p}_{\text{fix}}(\mu \to \mu_m; 0) \left(\frac{U_b'}{s_b'}\right)^k \left\{ \frac{v}{s_b'\sqrt{2\pi v}} \int_{x_{cm}-(k+1)s_b'}^{x_{cm}-ks_b'-\zeta} \frac{e^{-\frac{(x-s_m)^2}{2v}}}{\Gamma(1+\frac{|x|}{s_b'})} \frac{x+(k+1)s_b'}{|x+ks_b'|} \frac{dx}{x_{cm}} \right.$$
$$\left. + \frac{e^{-\frac{ks_b'\Delta}{v}-\frac{\Delta^2}{2v}+\frac{(2k+1)s_b'^2}{2v}-\frac{x_{cm}^2}{2v}}}{(k+1)!} \frac{x_{cm}\sqrt{v}}{s_b'^2\sqrt{2\pi}} \int_{x_{cm}-(k+2)s_b'}^{x_{cm}-(k+1)s_b'} e^{\frac{x(s_b'-\Delta)}{v}} \Phi\left(\frac{x+(k+1)s_b'}{\sqrt{v}}\right) \frac{dx}{x_{cm}} \right\}$$
$$\approx p_{\text{fix}}(\mu \to \mu_m; 0) \left(\frac{U_b'}{s_b'}\right)^k \left\{ \frac{\frac{v}{s_b'\sqrt{2\pi v}} v e^{-\frac{(\Delta-s_b'+x_{cm})^2}{2v}}}{(\Delta - s_b' + x_{cm})(s_b' - x_{cm})\Gamma(k+2-\frac{x_{cm}}{s_b'})} \right.$$
$$\left. + \frac{\frac{v}{s_b'\sqrt{2\pi v}} v e^{-\frac{(\Delta-s_b'+x_{cm})^2}{2v}}}{s_b'(s_b'-\Delta)\Gamma(k+2)} \right\} \tag{S142}$$

which is valid for $s_b' - x_{cm} + O(\sqrt{v}) \lesssim \Delta \lesssim s_b' - O(s_b')$. The leading order solution for the fixation probability can therefore be summarized as

$$\log \tilde{p}_{\text{fix}}(\mu \to \mu_m; s_m) \approx \begin{cases} \frac{x_c^2}{2v} - \frac{x_{cm}^2}{2v} + \frac{x_{cm}s_m}{v} & \text{if } |s_m| \ll x_c, s_m > 0, \\ \frac{x_c^2}{2v} - \frac{x_{cm}^2}{2v} - \log\left(1 - \frac{|s_m|}{s_b'}\right) + \log\left(\frac{v}{x_{cm}|s_m|}\right) & \text{if } \sqrt{v} \ll |s_m| \ll s_b' - x_{cm}, s_m < 0, \\ \frac{x_c^2}{2v} - \frac{x_{cm}^2}{2v} - \frac{|s_m|}{s_b'}\log\left(\frac{U_b'}{|s_m|}\right) & \text{if } |s_m| \gg s_b', s_m < 0. \end{cases} \tag{S143}$$

5 Extension to continuous distributions of fitness effects

So far, we have focused on a simplified model of the mutation spectrum, where all new mutations confer the same characteristic fitness benefit. In reality, beneficial mutations can have a range of different fitness benefits, so a general evolvability modifier will involve an arbitrary perturbation to the distribution of fitness effects,

$$\delta\mu(s) = \mu_m(s) - \mu(s) \,, \tag{S144}$$

which could include the addition or subtraction of mutations with a range of different fitness benefits. This forces us to return to the full solution in Eq. (S43) of SI Section 3,

$$w_m(x) = 2x_{cm} e^{\frac{x^2}{2v} - \frac{x_{cm}^2}{2v}}$$
$$\times \int_0^\infty ds \int_{-\infty}^{x+s} \frac{\mu(s) + \delta\mu(s)}{v} \left(\frac{e^{\frac{ys}{v} - \frac{s^2}{2v}} \theta(y - x_{\min})}{1 + \frac{x_{cm}}{y} e^{\frac{y^2}{2v} - \frac{x_{cm}^2}{2v}}} + \frac{w_m(y)\theta(x_{\min} - y)}{2x_{cm} e^{\frac{(y-s)^2}{2v} - \frac{x_{cm}^2}{2v}}} \right) dy \,, \tag{S145}$$

along with the corresponding condition for x_{cm},

$$1 \approx \int_0^\infty ds \int_{-\infty}^{x_{cm}+s} \frac{\mu(s) + \delta\mu(s)}{v} \left(\frac{e^{\frac{ys}{v} - \frac{s^2}{2v}} \theta(y - x_{\min})}{1 + \frac{x_{cm}}{y} e^{\frac{y^2}{2v} - \frac{x_{cm}^2}{2v}}} + \frac{w_m(y)\theta(x_{\min} - y)}{2x_{cm} e^{\frac{(y-s)^2}{2v} - \frac{x_{cm}^2}{2v}}} \right) dy \,, \tag{S146}$$

both of which now include an integral over the fitness benefit of the next mutation (s) in addition to its landing fitness (y). The solutions to this equation are more complicated than the single-effect model in SI Section 4, since the y values that contribute the most to each integral will depend on the corresponding value of s (and vice versa).

However, previous work (39, 40) has shown that for a broad class of distributions with exponentially bounded tails, the integrals over s are dominated by a characteristic fitness effect s'_b when $x \geq x_c - s'_b$. This is a crucial simplification, as it implies that for this range of fitnesses, the shape of $w_m(x)$ and the location of x_{cm} can be well approximated by the single-s solutions in SI Section 4 for some appropriately chosen values of s'_b and U'_b (39, 40, 47, 48). These effective parameters will depend on the underlying values of $\mu_m(s)$, x_{cm}, and v, and must be determined self-consistently using the procedures described below.

In this section, we ask whether this same idea can be extended to the full range of x values in Eq. (S145), which is necessary for predicting the modifier fixation probability, $p_{\text{fix}}(\mu \to \mu_m; s_m)$. We will see that the single-s approximation holds for certain classes of modifiers, but not others, and we will develop methods for approximating $p_{\text{fix}}(\mu \to \mu_m; s_m)$ in both cases.

To carry out this analysis, we note that the integral formulation in Eqs. (S145) and (S146) naturally decomposes the fixation probability into contributions from different fitness benefits. By splitting $\mu_m(s)$ into its components, $\mu_m(s) = \mu(s) + \delta\mu(s)$, it also allows us to infer the relative contributions of $\mu(s)$ and $\delta\mu(s)$. We can make this connection even more explicit by rewriting Eq. (S146) in the form,

$$1 = \frac{\sqrt{2\pi v}}{2x_{cm} v e^{-\frac{x_{cm}^2}{2v}}} \int \mu(s) p_{\text{fix}}(s|x_{cm}, v)\, ds + \frac{\sqrt{2\pi v}}{2x_{cm} v e^{-\frac{x_{cm}^2}{2v}}} \int \delta\mu(s) p_{\text{fix}}(s|x_{cm}, v)\, ds\,, \tag{S147}$$

where

$$p_{\text{fix}}(s|x_{cm}, v) = \int_{-\infty}^{x_{cm}} \frac{e^{-\frac{x^2}{2v}}}{\sqrt{2\pi v}} w_m(x+s) \tag{S148}$$

is the fixation probability of a first-order mutation in the "dual" population defined by $v(\mu_m, N^*) = v(\mu, N)$ (see Eq. S37 in SI Section 3).

The decomposition in Eq. (S147) allows us to distinguish between two broad regimes depending on the relative contributions of $\mu(s)$ and $\delta\mu(s)$. When the fixation probability of the modifier is dominated by mutations from $\mu(s)$, then s'_b and x_{cm} will remain close to their wildtype values. This allows us to calculate x_{cm} and $p_{\text{fix}}(\mu \to \mu_m; s_m)$ perturbatively, by treating the $\delta\mu(s)$ term as a small correction. We will refer to this case as the *perturbative regime*, since it will only apply for modifiers with weak-to-moderate fixation probabilities. For stronger second-order selection pressures, the fixation probability of the modifier will be dominated by mutations from the $\delta\mu(s)$ portion of $\mu_m(s)$. In this *modifier-dominated regime*, we will need to turn to the full solution of $w_m(x)$, with $\mu_m(s) \approx \delta\mu(s)$. We will discuss each of these cases in turn, after quickly reviewing the wildtype dynamics when $\delta\mu(s) = 0$.

5.1 Location of the interference threshold for the wildtype population

When $\delta\mu(s) = 0$, Eq. (S146) determines the location of interference threshold for the wildtype population,

$$1 = \int_0^\infty ds \int_{-\infty}^{x_c+s} \frac{\mu(s)}{v} \left(\frac{e^{\frac{ys}{v}}}{1 + \frac{x_c}{y} e^{\frac{y^2}{2v} - \frac{x_c^2}{2v}}} + \frac{w_m(y)\theta(x_{\min} - y)}{2x_c e^{\frac{(y-s)^2}{2v} - \frac{x_c^2}{2v}}} \right) dy\,. \tag{S149}$$

The solution to this equation was previously described by Ref. (39). We reproduce it here for completeness, since we will build on this result in the following sections.

Ref. (39) showed that in our parameter regime of interest, the s integral in Eq. (S149) will be strongly peaked at a characteristic value, s_b, which will be determined self-consistently below. This observation allows us to evaluate the inner integral over y for s close to s_b. When $x_c - s_b \gg \sqrt{v}$, the contributions from $y < x_c$ will be dominated by y within $O(v/s)$ of x_c, while the contributions from $y > x_c$ will be dominated by fitnesses within $O(v/(x_c - s))$ of x_c. Equation (S149) then reduces to a single integral over s,

$$1 \approx \int_0^\infty \mu(s) e^{\frac{x_c s}{v} - \frac{s^2}{2v}} \left[\frac{1}{s} + \frac{1}{x_c - s}\right] ds \,, \tag{S150}$$

where the rapidly increasing exponential term is eventually counteracted by the rapidly decreasing $\mu(s)$ term. When $x_c - s_b \gg \sqrt{v}$, the competition between these two factors produces a sharp peak around $s \approx s_b \pm \Delta s_b$, where s_b is the value of s that maximizes the rapidly varying portion of the integrand,

$$x_c - s_b + v \frac{\partial \log \mu(s_b)}{\partial s} = 0 \,, \tag{S151}$$

and Δs_b is the characteristic width,

$$\Delta s_b = \left[\frac{1}{v} - \frac{\partial^2 \log \mu(s_b)}{\partial s^2}\right]^{-1/2} . \tag{S152}$$

When $s_b \gg \Delta s_b$, we can evaluate Eq. (S150) with a Gaussian Laplace approximation to obtain

$$1 = \frac{U_b}{s_b} \left[1 - \frac{s_b}{x_c}\right]^{-1} e^{\frac{x_c s_b}{v} - \frac{s_b^2}{2v}} \,, \tag{S153}$$

where the effective mutation rate,

$$U_b = \sqrt{2\pi \Delta s_b} \mu(s_b) \,, \tag{S154}$$

denotes the total rate of producing "driver" mutations within Δs_b of s_b. This matches the corresponding expression for the single-effect model in SI Section 4 if the effective selection strength and mutation rate are defined by Eqs. (S151) and (S154). Unlike SI Section 4, these effective parameters will now depend on the underlying values of $\mu(s)$ and N, and can shift if one or the other is varied (39, 40, 47, 48). This behavior will be crucial for understanding the fates of modifier mutations below.

Application to stretched exponential distributions. Following previous work (39, 40, 47), we can gain some intuition for these expressions by considering the class of stretched exponential distributions,

$$\mu(s) = \frac{U_0}{s_0} \cdot \frac{e^{-(s/s_0)^\beta}}{\Gamma(1 + \beta^{-1})} \,, \tag{S155}$$

as a function of the tail parameter $\beta \geq 1$. Ref. (39) showed that the effective parameters for the exponential case ($\beta = 1$) are given by

$$s_b = x_c - \frac{v}{s_0} \approx x_c \,, \qquad \Delta s_b = \sqrt{v} \,, \tag{S156a}$$

leading to the approximate scaling,

$$x_c \sim s_0 \log(N U_0) \,, \qquad v \sim \frac{s_0^2 \log^2(N U_0)}{2 \log(s_0/U_0)} \,. \tag{S156b}$$

Similarly, the solution for the $\beta \gg 1$ case is given by

$$s_b \approx \left(\frac{x_c s_0^\beta}{\beta v}\right)^{\frac{1}{\beta-1}} \left[1 - \frac{1}{1+\beta(\beta-1)\frac{v}{s_0^2}\left(\frac{x_c}{\beta v}\right)^{\frac{\beta-2}{\beta-1}}}\right] \approx \left(\frac{x_c s_0^\beta}{\beta v}\right)^{\frac{1}{\beta-1}}, \tag{S157a}$$

$$\Delta s_b = \left[\frac{1}{v} + \frac{\beta(\beta-1)}{s_0^2}\left(\frac{s_b'}{s_0}\right)^{\beta-2}\right]^{-1/2} \approx \frac{s_0}{\sqrt{\beta(\beta-1)\left(\frac{x_c s_0}{\beta v}\right)^{\frac{\beta-2}{\beta-1}}}}, \tag{S157b}$$

with

$$x_c \sim \frac{2s_0 \log(Ns_0)}{\log(\sqrt{\beta}s_0/U_0)}, \qquad v \sim \frac{2s_0^2 \log(Ns_0)}{\log^2(\sqrt{\beta}s_0/U_0)}. \tag{S157c}$$

In our analysis below, we will find that the $\beta = 2$ case (corresponding to a half-Gaussian distribution) will serve as an important boundary case, so it will be useful to derive an explicit expression for this case as well. Equations (S151) and (S152) imply that

$$s_b = \frac{x_c}{1+2\left(\frac{v}{s_0^2}\right)}, \qquad \Delta s_b = \sqrt{v\left[\frac{1}{1+2\left(\frac{v}{s_0'^2}\right)}\right]}, \tag{S158a}$$

which yields the asymptotic approximations

$$v \sim \frac{s_0^2 \log(NU_0)}{2\log(s_0/U_0)}, \qquad x_c \sim \sqrt{\frac{\log(Ns_0)\log(NU_0)}{\log(s_0/U_0)}}. \tag{S158b}$$

When $v/s_0^2 \ll 1$, this solution behaves like the exponential case above (with $s_b \approx x_c$). This regime will be valid provided that

$$1 \ll \log(NU_0) \ll \log(s_0/U_0). \tag{S159}$$

On the other hand, when $v/s_0^2 \gg 1$, the solution will behave more like the $\beta \gg 1$ case ($s_b \ll x_c$). This regime will be valid provided that

$$1 \ll \log(s_0/U_0) \ll \log(Ns_b) \ll \log^2(s_0/U_0). \tag{S160}$$

which mirrors the conditions of validity for the single-effect model in SI Section 4.

5.2 Perturbative regime

5.2.1 Location of the interference threshold

In the perturbative regime, we assume that the dominant contribution to Eq. (S146) still comes from the $\mu(s)$ term. This suggests that x_{cm} will remain close to the wildtype interference threshold,

$$x_{cm} = x_c + \delta x_c, \tag{S161}$$

where δx_c is a small correction whose magnitude will be determined self-consistently below.

If $\delta x_c \ll v/s_b$, then the integral over s will remain strongly peaked around the same value as the wildtype integral in Eq. (S150). This implies that

$$s_b(N^*) \approx s_b(N)\,, \quad U_b(N^*) = U_b(N)\,, \tag{S162}$$

where s_b and U_b are defined as in Eqs. (S151) and (S154), and that

$$\int \mu(s) e^{\frac{(x_c+\delta x_c)s}{v} - \frac{s^2}{2v}}\, ds \approx U_b \tilde{p}_{\text{fix}}(s_b|x_c)\left(1 + \frac{s_b \delta x_c}{v}\right) = U_b \tilde{p}_{\text{fix}}(s_b|x_c + \delta x_c)\,. \tag{S163}$$

From the self-consistency condition for N in Eq. (S71), we also have

$$\frac{2N^*(x_c+\delta x_c)e^{-\frac{(x_c+\delta x_c)^2}{2v}}}{\sqrt{2\pi v}} = \frac{1}{s_b} = \frac{2N x_c e^{-\frac{x_c^2}{2v}}}{\sqrt{2\pi v}}\,, \tag{S164}$$

and

$$\frac{N^*}{N} \approx e^{\frac{x_c \cdot \delta x_c}{v}}\,. \tag{S165}$$

Substituting these results into Eq. (S147), we obtain

$$1 = \frac{s_b}{v} \cdot U_b \tilde{p}_{\text{fix}}(s|x_c)\left(1 + \frac{s_b \delta x_c}{v}\right)\left[1 + I(\delta\mu, \delta x_c)\right]\,, \tag{S166}$$

where

$$I(\delta\mu, \delta x_c) \equiv \int \frac{\delta\mu(s)}{U_b} e^{\log \tilde{p}_{\text{fix}}(s|x_c+\delta x_c) - \log \tilde{p}_{\text{fix}}(s_b|x_c+\delta x_c)}\, ds \tag{S167}$$

is the integral from Eq. (8) in the main text. Equation (S166) shows that $I(\delta\mu, \delta x_c)$ can be interpreted as the relative contribution of mutations from $\delta\mu(s)$ vs $\mu(s)$ when $\delta x_c \ll v/s_b$. Substituting for the wildtype x_c condition $[U_b \tilde{p}_{\text{fix}}(s_b) = v/s_b]$, we can rearrange Eq. (S166) to obtain

$$\delta x_c = -\frac{v}{s_b} I(\delta\mu, \delta x_c)\,, \tag{S168}$$

which is valid when $\delta x_c \ll v/s_b$. This shows that $\delta x_c \ll v/s_b$ is equivalent to the condition that $I(\delta\mu, \delta x_c) \ll 1$. If $\delta x_c \ll v/s$ for the fitness effects that dominate Eq. (S167), then we can perform the same Taylor expansion obtain an analytical expression for δx_c,

$$\delta x_c = -\frac{v}{s_b} I(\delta\mu, 0) = -\frac{v}{s_b} \int \frac{\delta\mu(s)}{U_b} e^{\log \tilde{p}_{\text{fix}}(s|x_c) - \log \tilde{p}_{\text{fix}}(s_b|x_c)}\, ds\,, \tag{S169}$$

which can be evaluated in terms of the wildtype interference threshold x_c. For stronger fitness benefits, δx_c must be obtained by solving the implicit relation in Eq. (S168).

5.2.2 Fixation probabilities of modifiers

When the above assumptions hold, the fixation probability of the modifier can be obtained from our previous calculations in Eqs. (S82) and (S92). Since $\delta x_c \ll v/s_b \ll \sqrt{v} \ll x_c$, the fixation probability of a modifier without direct costs or benefits will be dominated by exceptionally fit genetic backgrounds ($x_c - s_b \lesssim x \lesssim x_c$), so that Eq. (S82) reduces to,

$$\tilde{p}_{\text{fix}}(\delta\mu) \approx \frac{x_{cm}}{x_c} e^{\frac{x_c^2}{2v} - \frac{x_{cm}^2}{2v}} \approx \exp\left[-\frac{x_c \delta x_c}{v}\right] . \tag{S170}$$

Substituting our expression for δx_c in Eq. (S168) then yields

$$\tilde{p}_{\text{fix}}(\delta\mu) \approx \exp\left[\frac{x_c}{s_b} \cdot I(\delta\mu, \delta x_c)\right] . \tag{S171}$$

This shows that the conditions of validity for the perturbative regime,

$$\delta x_c \ll v/s_b \iff I(\delta\mu, \delta x_c) \ll 1 , \tag{S172}$$

are equivalent to the assumption that

$$|\log \tilde{p}_{\text{fix}}(\delta\mu)| \ll x_c/s_b . \tag{S173}$$

Since x_c can be asymptotically larger than s_b, this implies that a small change to the distribution of fitness effects can be strongly selected ($|\log \tilde{p}_{\text{fix}}(\delta\mu)| \gg 1$) even when it leads to a negligible change in s_b and U_b. The perturbative regime is therefore an example of a modifier mutation that cannot be reduced to the single-effect model in SI Section 4.

It is straightforward extend this calculation to consider modifiers that also include direct costs or benefits. Since the second-order selection pressures are bounded by Eq. (S173), most of the interesting behavior will occur when $|s_m| \sim O(x_c/s_b) \ll s_b$. The fixation probability of these modifiers will continue to be dominated by exceptionally fit genetic backgrounds ($x_c - s_b \lesssim x \lesssim x_c$), so that Eq. (S92) reduces to

$$\log \tilde{p}_{\text{fix}}(\delta\mu, s_m) \approx \log \tilde{p}_{\text{fix}}(\delta\mu) + \log \tilde{p}_{\text{fix}}(s_m) , \tag{S174}$$

where $\tilde{p}_{\text{fix}}(s_m)$ is the fixation probability of a first-order mutation from SI Section 4.1.3.

5.3 Modifier-dominated regime (multiple mutations)

For stronger second-order selection pressures ($|\log \tilde{p}_{\text{fix}}(\delta\mu)| \gtrsim x_c/s_b$), the fixation probability in Eq. (S145) will start to depend more sensitively on the mutations that are added (or removed) by $\delta\mu(s)$. For a strongly beneficial modifier, the fixation probability will be dominated by the new mutations that are added by $\delta\mu(s)$. In contrast, a strongly deleterious modifier will be dominated by the mutations that remain once $\delta\mu(s)$ is subtracted from $\mu(s)$. In both cases, we will need to develop a general solution for $w_m(x)$ for arbitrary distributions $\mu_m(s) \neq \mu(s)$.

Following our analysis in SI Section 4, it will be useful to distinguish between two characteristic regimes, depending on whether the dominant fitness benefits are smaller or larger than x_{cm}. We will continue to refer to these as the *multiple mutations* and *quasi-sweeps* regimes, respectively, since they will qualitatively resemble their analogues in SI Section 4. We will start by analyzing the multiple mutations regime, while the quasi-sweeps case will be considered in SI Section 5.4 below.

5.3.1 Location of the interference threshold

When the relevant fitness benefits in $\mu_m(s)$ are small compared to x_{cm}, we can repeat the calculation in SI Section 5.1 to obtain an analogous condition for x_{cm},

$$1 \approx \int_0^\infty \mu_m(s) e^{\frac{x_{cm}s}{v} - \frac{s^2}{2v}} \left[\frac{1}{s} + \frac{1}{x_{cm} - s}\right] ds \,. \tag{S175}$$

This integral will be peaked around a new value s_b' that satisfies

$$x_{cm} - s_b' + v \frac{\partial \log \mu_m(s_b')}{\partial s} = 0 \,, \tag{S176}$$

along with a new characteristic width,

$$\Delta s_b' = \left[\frac{1}{v} - \frac{\partial^2 \log \mu_m(s_b')}{\partial s^2}\right]^{-1/2} \,, \tag{S177}$$

which will generally differ from the wildtype values of s_b and Δs_b above. When $s_b' \gg \Delta s_b'$ an analogous Gaussian Laplace approximation yields a simple expression for x_{cm},

$$1 = \frac{U_b'}{s_b'} \left[1 - \frac{s_b'}{x_{cm}}\right]^{-1} e^{\frac{x_{cm}s_b'}{v} - \frac{s_b'^2}{2v}} \,, \tag{S178}$$

where we have defined the effective mutation rate

$$U_b' = \sqrt{2\pi \Delta s_b'^2} \mu_m(s_b') \,. \tag{S179}$$

Together, Eqs. (S176-S178) allow us to solve for s_b', U_b', and x_{cm} as a function of $\mu_m(s)$ and v. These expressions will be self-consistently valid when $x_{cm} - s_b' \gg \sqrt{v}$.

5.3.2 Extending the shoulder solution to lower fitness values.

With the location of x_{cm} fixed by Eq. (S178), we can repeat our calculation in SI Section 4.1.2 to extend the shoulder solution for $w_m(x)$ to progressively lower fitness values. Provided that $x + s_b' \gtrsim x_{cm} + O(\Delta s_b')$, the dominant contribution to Eq. (S145) will also be contained within the region of integration. This demonstrates that the shoulder solution will continue to be valid for fitnesses as low as

$$x_{\min}(\mu_m) = x_{cm} - s_b' + \Delta s_b' \,. \tag{S180}$$

In particular, it shows that for $x > x_{\min}(\mu_m)$, the shape of $w_m(x)$ can be approximated by a single-effect model with s_b' and U_b' defined by Eqs. (S176) and (S179) above.

Hopping vs leapfrogging. When $x \lesssim x_{\min}(\mu_m)$, the upper limit in the landing fitness integral in Eq. (S145) will again become important. However, the crucial difference from our previous analysis in SI Section 4.1 is that this upper limit is not a fixed parameter, but can vary as a function of s. In other words, it is theoretically possible for the mutant offspring to land at *any* relative fitness value if it acquires a sufficiently large beneficial mutation from $\mu_m(s)$. This is a new effect that is only present when we allow for a continuous distribution of fitness effects. This will lead to two distinct classes of solutions for $w_m(x)$ depending on the tail of $\mu_m(s)$.

If the tail of $\mu_m(s)$ decays sufficiently rapidly, then a modifier lineage that arises below $x_{cm} - s'_b$ will typically fix by "hopping" to higher relative fitnesses through a sequence of smaller mutations, similar to the single-effect models in SI Section 4. However, if the tail of $\mu_m(s)$ decays more slowly, then a modifier arising below $x_{cm} - s'_b$ will typically fix by generating a *single* large driver mutation that bypasses these intermediate fitness values, and lands at or above interference threshold at x_{cm}. This "leapfrogging" regime is potentially empirically relevant, since it will apply when $\mu_m(s)$ decays like an exponential distribution.

By splitting the integral in Eq. (S145) into contributions from above and below the interference threshold, we can obtain a corresponding expression for the fixation probability,

$$w_m(x) \approx 2x_{cm} e^{\frac{x^2 - x_{cm}^2}{2v}} \underbrace{\int_{x_{cm}-x}^{\infty} \frac{\mu_m(s) 2(x+s)}{2x_{cm} e^{\frac{x^2-x_{cm}^2}{2v}}} \frac{ds}{|x|}}_{\text{landing above } x_{cm}} + 2x_{cm} e^{\frac{x^2-x_{cm}^2}{2v}} \underbrace{\int_{x_{\min}-x}^{x_{cm}-x} \mu_m(s) e^{\frac{xs}{v} + \frac{s^2}{2v}} \frac{ds}{s}}_{\text{landing below } x_{cm}}, \tag{S181}$$

which will be valid for $x \gtrsim x_{cm} - 2s'_b + O(v/s'_b + \Delta s_b)$. It will be helpful to evaluate these expressions by defining a new variable,

$$\Delta x \equiv x_{cm} - s'_b - x\,, \tag{S182}$$

such that $\Delta x \in (0, s'_b)$.

When $\Delta x \gtrsim O(\Delta s_b)$, then the lower limit of the integral in the first term in Eq. (S181) will lie in the high fitness tail of $\mu_m(s)$. An exponential Laplace approximation around $s = x_{cm} - x = s'_b + \Delta x$ then yields

$$2x_{cm} e^{\frac{x^2-x_{cm}^2}{2v}} \underbrace{\int_{x_{cm}-x}^{\infty} \frac{\mu_m(s) 2(x+s)}{2x_{cm} e^{\frac{x^2-x_{cm}^2}{2v}}} \frac{ds}{|x|}}_{\text{landing above } x_{cm}} \approx \frac{2x_{cm}}{|x|} \frac{\mu_m(s'_b + \Delta x)}{-\partial_s \log \mu_m(s'_b + \Delta x)}\,. \tag{S183}$$

The integrand in the second term in Eq. (S181) has a critical point at $s = s^*(\Delta x)$, which satisfies,

$$x_{cm} - s'_b - \Delta x + s^* + v \frac{\partial \log \mu_m(s^*)}{\partial s} = 0\,. \tag{S184}$$

This critical point will correspond to a local maximum if the second derivative is negative,

$$1 + v \frac{\partial^2 \log \mu_m(s^*)}{\partial s^2} < 0\,. \tag{S185}$$

The value of this integral will therefore crucially depend on the tail of $\mu_m(s)$.

If the curvature of $\mu_m(s)$ is sufficiently small $[\partial_s^2 \log \mu_m(s^*) > -1/v]$, then s^* will correspond to a local minimum of the integrand, and the integral will therefore be dominated by the upper limit of integration. An exponential Laplace approximation then yields,

$$w_m(x) \approx \frac{2x_{cm}}{|x|} \cdot \frac{\mu_m(s'_b + \Delta x)}{-\partial_s \log \mu_m(s'_b + \Delta x)} + \frac{2x_{cm}}{s'_b + \Delta x} \cdot \frac{\mu_m(s'_b + \Delta x)}{\frac{x_{cm}}{v} + \partial_s \log \mu_m(s'_b + \Delta x)}\,. \tag{S186}$$

Intuitively speaking, this says that the fixation probability is dominated by mutations of size $s \approx s'_b + \Delta x$ that result in a landing fitness very close to the interference threshold x_{cm}. Since Δx can be as large as s'_b (which is $\gg \Delta s'_b$), this constitutes an example of the "leapfrogging" behavior above, since it will involve a jump much larger than s'_b. From the form of the condition in Eq. (S185), we can see that this leapfrogging behavior

will always occur for the exponential distribution in Eq. (S155), as well as the Gaussian case ($\beta = 2$) when $v/s_0^2 \ll 1$. In these cases, we the single-s equivalence will break down for initial fitness below $x_{cm} - s'_b$, and we must use Eq. (S186) instead.

On the other hand, if the curvature of $\mu_m(s)$ is large and negative $[-\partial_s^2 \log \mu_m(s^*) \gg 1/v]$, then s^* will be a local maximum, and a Gaussian Laplace approximation yields

$$\int_{x_{\min}-x}^{x_{cm}-x} \mu_m(s) e^{\frac{xs}{v}+\frac{s^2}{2v}} \frac{ds}{s} \approx \sqrt{\frac{2\pi}{-\partial_s^2 \log \mu_m(s^*)}} \frac{\mu_m(s^*)}{s^*} e^{\frac{xs^*}{v}-\frac{s^{*2}}{2v}} . \tag{S187}$$

When $-\partial_s^2 \log \mu_m(s^*) \gg 1/v$, we can also solve Eq. (S184) perturbatively to show that $s^*(\Delta x)$ is close to s'_b:

$$s^* \approx s'_b \left(1 + \frac{1 - \frac{\Delta x}{s'_b}}{-v\partial_s^2 \log \mu_m(s'_b)}\right) \approx s'_b , \tag{S188}$$

so that Eq. (S181) reduces to

$$w_m(x) \approx \underbrace{\frac{2x_{cm}}{|x|} \cdot \frac{\mu_m(x_{cm}-x)}{-\partial_s \log \mu_m(s'_b + \Delta x)}}_{\substack{\text{landing above } x_{cm} \\ \text{“leapfrogging”}}} + \underbrace{2x_{cm} e^{\frac{x^2-x_{cm}^2}{2v}+\frac{xs'_b}{v}+\frac{s'^2_b}{2v}} \frac{\sqrt{2\pi\Delta s'^2_b}\mu_m(s'_b)}{s'_b}}_{\substack{\text{landing below } x_{cm} \\ \text{“hopping”}}} . \tag{S189}$$

The balance between these two terms will determine whether the typical successful lineage “hops” or “leapfrogs” to fixation. Comparing the magnitudes of these terms, we find that the hopping term in Eq. (S189) will dominate if

$$\log\left[\frac{\mu_m(s'_b)}{\mu_m(s'_b + \Delta x)}\right] - \frac{x_{cm}\Delta x}{v} + \frac{\Delta x^2}{2v} \gtrsim O(1) . \tag{S190}$$

This condition is more stringent than the one defining s'_b in Eq. (S175), which has an additional $s'_b \Delta x/v$ term on the left hand side. This means that even if mutations above s'_b are negligible for $x \gtrsim x_{cm} - s'_b$, they may become relevant for larger direct costs. When $\beta \geq 2$, the decay of $\mu_m(s)$ will still be bounded by its Gaussian approximation,

$$-\log \mu_m(s'_b + \Delta x) \geq -\log \mu_m(s'_b) - \frac{(s'_b - x_{cm})\Delta x}{v} + \frac{\Delta x^2}{2}\left(\frac{1}{\Delta s'^2_b} - \frac{1}{v}\right) , \tag{S191}$$

so Eq. (S190) reduces to

$$-\frac{s'_b \Delta x}{v} + \frac{\Delta x^2}{2\Delta s'^2_b} \gtrsim O(1) . \tag{S192}$$

The leapfrogging term will therefore be negligible for

$$\Delta x \gtrsim O(s'_b \cdot \Delta s'^2_b / v) , \tag{S193}$$

which will constitute the majority of the interval $(0, s'_b)$ when $v \gg \Delta s'^2_b$ — this is the same condition used to evaluate the hopping term above. For more general choices of $\mu_m(s)$, the Gaussian approximation may no

longer bound the value of $\mu_m(s_b' + \Delta x)$, and full condition in Eq. (S190) must be used instead. When these conditions are met, the fixation probability in Eq. (S189) will be well-approximated by the hopping term,

$$w_m(x) \approx 2x_{cm} e^{\frac{x^2 - x_{cm}^2}{2v} - \frac{s_b'(x_{cm} - s_b' - x)}{v}} \left(\frac{x_{cm} - s_b'}{x_{cm}} \right) , \tag{S194}$$

where we have used the auxilliary condition in Eq. (S178) to eliminate the explicit dependence on $\mu_m(s_b')$. This expression is identical to the result in SI Section 4.1.2, which shows that the single-s mapping continues to apply for $x \gtrsim x_{cm} - 2s_b'$.

We can continue these calculations to extend $w_m(x)$ to lower fitness values. At the next rung $[x_{cm} - 3s_b' \lesssim x \lesssim x_{cm} - 2s_b']$, we can once again write x as

$$x = x_{cm} - 2s_b' - \Delta x , \tag{S195}$$

for some $\Delta x \in (0, s_b')$, so that Eq. (S145) reduces to

$$\begin{aligned} w_m(x) \approx\ & 2x_{cm} e^{\frac{x^2 - x_{cm}^2}{2v}} \underbrace{\int_{2s_b' + \Delta x}^{\infty} \frac{\mu_m(s) 2(x+s)}{2x_{cm} e^{\frac{x^2 - x_{cm}^2}{2v}}} \frac{ds}{|x|}}_{\text{landing above } x_{cm}} \\ & + 2x_{cm} e^{\frac{x^2 - x_{cm}^2}{2v}} \underbrace{\int_{s_b' + \Delta x}^{2s_b' + \Delta x} \mu_m(s) e^{\frac{(x_{cm} - 2s_b' - \Delta x)s}{v} + \frac{s^2}{2v}} \frac{ds}{s}}_{\text{landing between } x_{cm} - s_b' \text{ and } x_{cm}} \\ & + 2x_{cm} e^{\frac{x^2 - x_{cm}^2}{2v} - \frac{s_b'(x_{cm} - s_b' - x)}{v}} \underbrace{\int_{\Delta x}^{s_b' + \Delta x} \mu_m(s) e^{\frac{s^2}{2v} + \frac{s(x_{cm} - s_b' - \Delta x)}{v}} \frac{ds}{s + s_b'}}_{\text{landing below } x_{cm} - s_b'} . \end{aligned} \tag{S196}$$

The first term in Eq. (S196) can be evaluated with a similar Laplace approximation as Eq. (S183), yielding

$$2x_{cm} e^{\frac{x^2 - x_{cm}^2}{2v}} \underbrace{\int_{2s_b' + \Delta x}^{\infty} \frac{\mu_m(s) 2(x+s)}{2x_{cm} e^{\frac{x^2 - x_{cm}^2}{2v}}} \frac{ds}{|x|}}_{\text{landing above } x_{cm}} \approx \frac{2x_{cm}}{|x|} \frac{\mu_m(2s_b' + \Delta x)}{-\partial_s \log \mu_m(2s_b' + \Delta x)} . \tag{S197}$$

The integral in the third term in Eq. (S196) is similar to the hopping term in Eq. (S181) (up to a prefactor). This implies that it will have a critical point at $s = s^*(\Delta x)$ with the same location and curvature as Eq. (S184). Finally, the second term in Eq. (S196) corresponds to intermediate leapfrogging events of size $s \in (s_b', 2s_b')$. The integrand in this term will have a critical point at a new value, s_2^*, defined by

$$x_{cm} - 2s_b' - \Delta x + s_2^* + v \frac{\partial \log \mu_m(s_2^*)}{\partial s} = 0 , \tag{S198}$$

which will generally be less than s^*. If the curvature of $\mu_m(s)$ is small $[-\partial_s^2 \log \mu_m(s_b') < 1/v]$, then the critical points in the second and third terms will once again correspond to local minima, and the fixation probability will be dominated by the first term,

$$w_m(x) \approx \frac{2x_{cm}}{|x|} \frac{\mu_m(2s_b' + \Delta x)}{-\partial_s \log \mu_m(2s_b' + \Delta x)} , \tag{S199}$$

similar to Eq. (S186) above. On the other hand, if $-\partial_s^2 \log \mu_m(s_b') \gg 1/v$, then $s^*(\Delta x) \approx s_2^*(\Delta x) \approx s_b'$ will both correspond to local maxima, and the fixation probability will be dominated by the first and third terms:

$$w_m(x) \approx \underbrace{\frac{2x_{cm}}{|x|} \frac{\mu_m(2s_b' + \Delta x)}{-\partial_s \log \mu_m(2s_b' + \Delta x)}}_{\text{leapfrogging}} + \underbrace{2x_{cm} e^{\frac{x^2 - x_{cm}^2}{2v} - \frac{s_b'(x_{cm} - \frac{5}{2}s_b' - 2x)}{v}} \left(\frac{x_{cm} - s_b'}{x_{cm}}\right) \frac{\mu_m(s_b')\sqrt{2\pi\Delta s_b'}}{2s_b'}}_{\text{hopping}} . \tag{S200}$$

By comparing the magnitudes of these terms, we can see that the hopping term will dominate if

$$\log\left[\frac{\mu_m(s_b')}{\mu_m(2s_b' + \Delta x)}\right] - \left[\frac{x_{cm}(s_b' + \Delta x)}{v} - \frac{s_b'^2 + \Delta x^2}{2v}\right] \gtrsim O(1) , \tag{S201}$$

and the corresponding fixation probability will be given by

$$w_m(x) \approx \frac{1}{2} \cdot 2x_{cm} e^{\frac{x^2 - x_{cm}^2}{2v} - \frac{(2x_{cm} - 3s_b' - 2x)s_b'}{v}} \left(\frac{x_{cm} - s_b'}{x_{cm}}\right)^2 , \tag{S202}$$

which matches the single-effect result from SI Section 4.1. For the class of stretched exponential distributions in Eq. (S155) with $\beta \geq 2$, the mutation spectrum term will again be bounded by its Gaussian approximation,

$$-\log \mu_m(2s_b' + \Delta x) \geq -\log \mu_m(s_b') - \frac{(s_b' - x_{cm})(s_b' + \Delta x)}{v} + \frac{(s_b' + \Delta x)^2}{2}\left(\frac{1}{\Delta s_b'^2} - \frac{1}{v}\right) , \tag{S203}$$

so that Eq. (S201) reduces to

$$-\frac{s_b'(s_b' + \Delta x)}{v} + \frac{s_b'^2 + \Delta x^2}{2\Delta s_b'^2} + \frac{s\Delta x}{\Delta s_b'^2}\left(1 - \frac{\Delta s_b'^2}{v}\right) \gtrsim O(1) . \tag{S204}$$

This shows that the hopping approximation will be valid whenever $v \gg \Delta s_b'^2$, which is the same condition that we obtained for the previous interval above. One can continue this argument to show that for relative fitnesses of the form,

$$x = x_{cm} - ns_b' - \Delta x , \tag{S205}$$

with $\Delta x \in (0, s_b')$, the hopping term will be given by

$$w_m(x) \approx \frac{1}{n!} \cdot 2x_{cm} e^{\frac{x^2 - x_{cm}^2}{2v} - \frac{ns_b'}{v}(x_{cm} - \frac{(n+1)}{2}s_b' - x)} \left(\frac{x_{cm} - s_b'}{x_{cm}}\right)^n , \tag{S206}$$

and will constitute the dominant contribution whenever

$$\log\left[\frac{\mu_m(s_b')}{\mu_m(ns_b' + \Delta x)}\right] - \left[\frac{x_{cm}(n-1)s_b' + x_{cm}\Delta x}{v} - \frac{(n-1)^2 s_b'^2 + \Delta x^2}{2v}\right] \gtrsim O(1) . \tag{S207}$$

This shows that for the stretched exponential distributions in Eq. (S155) with $\beta \geq 2$, the single-s approximation will continue to apply for arbitrarily large fitness costs as long as $v \gg \Delta s_b'^2$.

5.3.3 Fixation probabilities of modifiers

We can now use our solution for $w_m(x)$ to evaluate the fixation probability of a modifier mutation. When the dynamics are dominated by hopping, we have seen that the functional form of $w_m(x)$ is well-approximated by the single-s model in SI Section 4.1. This implies that the fixation probability of the modifier can be predicted using our earlier expressions in SI Sections 4.1.4 and 4.1.5

$$\tilde{p}_{\text{fix}}(\mu \to \mu_m; s_m) = \tilde{p}_{\text{fix}}(s_b, U_b \to s_b', U_b'; s_m) \,, \tag{S208}$$

if (s_b, U_b) and (s_b', U_b') are chosen to coincide with the effective parameters defined in SI Sections 5.1 and 5.3.1 above. This approximation will be valid when $\log \tilde{p}_{\text{fix}}(\mu_m) \gtrsim x_c/s_b$ and $x_{cm} - s_b' \gtrsim \sqrt{v}$. We used this expression to obtain the theoretical predictions for the exponential distribution in Fig. 4B,C.

This single-s approximation will also be valid for $\mu_m(s)$ that allow for leapfrogging, as long as the direct or indirect costs of the modifier are not too strong ($s_m + x_c - x_{cm} \gtrsim -s_b'$). For larger costs ($s_m + x_c - x_{cm} \lesssim -s_b'$), we will need to modify Eq. (S208) to include the additional contributions from the leapfrogging term. Substituting the solution for $w_m(x)$ in Eq. (S186) into Eq. (S92), we obtain

$$\begin{aligned}
\tilde{p}_{\text{fix}}(\mu \to \mu_m; s_m) &\approx \frac{x_{cm}}{x_c s_b} e^{\frac{x_c^2}{2v}} \int_{-\infty}^{x_c+s_m} \frac{\mu_m(x_{cm}-x)}{-\partial_s \log \mu_m(x_{cm}-x)} e^{-\frac{(x-s_m)^2}{2v}} \frac{dx}{|x|} \,, \\
&\approx \frac{(x_{cm}-s_b')s_b'}{x_c s_b} e^{\frac{x_c^2 - x_{cm}^2}{2v}} \int_{x_{cm}-(x_c+s_m)}^{\infty} \frac{\mu_m(u)}{\mu_m(s_b')} \cdot \frac{e^{\frac{(x_{cm}-s_b')^2}{2v} - \frac{(u-(x_{cm}-s_m))^2}{2v}}}{-|x_{cm}-u| \cdot \partial_s \log \mu_m(u)} \frac{du}{\sqrt{2\pi \Delta s_b'^2}} \,,
\end{aligned} \tag{S209}$$

where we have substituted the auxiliary condition for x_{cm}. This integral will have a local maximum at u^*, where u^* satisfies

$$x_{cm} - s_m - u^* + v\partial_s \log \mu_m(u^*) = 0 \,, \tag{S210}$$

and a characteristic width Δ_u,

$$\Delta_u = \left[\frac{1}{v} - \frac{\partial^2 \log \mu_m(u^*)}{\partial s^2} \right]^{-1/2} \,. \tag{S211}$$

We can therefore evaluate it with a Gaussian Laplace approximation to obtain,

$$\tilde{p}_{\text{fix}}(\mu \to \mu_m; s_m) \approx \frac{(x_{cm}-s_b')s_b'}{x_c s_b} e^{\frac{x_c^2 - x_{cm}^2}{2v}} \frac{\mu_m(u^*)}{\mu_m(s_b')} \cdot \frac{e^{\frac{(x_{cm}-s_b')^2}{2v} - \frac{(u^*-(x_{cm}-s_m))^2}{2v}}}{-|x_{cm}-u^*| \cdot \partial_s \log \mu_m(u^*)} \frac{\sqrt{2\pi\Delta_u^2}}{\sqrt{2\pi \Delta s_b'^2}} \,. \tag{S212}$$

In the case of an exponential distribution, $u^* = s_b' - s_m$ and $\Delta_u = \sqrt{v}$, so Eq. (S212) reduces to,

$$\tilde{p}_{\text{fix}}(\mu \to \mu_m; s_m) \approx \frac{v s_b'}{x_c s_b |s_m|} e^{\frac{x_c^2 - x_{cm}^2}{2v} + \frac{s_m}{s_0'}} \,. \tag{S213}$$

For a half-Gaussian distribution, $u^* = s_b'(1 - s_m/x_{cm})$ and $\Delta_u = \sqrt{\frac{1}{v} + \frac{2}{s_0'^2}}$, so that

$$\tilde{p}_{\text{fix}}(\mu \to \mu_m; s_m) \approx \frac{v x_{cm}^2}{x_c s_b s_b' s_m^2} e^{\frac{x_c^2 - x_{cm}^2}{2v} + \frac{2 s_m x_{cm} - \delta^2}{s_0'^2 + 2v}} \approx \frac{v x_{cm}^2}{x_c s_b s_b' s_m^2} e^{\frac{x_c^2 - x_{cm}^2}{2v} - \frac{s_m^2}{s_0'^2}} \tag{S214}$$

where the last approximation follows when $s_0' \gg \sqrt{v}$. These results imply that the leading order scaling for large direct costs when leapfrogging is valid is given by

$$\log \tilde{p}_{\text{fix}} \approx \frac{x_c^2}{2v} - \frac{x_{cm}^2}{2v} + \log(\mu_m(s_m)s_0'/U_b') \,. \tag{S215}$$

5.4 Modifier-dominated regime (quasi-sweeps)

The results in the previous section apply when $x_{cm} - s_b' \gg \sqrt{v}$. As the gap between x_{cm} and s_b' shrinks, we will eventually reach a point where a single established mutation will be sufficient to drive the modifier lineage to fixation. To solve for the shape of $w_m(x)$ in this "quasi-sweeps" regime, we must revisit the equations for $w_m(x)$ and x_{cm} in Eqs. (S145) and (S146).

5.4.1 Location of the interference threshold

Motivated by our solution for the quasi-sweeps regime in SI Section 4.2, we anticipate that the dominant contributions to the y-integral in Eq. (S146) come from the Haldane region of the shoulder solution ($y > x_{cm}$). When $s \gtrsim x_{cm} + O(\sqrt{v})$, the integrand will be peaked around a characteristic value $y^* = s \pm O(\sqrt{v})$. Performing a Gaussian Laplace approximation around this maximum yields,

$$1 \approx \frac{\sqrt{2\pi}}{x_{cm}\sqrt{v}} e^{\frac{x_{cm}^2}{2v}} \int_{x_{cm}}^{\infty} s\mu_m(s)\, ds \,. \tag{S216}$$

This is similar to the single-effect model in SI Section 4.2.1 if we take

$$U_b' = \int_{x_{cm}}^{\infty} \mu_m(s) ds \,, \quad s_b' = \frac{1}{U_b'} \int_{x_{cm}}^{\infty} s\mu_m(s)\, ds \,. \tag{S217}$$

The decomposition in Eqs. (S145) and (S146) shows that this s_b' integral coincides with the distribution of fixed mutations in the modifier lineage. In contrast to the multiple mutations regime in SI Section 5.3, the distribution in Eq. (S217) may have a typical scale but will generally not be strongly peaked.

This quasi-sweeps regime will be self-consistently valid if Eq. (S216) is much larger than the remaining contributions from $y \leq x_{cm}$:

$$\int^{x_{cm}} \frac{\mu_m(s)}{v} e^{\frac{x_{cm}s}{v} - \frac{s^2}{2v}} \ll \frac{\sqrt{2\pi}}{x_{cm}\sqrt{v}} e^{\frac{x_{cm}^2}{2v}} \int_{x_{cm}}^{\infty} s\mu_m(s)\, ds \,. \tag{S218}$$

For the stretched exponential distributions in Eq. (S155), this will only be true if $x_{cm} \lesssim s_0'$, so that $U_b' \sim O(U_0')$ and $s_b' \sim O(s_0')$. This shows that the quasi-sweeps regime will be valid when

$$s_0' \gg x_{cm} \sim \sqrt{2v \log\left(\frac{v}{U_0' s_0'}\right)} \,. \tag{S219}$$

5.4.2 Extending the shoulder solution to lower fitness values

We can continue this line of reasoning to compute the shape of $w_m(x)$ for fitness values below x_{cm}. For mutations with fitness effects $s \gtrsim x_{cm} - x + O(\sqrt{v})$, the y-integral in Eq. (S145) will continue to have a

large contribution from the Haldane region of the shoulder solution. For these values of s, we can repeat the Gaussian integration in Eq. (S127) to obtain

$$w_m(x) \approx 2x_{cm} e^{\frac{x^2}{2v}} \int_{x_{cm}-x+O(\sqrt{v})}^{\infty} \mu_m(s) \left\{ \frac{s\sqrt{2\pi}}{x_{cm}\sqrt{v}} \cdot \left[\Phi\left(\frac{x}{\sqrt{v}}\right) - \Phi\left(\frac{x_{cm}-s}{\sqrt{v}}\right) \right] - \frac{e^{\frac{-x^2}{2v}} - e^{\frac{-(x_{cm}-s)^2}{2v}}}{x_{cm}} \right\} ds\,, \tag{S220}$$

where $\Phi(z) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{z} e^{-u^2/2}\, du$ is the Gaussian cumulative distribution function. When $x \gtrsim O(\sqrt{v})$, the Gaussian cumulative function is close to one, and Eq. (S220) reduces to

$$w_m(x) \approx 2x_{cm} e^{\frac{x^2-x_{cm}^2}{2v}} \frac{\sqrt{2\pi}}{x_{cm}\sqrt{v}} \int_{x_{cm}}^{\infty} \mu_m(s) s\, ds \;= 2x_{cm} e^{\frac{x^2-x_{cm}^2}{2v}}\,, \tag{S221}$$

where the last line follows from substituting the auxiliary condition for x_{cm} in Eq. (S216). This shows that the shoulder solution will be valid down to $x_{\min} = O(\sqrt{v})$, similar to the single-effect model in SI Section 4.2.2.

When $x \lesssim O(\sqrt{v})$, the shape of $\Phi(x)$ will start to become important. At this point, the boundary terms at $x_{cm} - s$ will become negligible, so that Eq. (S220) reduces to

$$\begin{aligned} w_m(x) &\approx 2x_{cm} e^{\frac{x^2-x_{cm}^2}{2v}} \int_{x_{cm}-x}^{\infty} \mu_m(s) \left[\frac{s\sqrt{2\pi v} e^{\frac{x_{cm}^2}{2v}}}{v x_{cm}} \cdot \Phi\left(\frac{x}{\sqrt{v}}\right) - \frac{e^{\frac{x_{cm}^2-x^2}{2v}}}{x_{cm}} \right] ds\,, \\ &\approx 2x_{cm} e^{\frac{x^2-x_{cm}^2}{2v}} \left[\Phi\left(\frac{x}{\sqrt{v}}\right) \left(1 - \frac{\int_{x_{cm}}^{x_{cm}-x} s\mu_m(s) ds}{\int_{x_{cm}}^{\infty} s\mu_m(s) ds} \right) - \frac{\sqrt{v}}{s_b'\sqrt{2\pi}} e^{-\frac{x^2}{2v}} \left(1 - \frac{\int_{x_{cm}}^{x_{cm}-x} \mu_m(s) ds}{\int_{x_{cm}}^{\infty} \mu_m(s) ds} \right) \right]\,, \end{aligned} \tag{S222}$$

where the second line follows from substituting the auxiliary condition for x_{cm} in Eq. (S216). This will coincide with the analogous expression for the single-s model in 4.1.2 if

$$\int_{x_{cm}}^{x_{cm}-x} \mu_m(s)\, ds \ll \int_{x_{cm}}^{\infty} \mu_m(s)\, ds\,. \tag{S223}$$

This condition will be satisfied if $x_{cm} - x \lesssim s_b'$, or alternatively, if $x \gtrsim x_{cm} - s_b'$.

For lower initial fitnesses, the contributions from mutations that land below the interference threshold ($x + s < x_{cm}$) can begin to become important. We can estimate the onset of these effects by considering initial fitnesses in the range $-O(\sqrt{v}) \gtrsim x \gtrsim x_{cm} - s_b'$. In this limit, Eq. (S222) reduces to

$$w_m(x) \approx \frac{2U_b'(s_b' - |x|)}{|x|} \approx \frac{2x_{cm}\sqrt{v}}{\sqrt{2\pi}|x|} \left(1 - \frac{|x|}{s_b'} \right) e^{-\frac{x_{cm}^2}{2v}}\,, \tag{S224}$$

where the second line follows from substituting the auxiliary condition for x_{cm} in Eq. (S216). The corresponding contributions from mutations with $x + s < x_{cm}$ in Eq. (S145) are dominated by the upper limit of the y-integral, which yields an additional correction of order

$$\delta w_m(x) \approx 2x_{cm} e^{\frac{x^2-x_{cm}^2}{2v}} \int_{O(\sqrt{v})-x}^{x_{cm}-x} \mu_m(s) e^{\frac{xs}{v} + \frac{s^2}{2v}} \frac{ds}{s}\,. \tag{S225}$$

If the curvature of $\mu_m(s)$ is not too strong $[\partial_s \log \mu_m(s'_b) \ll s'_b/v \ll s'_b]$, this s-integral will be dominated by the upper limit of integration, so that

$$\delta w_m(x) \approx \frac{2v \cdot \mu(x_{cm} - x)}{x_{cm} - x} \,. \tag{S226}$$

This will be small compared to Eq. (S224) as long as $s'_b \gtrsim x_{cm} \gg \sqrt{v}$, which shows that Eq. (S222) will be a good approximation for $w_m(x)$ as long as $x \gtrsim x_{cm} - s'_b$.

For lower initial fitnesses ($x \lesssim x_{cm} - s'_b$), these additional contributions will start to become important. In principle, we can generalize the recursive approach in SI Section 4.2.2 to extend our solution for $w_m(x)$ below this point. However, these corrections will no longer be universal, and will depend on other features the mutation spectrum beyond the first two moments in Eq. (S217). Since our results in the main text will only require the portion of the solution for $x \gtrsim x - s'_b$, we will leave these calculations for future work.

6 Incorporating deleterious mutations

While the previous sections have focused on changes to the beneficial mutation spectrum, the vast majority of new mutations are neutral or deleterious (45). The most general evolvability modifiers could alter the rates and fitness costs of these deleterious mutations as well.

Previous work (46, 103, 105) has shown in our parameter regime of interest, deleterious mutations can be divided into two broad categories: (i) quasi-neutral mutations ($|s| \ll v/x_c$), which can frequently hitchhike to fixation with beneficial mutations, and (ii) "purgeable" deleterious mutations ($|s| \gg v/x_c$), which rarely reach high frequencies, but can collectively reduce the effective population size if their mutation rate is sufficiently high. Previous estimates suggest that the overall mutation rates in many natural and laboratory microbial populations are sufficiently low ($U \lesssim 10^{-3}$ (106)) that both categories have a negligible impact on the overall the rate of adaptation (46, 103). However, differences in the deleterious distribution of fitness effects can still have a strong influence on the fixation probability of a modifier mutation (24).

We can incorporate purgeable deleterious mutations into our framework using the argument outlined in Ref. (24), which we reproduce here for completeness. Ref. (24) showed that it is useful to rewrite the mutation term in Eq. (S33) as separate beneficial and deleterious contributions,

$$\begin{aligned} 0 = \underbrace{x \cdot w_m(x)}_{\text{selection}} - \underbrace{v \cdot \partial_x w_m(x)}_{\text{competition w/ wildtype}} - \underbrace{\frac{1}{2} \cdot w_m(x)^2}_{\text{drift while rare}} + \underbrace{\int_0^\infty \mu_m(s)\,[w_m(x+s) - w_m(x)]\; ds}_{\text{beneficial mutations}} \\ + \underbrace{\int_0^\infty \mu_m(-\delta)\,[w_m(x-\delta) - w_m(x)]\; d\delta}_{\text{deleterious mutations}} \end{aligned} \tag{S227}$$

where we are indexing the deleterious mutations by their overall magnitude $\delta \equiv -|s|$. The definition of a purgeable mutation is that they rarely reach high frequencies, which implies that the fixation probability after acquiring a purgeable mutation must be close to zero. If all of the deleterious mutations in μ_m fall in this purgeable category, we can therefore neglect the $w_m(x-\delta)$ term in Eq. (S227). This allows us to rewrite Eq. (S227) in the simpler form,

$$0 = (x - U'_d) w_m(x) - v \cdot \partial_x w_m(x) - \frac{1}{2} \cdot w_m(x)^2 + \int_0^\infty \mu_m(s)\,[w_m(x+s) - w_m(x)]\; ds \tag{S228a}$$

where U'_d denotes the total mutation rate for purgeable mutations:

$$U'_d = \int_0^\infty \mu_m(-|\delta|)d\delta \,. \tag{S228b}$$

This equation has the same form as the purely beneficial case analyzed above, but with the x coordinate shifted by $-U'_d$. The solutions can therefore be expressed in the simple form

$$w_m(x) = w_m^b(x - U'_d) \,, \tag{S229}$$

where $w_m^b(x)$ denotes our existing results for the purely beneficial case. Likewise, previous work (24, 46) has shown that the background fitness distribution satisfies an analogous formula,

$$f(x) = f^b(x - U_d) \,, \tag{S230a}$$

where U_d is the rate of producing purgeable mutations in the wildtype background,

$$U_d = \int_0^\infty \mu(-|\delta|)d\delta \,. \tag{S230b}$$

Substituting these expressions into Eq. (2), we find that the fixation probability of the modifier can be written as

$$p_{\text{fix}}(\mu \to \mu_m; s_m) = \int f^b(x - U_d) \cdot w_m^b(x - U'_d + s_m)\, dx = \int f^b(y) \cdot w_m^b(y + s_m - U'_d + U_d)\, dy \,, \tag{S231}$$

which implies that differences in the purgeable mutation rate behave like effective direct cost,

$$s_m^{\text{eff}} = U_d - U'_d \,. \tag{S232}$$

We validated this approximation using Wright-Fisher simulations (Fig. S2), and found that it encompasses a broad range of deleterious fitness effects, including those much smaller than a typical beneficial driver mutation (Fig. S2B). Deviations eventually occur for more weakly selected mutations, which have a lower impact on the fixation probability than Eq. (S232) would predict. This suggests that purgeable mutations represent an upper bound on the strength of second-order selection.

Interestingly, the mapping in Eq. (S232) implies that the same modulation effect that occurs between first- and second-order selection will also apply for selection on simultaneous changes to robustness ($U_d \to U_d^{'}$) and evolvability ($\mu \to \mu_m$). In particular, it implies that larger populations are more likely to trade reduced robustness for long-term gains in evolvability. Since empirical deleterious mutation rates are often comparatively low ($|s_m^{\text{eff}}| \lesssim s_b$), even the maximum possible reduction in robustness will be overpowered by marginal increases in evolvability (Fig. S2C). The opposite scenario – greedily selecting for increased robustness despite long-term reductions in evolvability – is also theoretically possible, but we expect that it will be less relevant in practice due to the low deleterious mutation rates in many organisms.

Finally, we note that robustness and evolvability do not have to trade off with one other – several recent studies have shown that they can sometimes interact synergistically as well (61, 63). Our analytical framework also provides new predictions for these synergistic mutations, allowing us to determine how each phenotype contributes to the lineage's long-term evolutionary fate (Fig. S2E). Our results show that larger populations will tend to weigh proportional enhancements in evolvability ($\Delta v/v$) more heavily than comparable increases in robustness ($\Delta U_d/U_d$). This is qualitatively different from the successive sweeps picture (SI Section 2), which predicts that the opposite ordering should occur. These examples illustrate how clonal interference can reshape our intuition about second-order selection for evolvability.

7 Extensions to more general fitness landscapes

Our analysis above focused on the simplest possible model of an evolvability modifier, where the local distribution of fitness effects could be approximated by a pair of fixed distributions,

$$\mu(s|\vec{g}) \approx \mu(s)\,, \quad \mu_m(s|\vec{g}) \approx \mu_m(s)\,, \tag{S233}$$

for all genotypes $\vec{g}$ (SI Section 3.1). In this section, we consider several extensions of our model that relax this assumption in different ways. In each case, we find that our existing theory provides a useful baseline for incorporating these new effects.

7.1 Weak Macroscopic Epistasis

It is clear that some deviations from Eq. (S233) will be too small to be relevant to natural selection. We can formalize this idea by considering subtle deviations of the form,

$$\mu(s|\vec{g}) = \mu(s) + \epsilon(s, \vec{g})\,, \tag{S234}$$

where the function $\epsilon(s, \vec{g})$ represents a small perturbation. In this case, our existing results in SI Section 5 allow us to precisely define what we mean by "small". In particular, if we let $\delta\mu(s) = \epsilon(s, \vec{g})$, then we can conclude that any perturbation for which the integral Eq. (8) is small $[I(\delta\mu) \lesssim s_b/x_c]$ will be essentially invisible to natural selection. Any deviations from Eq. (S233) that fall below this minimum resolution will therefore not affect our main results.

We note that this space of "negligible" perturbations can be quite large from the perspective of the original DFE. Fig. 4 shows that even large fluctuations in $\mu(s)$ can be tolerated for fitness effects $\lesssim s_b(\mu, N)$, even if they cause in large shifts in the overall mean and height of the DFE. Conversely, much smaller perturbations at fitness effects $\gtrsim s_b(\mu, N)$ can lead to important deviations from our original model, even if they nominally appear to satisfy Eq. (S233).

We also note that this argument constitutes an upper bound on the impact of a given deviation from Eq. (S233), since our derivation of Eq. (8) assumed that the evolvability differences were permanent. This means that some values of $\epsilon(s, \vec{g})$ that exceed the resolution limit above could still have a small influence on the results because their evolvability differences are only transient. We consider such cases in more detail below.

7.2 Transient differences in evolvability

In many cases of interest, the evolvability benefits of a modifier will not persist indefinitely (as in Eq. S233) but will only apply within some local region of genotype space. For example, in the stability-activity landscape in Fig. S1C, the stability enhancing modifier creates an opportunity for K additional mutations to accumulate. At a formal level, this constitutes a large deviation from Eq. (S233).

However, it is also clear that such permanent shifts are not truly necessary. Second-order selection can only take place while the modifier is competing with the wildtype; once one of these lineages fixes, future changes in $\mu(s|\vec{g})$ can no longer contribute to the fixation probability. Our results above allow us to estimate the size of this critical window. In particular, our heuristic analysis in SI Section 4.1.4 shows that the benefits of an evolvability modifier accumulate over $K^* \sim \max\{x_{cm}/s'_b, 1\}$ mutational steps. This suggests that changes to the DFE that occur outside this horizon will have a negligible impact on our results.

We tested this prediction by considering a simple model where the modifier reverts to the wildtype DFE after K mutational steps:

$$\mu_m(s|\vec{g}) = \begin{cases} \mu_m(s) & \text{if } |\vec{g} - \vec{g}_m| < K, \\ \mu(s) & \text{else.} \end{cases} \tag{S235}$$

Simulations of this model show that our existing theory continues to provide a good approximation even when K is as low as 2 or 3 (Fig. S3A), consistent with the moderate values of $q \equiv x_c/s_b$ that are attained for many empirically relevant parameter values (39, 47). Similar results are also observed if the modifier reverts to a dead-end rather than the wildtype (Fig. S3B), with differences arising only for the smallest values of K. In both cases, we find that the evolvability benefits still confer exponentially large advantages over the classical SSWM expectation (Fig. S3A), even when they last for just a single mutation ($K = 1$).

7.3 Global diminishing returns epistasis

Many evolving populations exhibit a form of diminishing returns epistasis, where the fitness benefits of new mutations systematically decline with a lineage's absolute fitness (9, 10, 41, 54, 55, 89). To understand how this phenomenon might impact our results, we considered a simple model of global epistasis inspired by the long-term evolution experiment in Fig. 1A (55), where the typical fitness benefits of new mutations decline exponentially with the total fitness,

$$\mu(s|\vec{g}) = U_b \cdot \delta(s - \tilde{s}_b e^{-X(\vec{g})/\theta}), \tag{S236}$$

where θ controls the strength of the diminishing returns effect. In the simplest scenario, we can assume that the modifier distribution exhibits a similar decline,

$$\mu_m(s|\vec{g}) = U'_b \cdot \delta(s - \tilde{s}'_b e^{-X(\vec{g})/\theta}), \tag{S237}$$

but with different values of U'_b and s'_b. This generalizes the simple toy model in Fig. 2 to allow for steadily declining fitness effects.

Adiabatic approximation. The simplest behavior occurs when θ is sufficiently large. Suppose that the modifier arises at time t_0 when the mean fitness of the population is $\overline{X}(t_0)$. If we could neglect the additional deceleration that occurs over the lifetime of a single mutation, we could simply apply our existing results with $s_b = \tilde{s}_b e^{-\overline{X}(t_0)/\theta}$ and $s'_b = \tilde{s}'_b e^{-\overline{X}(t_0)/\theta}$. We can use our heuristic picture in SI Section 4.1.4 to determine when this "adiabatic approximation" will be valid. The fate of a mutation is determined over $T_c \sim x_c/v$ generations, during which time the total fitness of the population increases by $\Delta\overline{X} \sim x_c$. If $\theta \gg x_c$, then diminishing returns epistasis will have a negligible effect over the lifetime of a single mutation, even if it had a large effect in setting the initial scale of selection ($\overline{X}(t_0) \gtrsim \theta$). This provides a self-consistent justification for our adiabatic approximation above.

This simple picture leads to novel predictions when combined with our existing results in Eqs. (3) and (4). In particular, it suggests that a selection-strength modifier will become less strongly selected over time as the population becomes better adapted, while the benefits of a mutator allele will remain roughly constant over the same time window. Since x_c is roughly proportional to s_b (Eq. S49), we expect that θ/x_c will grow increasingly large at long times as s_b becomes progressively smaller (Fig. S4). This suggests that the adiabatic approximation will often be useful for understanding the long-term dynamics of a population (41).

Beyond the adiabatic approximation. The situation becomes more complicated when x_c is comparable to θ, since the fitness benefits of new mutations will now decline within the scale of the population fitness distribution. The dynamics in this regime are poorly understood even in the absence of second-order selection (41). However, we can still account for these effects in a crude way by leveraging our heuristic picture in SI Section 4.1.4, and focusing on the leading-order corrections when x_c/θ is small but finite. We briefly review the results for our original model below, and then show how they can be extended to calculate the leading-order corrections from diminishing returns epistasis.

Recall that for small changes in the selection coefficient ($s'_b - s_b \ll s'_b$), successful modifiers will typically arise in the high-fitness nose of the population ($x \approx x_c$) and acquire x_c/s_b mutations before reaching appreciable frequencies (SI Section 4.1.4). In each of these steps ($j = 1, \ldots, x_c/s_b$), a selection-strength modifier will grow for $\tau_{\text{est}} \approx s_b/v$ generations while the next nose establishes. During this establishment time, the modifier will produce $\sim \exp\left[\frac{s_b}{v} \cdot j(s'_b - s_b)\right]$ more mutations than a wildtype individual with the same fitness, leading to the scaling

$$\log \tilde{p}_{\text{fix}} \sim \frac{s_b}{2v} \cdot \frac{x_c}{s_b}\left(\frac{x_c}{s_b} - 1\right)\left(s'_b - s_b\right) . \tag{S238}$$

This picture becomes more complicated with diminishing returns epistasis since the wildtype and modifier selection coefficients will decline with each subsequent mutation,

$$s'_{b,j} \equiv s'_b(t_0) e^{-\sum_{i=1}^{j-1} s'_{b,i}/\theta} , \tag{S239a}$$

$$s_{b,j} \equiv s_b(t_0) e^{-\sum_{i=1}^{j-1} s_{b,i}/\theta} , \tag{S239b}$$

where $s'_b(t_0)$ and $s_b(t_0)$ denote the effective selection coefficients at the time that the modifier arises. As a result, the modifier's growth advantage, and the time it spends growing, will also vary with each mutation. We can straightforwardly extend our heuristic prediction to account for these differences,

$$\log \tilde{p}_{\text{fix}} \sim \sum_{i=1}^{\frac{x_c}{s_b}-1} \sum_{j=1}^{i} \tau_{\text{est},j} \cdot s_{b,j} \cdot \left(\frac{s'_{b,j}}{s_{b,j}} - 1\right) , \tag{S240}$$

where $\tau_{\text{est},j}$ denotes the establishment time of the jth mutational step. These establishment times will emerge from the non-equilibrium dynamics imposed by diminishing returns epistasis, including the declining rate of fitness increase and width of the fitness distribution. While these dynamics remain poorly characterized even in the absence of second order selection, we find that we can obtain accurate predictions by assuming that the product $\tau_{\text{est},j} \cdot s_{b,j}$ remains approximately constant during the lifetime of the modifier:

$$\tau_{\text{est},j} \cdot s_{b,j} \approx \tau_{\text{est}}(t_0) \cdot s_b(t_0) . \tag{S241}$$

This assumption is motivated by the logarithmic dependence of $\tau_{\text{est}} \cdot s_b$ on s_b in the absence of diminishing returns epistasis (47; Eq. S49), which suggests that treating it as constant over fixation timescales will often be a good approximation. With this assumption, Eq. (S240) reduces to the simpler form

$$\log \tilde{p}_{\text{fix}} \sim \sum_{i=1}^{\frac{x_c}{s_b}-1} \sum_{j=1}^{i} \tau_{\text{est}}(t_0) \cdot s_b(t_0) \cdot \left(\frac{s'_{b,j}}{s_{b,j}} - 1\right) . \tag{S242}$$

Under the model in Eq. (S239), the diminishing returns effects on $s'_{b,j}$ and $s_{b,j}$ exactly cancel each other, so that we end up with the same adiabatic approximation above. This suggests that the next-order corrections remain small when the modifier and wildtype follow similar diminishing returns schedules.

We can also consider modifiers that alter the form of the diminishing returns epistasis itself. The simplest example is one that changes the diminshing returns parameter θ to a new value θ_m once the modifier mutation arises:

$$\mu_m(s|\vec{g}) = U_b' \cdot \delta\left(s - \tilde{s}_b' e^{-X(\vec{g}_m)/\theta - [X(\vec{g}) - X(\vec{g}_m)]/\theta_m}\right), \tag{S243}$$

where $\vec{g}_m$ denotes the founding genotype of the modifier lineage. The difference between θ and θ_m is invisible in the adiabatic approximation, but the next order corrections can be obtained from Eq. (S242), by substituting

$$s_{b,j}' \equiv s_b'(t_0) e^{-\sum_{i=1}^{j-1} s_{b,i}'/\theta_m}, \tag{S244a}$$

$$s_{b,j} \equiv s_b(t_0) e^{-\sum_{i=1}^{j-1} s_{b,i}/\theta}. \tag{S244b}$$

The diminishing returns contributions no longer cancel, leading to modest corrections to our existing theory (blue lines, Fig. S4D). In particular, when $\theta_m \gg \theta$, these diminishing returns modifiers can be positively selected even when their initial DFEs are the same $[s_b'(t_0) = s_b(t_0)]$, since the modifier experiences fewer diminishing returns effects at later times. These differences disappear once $s_b'(t_0)$ is sufficiently large that we enter the quasi-sweeps regime. In this case, we have seen that the success of the modifier is driven by its ability to acquire its first additional mutation (SI Section 4.2.3), where the new value of θ_m has not yet taken effect. These extensions of our heuristic analysis suggest that the simple model we have studied in this work may be a powerful tool to understand selection for evolvability on more general fitness landscapes.

8 Forward-Time Simulations and Numerical Methods

8.1 Forward-time simulations

We validated our theoretical predictions by comparing them to forward-time, Wright-Fisher-like simulations similar to those employed by Ref. (24). We begin with a population of $N_{t=0} \equiv N$ individuals at generation 0. In each generation, every individual i in the population produces a Poisson number of identical offspring with mean

$$\lambda_0 = C(t) \cdot (1 + X_i - \overline{X}(t))[1 - \int \mu(s) ds], \tag{S245a}$$

and a Poisson number of mutant offspring with mean

$$\lambda_m = C(t) \cdot (1 + X_i - \overline{X}(t)) \int \mu(s)\, ds, \tag{S245b}$$

where X_i is the absolute fitness of individual i, $\overline{X} = \sum_i X_i / \sum_i 1$ is the mean fitness of the population, and $C(t) = N/N_t$ is a normalization constant that ensures that the expected population size in each generation is equal to N. Each mutant offspring is assigned a new fitness value $X_i + s$, where s is randomly sampled from the normalized version of $\mu(s)$. After initialization, each simulation is allowed to "burn in" for $\Delta t = 2 \cdot 10^4$ generations.

To measure the fixation probability of a modifier lineage, we continue this basic algorithm while allowing individuals in the background population to produce new modifier offspring at a per capita rate U_m. These modifier individuals produce offspring according to an analogous version of Eq. (S245), with $\mu(s) \to \mu_m(s)$. Reversions from modifier to wildtype genotypes are not allowed. Following the burn in period, we record the

number of generations T that elapse until a modifier lineage fixes in the population. In the limit that $U_m \to 0$, the fixation time T is related to the fixation probability through the relation,

$$\tilde{p}_{\text{fix}}(\mu \to \mu_m) = \frac{1}{U_m T} \,. \tag{S246}$$

For the estimator in Eq. (S246) to apply, U_m must be small enough that the total time to fixation time T is much larger than the predicted sweep time of a successful modifier ($\sim x_{cm}/v$). To ensure that this is true, we repeat this process of measuring T for $M = 60$ replicate populations with a sequence of modifier mutation rates $U_{m,1}, U_{m,1}, ...U_{m,M}$, yielding a sequence of fixation times $T_1, T_2, \ldots T_m$; the mutation rate at step i is chosen based on the previous T_{i-1},

$$U_{m,i} \equiv \min\left\{\frac{U_{m,i-1}T_{i-1}}{T^*}, 10^{-2}\right\} \,, \tag{S247}$$

so that the predicted value of T_i is $T^* \equiv 250 \cdot x_{cm}/v$ generations. The mutation rate is also capped at 10^{-2} so that a successful modifier lineages always competes against the background population while small. The sequence is started at $U_{m,i-1} = 1/T^*$ and allowed to "burn-in" for 10 iterations before the T_i's are recorded. The fixation probability is calculated from the maximum likelihood estimator,

$$\tilde{p}_{\text{fix}}(\mu \to \mu_m) \approx \frac{1}{\sum_{i=10}^{60} U_{m,i}T_i} \,. \tag{S248}$$

A copy of our implementation in Python is available in the associated Github repository (`https://github.com/bgoodlab/evolution_of_evolvability`).

8.2 Theoretical predictions

The theoretical predictions in Figs. 2-4 were generated using the following procedures.

Figure 2. To generate the theoretical predictions in Fig. 2, we first solved for x_{cm} numerically as a function of s'_b, U'_b, and v. We found that numerical accuracy was improved by using a modified version of Eqs. (S56) and (S118),

$$1 = \frac{U'_b}{s'_b}\left[e^{\frac{x_{cm}s'_b}{v} - \frac{s'^2_b}{2v}} + \frac{\sqrt{2\pi}s'^2_b}{x_{cm}\sqrt{v}} e^{\frac{x^2_{cm}}{2v}} \Phi\left(\frac{s'_b - x_{cm}}{\sqrt{v}}\right)\right] \,, \tag{S249}$$

which has the same asymptotic behavior, but smoothly captures the transition between the multiple-mutations ($x_{cm} - s'_b \gg \sqrt{v}$) and quasi-sweeps regimes ($s'_b - x_{cm} \gg \sqrt{v}$). Numerical solutions to Eq. (S249) were obtained using the `fsolve` function from the SciPy library (107), with the measured value of v obtained from our forward-time simulations. We used the same procedure to solve for the wildtype interference threshold x_c using the analogous version of Eq. (S249). If $x_{cm} > s'_b$, we estimated the fixation probability of the modifier using the multiple-mutations expression in Eq. (S84), while the quasi-sweeps prediction in Eq. (S135) was used for $x_{cm} < s'_b$. We used this procedure to generate all of the theory lines in panels A-C.

The boundary for the gray region in Fig. 2D was obtained by identifying the values of U'_b and s'_b that minimize the true fixation probability $\tilde{p}_{\text{fix}}(s_b, U_b \to s'_b, U'_b)$ at a fixed value of $x = \tilde{p}_{\text{fix}}(s_b \to s'_b) \cdot \tilde{p}_{\text{fix}}(s_b \to s'_b)$. Using the leading-order approximations in Eqs. (S88), (S89), and (S90), the fixation probability of the compound modifier can be approximated as

$$\tilde{p}_{\text{fix}}(s_b, U_b \to s'_b, U'_b) \approx x^{s^2_b/s'^2_b} \cdot \tilde{p}_{\text{fix}}(s_b \to s'_b)^{1-s^2_b/s'^2_b} \,, \tag{S250}$$

which eliminates the explicit dependence on U_b'. Minimizing this function with respect to s_b' yields an analytical approximation for the boundary,

$$y(x) = \min_{s_b'|x} \left\{ \log \tilde{p}_{\text{fix}}(s_b, U_b \to s_b', U_b') \right\} \approx \log(x) - \frac{1}{4} \log^2(x) \left[\frac{x_c^2}{2v} - \frac{x_c s_b}{2v} \right] , \tag{S251}$$

which was used to define the gray region in Fig. 2D.

Figure 3. The theoretical predictions for Fig. 3 were obtained using a similar procedure as in Fig. 2. For a modifier with a direct cost ($s_m < 0$; Fig. 3A,C), we used the multiple-mutations expression in Eq. (S99) if $x_{cm} > s_b'$, and the quasi-sweeps expression in Eq. (S139) if $x_{cm} < s_b'$; the integrals in Eq. (S139) were computed numerically using the `quad` function from the SciPy library (107). For a modifier with a direct benefit ($s_m > 0$; Fig. 3B,C), we used the corresponding predictions in Eq. (S97) when $x_{cm} < x^*$, or the dead-end predictions in Eq. (S115) otherwise, with x^* defined by Eq. (S111).

The transition line in Fig. 3C was obtained by solving for the critical value of s_m that satisfies

$$\tilde{p}_{\text{fix}}(\mu \to \mu_m, s_m) = 1 \, . \tag{S252}$$

The theoretical line was obtained by numerically solving Eq. (S252) using the fixation probability predictions described above. The simulated values were calculated by linearly interpolating the observed fixation probabilities using the `polyfit` package in the Numpy library.

Figure 4. Predictions for the continuous distributions of fitness effects in Fig. 4 were obtained using the single-s mapping described in SI Section 5. We first calculated the value of x_c by numerically solving the wildtype version of Eq. (S249), with the effective parameters s_b and U_b defined by Eqs. (S151) and (S154). Using these estimates, we next checked whether the modifier was in the perturbative regime by numerically solving for δx_c using Eq. (S168). If $\delta x_c < v/s_b$, we used the corresponding predictions from the perturbative regime in Eq. (S174), with $s_b' \approx s_b$. If $\delta x_c > v/s_b$, we turned to the corresponding predictions from the modifier-dominated regime in SI Sections 5.3 and 5.4.

To do so, we numerically solved for x_{cm} using Eq. (S249) with s_b' and U_b' defined by Eqs. (S176) and (S179). For the bimodal distributions in Fig. 4, the solutions to Eqs. (S176) and (S179) can be expressed as a piecewise function,

$$s_b', U_b' = \begin{cases} s_b^\mu(x_{cm}), U_b^\mu(x_{cm}) & \text{if } U_b^\mu \tilde{p}_{\text{fix}}(s_b^\mu | x_{cm}, v) > U_1 \tilde{p}_{\text{fix}}(s_1 | x_{cm}, v), \\ s_1, U_1 & \text{else,} \end{cases} \tag{S253}$$

where $s_b^\mu(x_{cm})$ and $U_b^\mu(x_{cm})$ refer to the effective parameters of the wildtype distribution, but evaluated at x_{cm} rather than x_c. The fixation probability of the modifier was then estimated as

$$\tilde{p}_{\text{fix}}(\mu \to \mu_m; s_m) \approx \tilde{p}_{\text{fix}}(s_b, U_b \to s_b', U_b'; s_m) \, , \tag{S254}$$

where the right hand side was calculated using the methods described for Figs. 2 and 3 above.

To capture the transition region between the perturbative- and modifier-dominated regimes, we continued to use our perturbative approximation if the modifier-dominated estimate of $|x_{cm} - x_c|$ was less than the perturbative calculation of $|\delta x_c|$. This convention ensures that our theoretical predictions are continuous at the border between the two regimes.

Supplementary Figure S2. The theoretical predictions for Fig. S2B,C were obtained using Eq. (S73). The theoretical predictions for Fig. S2D,E were obtained using the procedure used in Fig. 2 for a beneficial modifier without a short term cost.

Supplementary Figure S3. The theoretical predictions for Fig. S3A,B were obtained using the same procedure in Fig. 2 for a beneficial modifier with no direct cost or benefit.

Supplementary Figure S4. The forward-time simulations with diminishing returns epistasis in Fig. S4 followed a similar procedure to that outlined in SI Section 8.1. To enhance reproducibility, the simulations were allowed to "burn-in" for $\Delta t = 2 \cdot 10^4$ generations before initiating diminishing returns. After this burn in period, diminishing returns was initiated and the mean fitness and typical selection coefficient, $s_b(\overline{X})$, were recorded every 50 generations. The simulation results in Fig. S4A,B are the mean and average of 10 simulations.

After the typical selection coefficient in the population had declined to the chosen value $s_b(\overline{X}) = 2.5 \cdot 10^{-2}$, the fitness and abundance of each lineage was saved. Modifier lineages were then introduced at a constant rate for $t_U = 1000$ generations. This value was capped to ensure a successful modifier arose in a population with the chosen selection coefficient. To generate the comparisons of the neutral modifier and background selection coefficients in Fig. S4C, $s_b(\overline{X} + x_c)$ was obtained using an estimate of x_c from the model without diminishing returns. This estimate was obtained using a similar procedure to that in Fig. 2 with v numerically evaluated in a population with a constant selection coefficient, $s_b = 2.5 \cdot 10^{-2}$. The results in Fig. S4C were then obtained using Eq. (S239).

After $t_U = 1000$, U_m was set to zero and the simulation continued until a modifier took over or all modifier lineages where purged. If a modifier did not take over, the population reverted back to the saved lineage distribution and this process was repeated n times until modifier took over. This allowed us to apply the same procedure in SI Section 8.1 to calculate the fixation probability of a modifier with $T = n \cdot t_U$. The theoretical predictions for the $\theta = \theta_m$ line were obtained using the same procedure in Fig. 2 for a selection strength modifier with $s_b = \tilde{s_b}$ and $s_b' = \tilde{s_b}'$. The theoretical predictions for the $\theta = \infty$ line were obtained using min{ Eq. (S242), Eq. (S135) }, where $s_{b,0} \cdot \tau_{\text{est},0}$ was calculated from the relation $s_{b,0} \cdot \tau_{\text{est},0} = s_{b,0}^2/v$. In this case, v was evaluated numerically in Wright Fisher simulations without diminishing returns for $s_b = s_b(\overline{X} + x_c)$.

8.3 Empirical example from Ref. (8)

The relative fitness estimates for the modifier example in Fig. 1A were obtained from Ref. (8). This study examined two strains of *E. coli* that were isolated from generation 500 of Lenski's long-term evolution experiment (42). Relative fitnesses of the two strains at the first timepoint in Fig. 1A were obtained from head-to-head competitions under the same conditions as the original experiment. Ref. (8) reported these relative fitness values using the metric,

$$W \equiv \frac{\log\left(\frac{N_2(\Delta t)\cdot 2^{\Delta t}}{N_2(0)}\right)}{\log\left(\frac{N_1(\Delta t)\cdot 2^{\Delta t}}{N_1(0)}\right)}, \tag{S255}$$

where $N_i(0)$ is the number of colonies of strain i observed at the beginning of the competition, $N_i(\Delta t)$ is the (adjusted) number of colonies observed at the end of the competition, and Δt is the length of the competition in generations. To enable more direct comparisons with our theory, we converted these W estimates to the

relative (log) fitness,

$$\Delta X \equiv \frac{1}{\Delta t} \log \left[\frac{f_2(\Delta t)}{f_1(\Delta t)} \cdot \frac{f_1(0)}{f_2(0)} \right] , \tag{S256}$$

where $f_i(t)$ is the relative frequency of strain i at generation t.

To perform this conversion, we assumed that the colony counts in Ref. (8) were adjusted so that a similar overall number of colonies were measured at both timepoints. This implies that $N_i(\Delta t)/N_i(0) = f_i(\Delta t)/f_i(0)$. If we also assume that the competition experiments were started at a 1:1 ratio, so that $f_1(0) \approx f_2(0) \approx 1/2$, we can obtain a relation between the ΔX and W metrics,

$$W = 1 + \frac{\Delta X \cdot \Delta t}{\log \left(\frac{2^{\Delta t+1}}{1+e^{\Delta X \cdot \Delta t}} \right)} , \tag{S257}$$

which depends on the duration Δt. We used this expression to convert the average and 95% confidence intervals reported in Ref. (8) into the relative fitness values in Fig. 1A:

Initial timepoint. The fitness measurements for the initial timepoint were conducted over $\Delta t = 47$ generations, and yielded a fitness disadvantage of W=0.937 (0.934,0.941) [mean and 95% confidence intervals from replicate competition assays]. Numerical conversion using Eq. (S257) yielded a log fitness difference of ΔX= -4.4% (-4.17%,-4.66%).

Second timepoint. The fitness measurements at the second timepoint were obtained after evolving each isolate for an additional 883 generations in 20 independent replay experiments. Pooled fitness measurements of the evolved populations were performed over $\Delta t = 6.6$ generations, and yielded a fitness benefit of W=1.02 (1.003, 1.039). Numerical conversion using Eq. (S257) yielded a log fitness difference of ΔX=1.4% (0.2%, 2.6%).