Evolutionarily stable levels of aposematic defence in prey populations

*Alan Scaramangas[1], Mark Broom[1], Graeme D Ruxton[2], Anna Rouviere[2]

[1]*School of Mathematics, Computer Science and Engineering, City, University of London, EC1V, 0HB London, UK.*
[2]*School of Biology, University of St Andrews KY16 9TH, St Andrews, UK*

February 2023

Abstract

Our understanding of aposematism (the conspicuous signalling of a defence for the deterrence of predators) has advanced notably since its first observation in the late nineteenth century. Indeed, it extends the scope of a well-established game-theoretical model of this very same process both from the analytical standpoint (by considering regimes of varying background mortality and colony size) and from the practical standpoint (by assessing its efficacy and limitations in predicting the evolution of prey traits in finite simulated populations). The nature of the manuscript at hand is more mathematical and it's aim is two-fold: first, to determine the relationship between evolutionarily stable levels of defence and signal strength under various regimes of background mortality and colony size. Second, to compare these predictions with simulations of finite prey populations that are subject to random local mutation. We compare the roles of absolute resident fitness, mutant fitness and stochasticity in the evolution of prey traits and discuss the importance of population size in the above.

Keywords: aposematism; ESS; chemical defence; numerical simulation.

1 The biology of aposematic defence

Aposematism is the signalling by prey individuals (animals or plants) to predators that they are unprofitable to consume. A succinct description of the process is provided in Wallace (1877), p. 651: *"...warning colours– are exceedingly interesting, because the object and effect of these is, not to conceal the object, but to make it conspicuous. To these creatures it is useful to be seen and recognised, the reason being that they have a means of defence which, if known, will prevent their enemies from attacking them, though it is generally not sufficient to save their lives if they are actually attacked."* Indeed, these warning colourations or signals are associated with some form of prey defence and may be manifest in a wide range of physical characteristics perceivable to the predator through sensory stimuli beyond just sight, but which may also include smell, touch, taste or combinations of these. The term *aposematism* was coined in Poulton (1890) and literally means *to keep predators at a distance*; it stems from the Greek words *apostasis* (meaning *distance*) and *sema* (*signal*). For a very brief, up-to-date description of aposematism a good first option is Rojas et al. (2021).

The observation of aposematism in the natural world would seem troubling from the evolutionary standpoint as it is sensible to surmise that conspicuous individuals run a clear disadvantage compared with their non-signalling relatives. Several hypotheses have been developed to explain its origins and maintenance, and biologists have provided insight into the physiological functioning of even the more intricate mechanisms of defence (see Ruxton et al., 2019 and Mappes et al., 2005 and the bibliographies therein for a succinct accounting of a vast body of research). These questions are of considerable importance but are beyond the scope of the manuscript at hand, in which we strive to address three questions perhaps less acknowledged in the current literature: (a) In a large enough (effectively infinite) population of prey is there a certain manner in which defence should be advertised to make the population more likely to retain its composition over successive generations and under the presence of mutation? (b) how might our answer in (a) change under different regimes of background levels of mortality? (c) how might our answers in (a) and (b) be adapted to a population of prey that is finite but large enough that its traits are not fully driven by stochasticity?

As first documented in Wallace (1877) and in the works of almost all naturalists and behavioural ecologists who have explored it since, aposematism is present in a rather broad range of taxa (although it is remarked in Santos et al., 2003, Vences et al., 2003 and Ruxton et al., 2019 that in most taxonomic groupings aposematic species are rare compared with their camouflaged counterparts). Perhaps for this reason, theoretical study of aposematism has involved an impressive number of approaches (see Summers et al., 2015 for a succinct cataloguing of these). Notice that our aim **here** is to explore the evolution of aposematic traits after they have become established in a population. We do not address the important but separate question of how aposematic traits initially evolve (see Mappes et al., 2005 for an overview of that topic).

In this manuscript we study the game-theoretical treatment of aposematism due to Broom et al. (2006), which builds on that of Leimar et al. (1986) and arguably constitutes one of the more inclusive descriptions of aposematism to date. The reader is encouraged to consult the original paper for a more detailed comparison with Leimar et al. (1986) and preceding descriptions, Broom et al. (2008) for a detailed demonstration of its direct implications and Scaramangas and Broom (2022) for a systematic review of the model's theoretical components and an exploration into some of **its** less anticipated functionings. Much like in Broom et al. (2006) and in several of the subsequent publications listed above, we introduce the background biology of the model with emphasis on the theory of defence (the reader may consult Ruxton et al., 2019 or Caro, 2005 for a concise survey of this topic from a more natural history perspective).

Defences may either be permanently present in the individuals that acquire them (***static/constitutive***) or deployed during conflict (*induced*). **The triggering of the latter typically requires time, during which predator attacks may be successful. For this reason induced defences tend to be most successful against attacks that are slow-paced and least successful against attacks that happen fast. The model of Broom et al. (2006) and which we present here assumes that predator attacks are mounted at a fast pace and that defences are not induced during conflict but are permanently present in the prey that deploy them.** Beyond the differences in the **time-scales at which they are deployed**, defences may be characterised (more fundamentally perhaps) in terms of their purpose as *primary* or *secondary*. Primary defences aim at reducing the frequency of predator encounters and are synonymous with either *camouflage* (prey remain invisible to the predator by resembling their background appearance) or *masquerade* (prey remain visible but inconspicuous by resembling objects that are perceived by the predator as neutral, such as leaves, twigs or bird droppings). On the other hand, secondary defences aim at reducing the chance that mounted attacks are lethal - see Ruxton et al. (2019), Broom et al. (2006) and the references therein. Secondary defences are widespread and heterogeneous and can be classified as *locomotor* (rapid escape, protean evasive flight), *morphological* (sharp spines, thick shells, etc.) and *chemical* (toxins, venoms, noxious secretions, etc.).

Broom et al. (2006) and subsequent manuscripts including this one are better suited in describing organisms that are chemically defended with internally stored toxins (these become apparent to the predator only after an attack has been attempted). Although some organisms possess defences (e.g. sharp spines) that are visually detectable by the predator at a distance and reliably signal unprofitability, chemically defended types of the sort that we consider require separate conspicuous signalling to achieve this effect. Poisonous frogs from the *Dendrobatidae* family exhibit brightly-coloured skin pigmentation with oftentimes impressively contrasting patterns and provide a good example of a system that closely matches the model description of Broom et al. (2006) - see Scaramangas and Broom (2022) for a more detailed discussion of this.

We establish that the process of warning colouration may effectively be phrased as "aposematic prey sacrifice their primary defence in favour of a signalling appearance that (i) signals to predators the presence of secondary defences and which in turn (ii) acts as a deterrent and hence a substituted form of primary defence." It is argued that both predators and prey can benefit from honest signalling of chemical defences if there are costs to both parties associated with prey capture prior to detection of defences (such as time and energy invested in chasing and fleeing, and/or risk of injury).

In the section that follows we introduce the game-theoretical model of aposematism as this was originally described in Broom et al. (2006). The reader is directed to Appendix I for a detailed comparison of the mentioned model and subsequent extensions (including this one). It should be remarked that other models including that of Leimar et al. (1986) have been crucial in refining our mathematical understanding of aposematism and we direct the reader to Broom et al. (2006) for an elaborate comparison of their interrelatedness. The theoretical foundations of the model are developed further in Scaramangas and Broom (2022) and also presently, where our theoretical analysis is complementary to novel numerical simulations.

2 A game-theoretic model for aposematism

In this section we describe a game-theoretical model for aposematism as first introduced in Broom et al. (2006). Following a general presentation in terms of prey strategies that are unrestricted we adopt a resident-mutant setup and discuss what it means for a resident strategy to be *locally uninvadable* or *evolutionarily stable*. These characterisations of stability are closely interwoven with the notion of *fitness* or *payoff*, which specifies the success of a certain type in retaining its traits (through mitigating predation without sacrificing reproduction) under the presence of predation and local mutation. Related to this, and of general interest is the discussion in Ruxton et al. (2009) on the trade-offs associated with aposematic strategies.

2.1 General description

We consider a habitat occupied by N individuals of a certain species that are potential prey to some predator population of size n. Prey individuals (indexed $i = 1, ..., N$) are aposematic and described in terms of their *conspicuousness* r_i and their *toxicity* t_i, both of which are real-valued and non-negative. Individuals with $r_i = 0$ are understood to be *maximally cryptic* (in the sense of *"close resemblance to a random sample of the background"* discussed in Endler, 1978), while individuals with $t_i = 0$ are completely *undefended.* Alongside prey resides a group of n predators who visit their habitat and make decisions on whether to attack based on some background level of *perceived aversiveness.*

There is ample evidence that investment in anti-predatory defences is costly (specific to *Poison dart frogs* is the discussion in Tarvin et al., 2017 among many others). We therefore assume that the *fecundity* or reproduction rate $F = F(t_i)$ is a declining function of t_i. Prey mortality can be attributed to either natural causes (we assume this happens with some fixed background mortality rate λ) or due to predation. Toward the latter, we imagine that predators encounter prey at some fixed rate and that their subsequent interaction is realised in the following way. Upon encounter, prey may or may not be detected, upon detection prey may or may not be attacked and finally a mounted attack may or may not be lethal depending on how defended the animal is.

It has been mentioned that the model of Broom et al. (2006) is static and does not impose specific dynamics to describe prey-predator interactions. We surmise that on average, predators encounter prey at some fixed rate σ. Detection of individual i is an event that is conditional on the predator encountering that prey such that the rate of detection $D(r_i)$ can be defined as the product of this mentioned (fixed) rate of encounter and the probability that i is detected given encounter has occurred. We write

$$\boldsymbol{D(r_i)} = \text{Rate of detection of } \ i = \sigma \times \mathfrak{P}(i \ \text{ is detected} \,|\, i \ \text{ is encountered}) \tag{1}$$

and assume that the detection rate $\boldsymbol{D(r_i)}$ is an increasing function of r_i, which tends to unity as prey conspicuousness assumes arbitrarily large values. **In addition, we assume that** $D(0) = d_0 > 0$, suggesting that even fully-cryptic prey can be detected. The probability that a mounted attack results in capture is given by $K = K(t_i)$, where K is declining with t_i, indicating that better defended individuals are harder to capture. A detected individual $\boldsymbol{i}$ is attacked with probability $Q = Q(I_i)$, where $\boldsymbol{I_i} \in \mathbb{R}$ represents the *average aversive information* that the average predator has on **that individual (so that the smaller/larger this value is the more attractive/aversive the individual is perceived as being). It is assumed that $\boldsymbol{Q}$ is a declining function of $\boldsymbol{I_i}$ so that the more/less aversive it is the less/more likely it is that it will be attacked. If individual $\boldsymbol{i}$ is perceived as very attractive it will almost certainly be attacked and we articulate this through the condition that $\boldsymbol{Q(I_i) \to 1^-}$ in the limit as $\boldsymbol{I_i \to -\infty}$. Conversely, if $\boldsymbol{i}$ is perceived as very aversive it will almost certainly not be attacked, so we write $\boldsymbol{Q(I_i) \to 0^+}$ in the limit as $\boldsymbol{I_i \to +\infty}$. In the special case that prey $\boldsymbol{i}$ is neutral it may or may not be attacked and this occurs at random with probability $\boldsymbol{Q(0) \in [0,1]}$.**

Predators assign I_i to individual i by comparing it to a certain (weighted) base-line level of aversive information, which is generated through **previous** encounters with the prey population, such that

$$I_i = \frac{1}{n} \sum_{j=1, j \neq i}^{N} L(r_j) H(t_j) \boldsymbol{S}(|\boldsymbol{r_i} - \boldsymbol{r_j}|). \tag{2}$$

We now introduce the terms present in the sum on the RHS of (2). Predators find chemically-defended prey aversive and the experience of consuming them is measured by $H = H(t_i)$, which is an increasing function of t_i and is zeroed at a critical value of the toxicity $t_i = t_c$. We write

$$H(t_i) \begin{cases} < 0, & t_i < t_c \\ = 0, & t_i = t_c \\ > 0, & t_i > t_c \end{cases} \tag{3}$$

and emphasize that this is an honest measure of the distastefulness of an experience as opposed its attractiveness. The level of defence of prey with $t_i < t_c$ is not sufficient to outweigh the nutritional benefit received from predators by consuming them and **such prey** are perceived as attractive or *negatively aversive*. By construction, the defence of prey with $t_i = t_c$ describes the limiting value at which the nutritional benefit is exactly outweighed by their distastefulness and such prey are perceived as *neutrally aversive*. Finally, prey with $t_i > t_c$ are perceived as *(positively) aversive* by the potential predator. The second term on the RHS of (2) requiring explanation is $L = L(r_i)$, which represents the rate at which encounters that have occurred can be recalled by predators. This is assumed to be a growing function of r_i indicating that encounters with more conspicuous prey can be better recalled. In much of the later work we will assume *perfect predator recollection*, which involves taking taking $L = D$.

The third term that warrants explanation on the RHS of (2) is the *similarity function* $\boldsymbol{S}$ **that describes** how a predator perceives/compares the visual appearance of individuals i and j differing in conspicuousness by amount $x = |r_i - r_j|$. The similarity function $x \mapsto S(x)$ is of class $\mathcal{C}^l$ with $l \geq 2$ (at least sufficiently near the origin) and has the following additional properties. i) $S(x) \in [0, 1]$ for all $x \geq 0$, namely that the perceived similarity of two individuals is assigned some real value between zero and unity. ii) $S(0) = 1$, which suggests that if individuals i and j share the same levels of conspicuousness the function S is evaluated at $x = 0$ where it assumes the value one. iii) $S(x) \to 0^+$ in the limit as $x \to +\infty$, suggesting that if i and j have vastly different levels of conspicuousness (i.e. $0 < r_i \ll r_j$ or $0 < r_j \ll r_i$) the similarity function is evaluated far from the origin where its value approaches zero. iv) $S'(x) \leq 0$ for all $x > 0$, namely that the similarity function is almost everywhere non-increasing except at $x = 0$ where v) $S'(0) < 0$. The reader is encouraged to consult Appendix IC for an in-depth discussion about properties iv) and v) and their association with the underlying predator psychology.

Having introduced L, H and S it remains for us to clarify those aspects of the prey-predator interaction that are implicit in the form of (2). We remark that while predator learning does feature strongly in our model it is not a process that we explicitly describe. That is, the model can best be thought to describe a composition of mostly educated predators who mount attacks on prey based on knowledge of prey traits gathered during an early, short and investigative period of their life. We also assume that prey reside in some extended habitat that is *territorially-divided* among the predators who occupy this: We imagine that the habitat is partitioned into (potentially large) geographical localities/sites so that each site is occupied by N prey (where N can be taken to be large) and visited by n predators, who visit one site only.

In the setting described above, it would be unnatural (subject to the relative prey/predator abundance not being exceptionally high/low - see below) to assume that individual predators (including insectivorous birds of the type discussed in Scaramangas and Broom, 2022) have had encounters with each prey individually (this may be due to the large turnover of prey and/or the vastness of the sites they occupy). Rather, we can conceive that the level of aversiveness of i corresponds to collectively generated experiences with its neighbours (i.e. by the entire group of predators) and is captured by the summation term on the RHS of (2), whose index j runs through $\{1, ..., N\} \setminus \{i\}$. In keeping with a description that is prey-focused we refrain from imposing specific restrictions on the distribution of knowledge among predators. Hence, the sum on the RHS of (2) is divided through by n so that the LHS (i.e. I_i) can be thought to describe the information on i corresponding to an individual predator in an idealised scenario in which collectively generated knowledge on i is shared equally. This is an assumption that is not unreasonable for avian predators whose nested family structures can allow for strong cultural transmission of foraging behaviour - see Mappes et al. (2005).

We establish that (2) describes learning as an averaging process that is static and in which prey occupy distinct geographical sites that are vast (even though the model does not feature an explicit spatial component to account for this). While this enables us to study aposematism from the point of view of evolutionary stability to considerable depth, the plethora of scenarios that this can describe can be limited by the relative prey/predator abundance. For instance, applying (2) to an exceptionally large prey population (i.e. with $N/n \to +\infty$) could yield an artificially large value for I_i (the sum on the RHS of (2) could increase without bound). It is unlikely, however, that such a situation would call for (2) since the assumption that predators have complete experience of i's neighbours may not be valid. Conversely, situations involving

an exceptionally large number of predators such that $N/n \to 0^+$ would result in an artificially low value of I_i (by averaging a finite quantity over a large population). Again the use of (2) **would be less relevant in such examples because direct experience of i is likelier than assuming knowledge of i indirectly through its neighbours. In the simulations discussed below it is important to note that we do not consider regimes with $N/n \to +\infty$ or $N/n \to 0^+$.**

By construction, attack is conditional on detection and capture is conditional on attack, therefore the *Law of Total Probability* suggests that the *predator-induced mortality rate* of i is evaluated as the product $P(Capture|Attack) \times P(Attack|Detection) \times (Rate\,of\,detection\,) = K(t_i)Q(I_i)D(r_i)$. Prey can also die from causes other than predation (in fact this component is central to our analysis is the next section) and therefore the total mortality rate of i is $\lambda + D(r_i)K(t_i)Q(I_i)$. **We emphasize that the latter specifies a rate such that its reciprocal $(1/(\lambda + D(r_i)K(t_i)Q(I_i)))$ has units of time and describes the average time taken for individual i to perish. Identifying this quantity as the (average) life-cycle for i and accounting for fecundity $F(t_i)$, we infer that the (average) number of offspring i produces per life-cycle is given by**

$$P(r_i, t_i) = \frac{F(t_i)}{\lambda + D(r_i)K(t_i)Q(I_i)} \tag{4}$$

and defines the *payoff* or *fitness* of individual i. This measure of fitness was introduced in Broom et al. (2006) and has also been used in subsequent works including those of Broom et al. (2008), and Scaramangas and Broom (2022). From definition (4) **it follows that high fitness individuals are distinguished as having longer life-cycles (by effectively mitigating predation) and/or producing more offspring in that time. As we discuss, there is a complex trade-off between selecting for one or the other. We are now in a position to describe the** evolutionary stability **of aposematism and do so** in the context of **a** resident-mutant setup of the prey population. Below, we provide a list of symbols and their meaning.

Symbol	Meaning
r	the conspicuousness of a prey individual
t	the toxicity of a prey individual
N	the size of the prey population
n	the size of the predator population
$D(r)$	the rate at which r-individuals are detected
$L(r)$	the rate at which r-individuals are detected and recalled
$S(x)$	the similarity function of individuals differing in conspicuousness by x
$H(t)$	the aversiveness of prey individuals with toxicity t
t_c	the critical level of toxicity such that $H(t_c) = 0$
$F(t)$	the fecundity of a prey individual with toxicity t
$K(t)$	the probability that an attacked t-individual is captured
$Q(I)$	the probability that a detected I-individual is attacked
I	the level of aversive information of an individual
λ	the prey background mortality rate (not due to predation)
a	the average relatedness of prey individuals in the population

Table 1: The parameters and functions of the model.

2.2 Resident-mutant description

In this subsection we explain how the model of Broom et al. (2006) can be used to determine when aposematic populations of prey can maintain their traits indefinitely in the absence of genetic drift and under the presence of local mutation. We consider a *resident-mutant* setup in which the prey population is made up of a majority playing some *resident strategy* (r_1, t_1) and a small minority of mutants playing a nearby strategy $(r, t) \in (r_1 - \delta r, r_1 + \delta r) \times (t_1 - \delta t, t_1 + \delta t)$. In this manuscript we deal exclusively with mutations of this type, which we identify as *local* (mutant traits are drawn from the local vicinity of the resident value) and best describe **heritable** random mutations that can arise in the parent genome. As described in Scaramangas and Broom (2022) and more briefly in Appendix I most habitat sites are occupied by residents, except for a small number containing mutants in local proportion a, which is known as the *average local relatedness.* Implicit in the description is that prey breed true so that (in the local area) there is proportion a of *exact copies*, with the remaining proportion being unrelated to the group.

Expression (2) can be used to determine the perceived aversiveness (see Appendix I) of the mutant type playing (r, t) in a site consisting of (local) proportion a of mutants as

$$I(r, t; r_1, t_1) = a\frac{N}{n}L(r)H(t) + (1-a)\frac{N}{n}L(r_1)H(t_1)S(|r - r_1|), \tag{5}$$

which we abbreviate to I. The majority of the habitat is assumed to be occupied by prey playing the resident strategy, whose perceived aversiveness (evaluated over the larger area) can be deduced from (2) as

$$I_1(r_1, t_1) = \frac{N}{n}L(r_1)H(t_1). \tag{6}$$

For purposes of consistency we remark that the mutant aversiveness of (5) coincides with the resident aversiveness provided the former is evaluated at the resident value so that $I(r = r_1, t = t_1; r_1, t_1) = I_1(r_1, t_1)$ Naturally, this result extends to the mutant fitness which on account of (4) reads

$$P(r, t; r_1, t_1) = \frac{F(t)}{\lambda + D(r)K(t)Q(I)} \tag{7}$$

and for the resident this is

$$P_1(r_1, t_1) = \frac{F(t_1)}{\lambda + D(r_1)K(t_1)Q(I_1)}. \tag{8}$$

We tend to use subscript $_1$ for resident-related quantities and tend to omit the resident traits in the arguments of mutant-related quantities. For instance, we write $P(r, t)$ for the mutant fitness and $\partial_r P(r_1, t_1)$ for the rate of change of the mutant fitness with respect to the mutant conspicuousness evaluated at $(r, t) = (r_1, t_1)$. We should also point out that for all **intents and purposes** (see (10)) the forms used to represent the functions F, D, K, Q, H and L are all of class $\mathcal{C}^l$ with $l \geq 2$ (this also applies to S, provided that this is evaluated sufficiently near the origin). This *smoothness* restriction mostly agrees with our general perception of the physical world; namely that organisms playing slightly different strategies tend to have very similar values for the different consequences of their strategies.

A resident strategy (r_1, t_1) is (locally) evolutionarily stable - local ESS - if it is best-response against itself and in particular, if no mutant strategy (r, t) can receive higher fitness when interacting with the ESS strategy than can the ESS strategy when interacting with itself. It follows immediately from the definition that a resident strategy (r_1, t_1) is **a** local ESS if it is a maximum of the mutant fitness defined on the infinitesimal rectangle $(r, t) \in (r_1 - \delta r, r_1 + \delta r) \times (t_1 - \delta t, t_1 + \delta t)$ centred at (r_1, t_1). We refer the reader to the related discussions in Broom et al. (2006) and Scaramangas and Broom (2022).

The precise conditions for a local maximum depend on where on the boundary-inclusive, right-upperhand plane $\{(\rho, \tau) : \rho \geq 0, \tau \geq 0\}$ the resident strategy is evaluated. We should clarify that our use of the generic variables ρ and τ for the conspicuousness and the defence are used exclusively to identify different subregions in the strategy space. We distinguish between the origin $\{(0,0)\}$, the boundaries $\{(\rho, 0) : \rho \geq 0\}$, $\{(0, \tau) : \tau \geq 0\}$ and the interior regions of the strategy space $\{(\rho, \tau) : \rho > 0, \tau > 0\}$. For the latter, it is shown in Broom et al. (2006) that the non-aversive subset of the interior $\{(\rho, \tau) : \rho > 0, \tau \leq t_c\}$ does not contain local ESSs and these regions are represented in grey-scale in Figures 2(a), 3(a) and 4**(a)**. The significance of this result is also explained in Appendix I.

For clarity, we mention here that on the interior subregion of the strategy space (which occupies the largest portion of our analysis) $\{(\rho, \tau) : \rho > 0, \tau > t_c\} \subset \{(\rho, \tau) : \rho > 0, \tau > 0\}$, the conditions for local ESS read

$$\partial_t P(r_1, t_1) = 0, \quad \partial_{tt} P(r_1, t_1) < 0, \quad \overleftarrow{\partial}_r P(r_1, t_1) > 0 \text{ and } \overrightarrow{\partial}_r P(r_1, t_1) < 0, \tag{9}$$

with specific definitions of the terms involved provided in Appendix I. As is clear from the discussion thus far, there is much freedom with respect to the functional forms**,** and as demonstrated in Scaramangas and Broom (2022) such choices can showcase different aspects of the model. The example functions used previously in Broom et al. (2008) are biologically plausible, sufficiently well-behaved and are the natural choice for use in the **simulation** model. These read

$$F(t) := f_0 e^{-ft}; \quad H(t) := t - t_c; \quad K(t) := \frac{k_0}{1 + kt}$$

$$L(r) = D(r) = \frac{1}{1 + e^{-r_1}}; \quad Q(I) := min\left(1, q_0 e^{-qI}\right); \quad S(x) = max(1 - vx, 0) \tag{10}$$

and are shown in Figure 1 below. We emphasise that while previous works including Broom et al., (2008) have considered (10) only alongside $\lambda = 0$ we presently account for scenarios with $\lambda = 0$ and $\lambda \neq 0$ alongside zero and non-zero levels of local clustering.

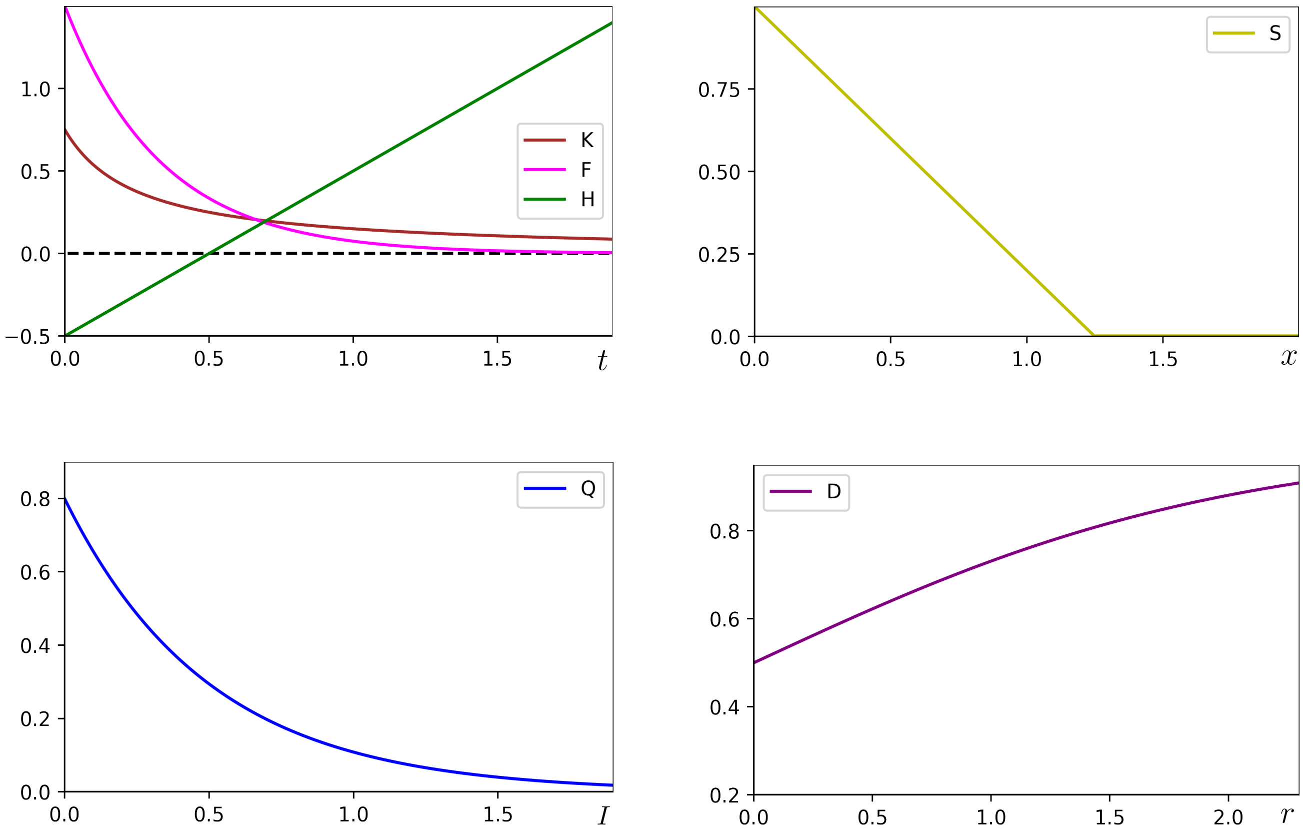

Figure 1: The example functions of (10) as used in the simulation model with specific parameter values chosen for purposes of demonstration only. (a) [Top left] The functional forms for the probability of escaping a mounted attack (K), the fecundity (F) and the aversiveness of an experience (H) plotted as functions of prey toxicity. Parameter values given as $k_0 = 0.75$ and $k = 4$; $f_0 = 1.5$ and $f = 3$ and $t_c = 0.5$ respectively. (b) [Top right] The functional form for the (uni-variate) similarity function S plotted with respect to the generic variable x and parameter $v = 0.8$. (c) [Bottom left] The form for the probability of attack Q with $q_0 = 0.8$ and $q = 2$ plotted as a function of the perceived aversiveness; (d) [Bottom right] The form for the rate of detection plotted as a function of prey conspicuousness D with $d_0 = 0.5$.

3 Evolutionarily stable outcomes

In this section we re-visit the example functions in Broom et al., (2008) - see (10) - and introduce non-zero level of background mortality alongside non-zero levels of local relatedness (a circumstance not previously explored). In Appendix I it is demonstrated that the relationship between predicted levels of aposematic defence and conspicuousness at ESS depends strongly on the levels background mortality. This is an important result that we recover presently in the context of the **simulation** model and which confirms a number of intuitive principles about the functioning of aposematic defence. The numerical results are showcased with situations of zero and non-zero rates of background mortality featuring in different subsections and with the simpler instances associated with zero levels of the local relatedness considered separately therein.

3.1 Solutions without background mortality $\lambda = 0$

In this subsection we consider the simplest scenario in which $\lambda = 0$ and treat the cases $a = 0$ and $a > 0$ separately. We make use of the theory developed **in** Appendix I and focus our attention on (i) the predicted form of **the** ESS, (ii) the resident fitness at equilibrium and (iii) **the** invasion fitness gradient (along r). This style of presentation exposes the reader to gradually increasing levels of complexity and is also adopted in the subsection following this, which deals with the case $\lambda > 0$. We should also remind the reader that the theoretical/predictions component of this section is based on the existing works of Broom et al. (2006) and Broom et al. (2008).

The $a \to 0$ limit

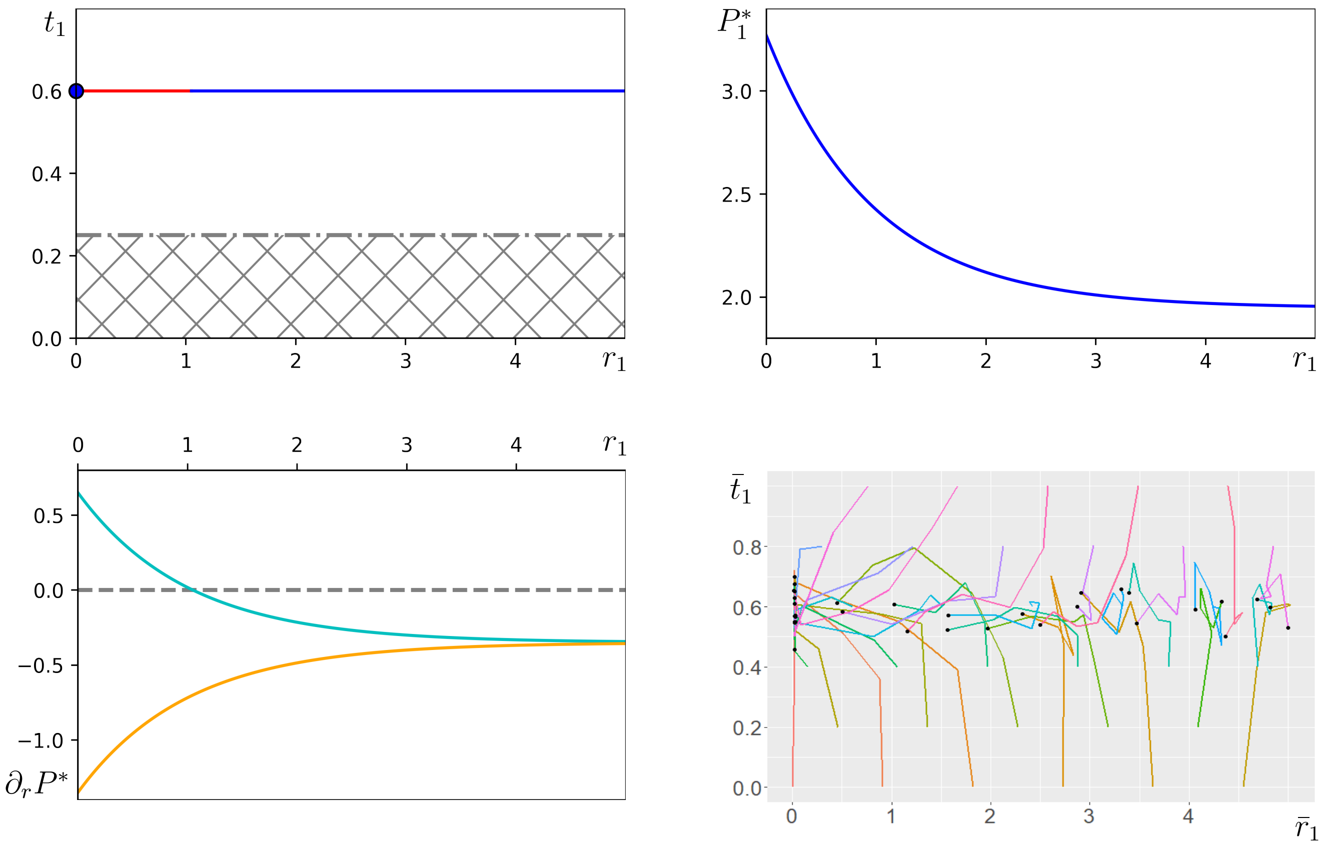

Figure 2: Parameter values $\lambda = 0, a = 0, f_0 = q_0 = k_0 = 1, f = 1, k = 2.5, t_c = 0.25, q = 0.1, N = 100, n = 10$ and $v = 1$ **(a)** [Top left] Strategies within the grey-shaded region $\{(r_1, t_1) : r_1 > 0, t_1 \leq 0.25\}$ risk invasion from less conspicuous mutants - see related discussion in Appendix I. Unique cryptic ESS at $(r_1, t_1) = (0, 0.6)$ shown with blue marker, succeeded by a line of equilibria that are unstable for $r_1 < 1.05$ and stable beyond that (blue section). **(b)** [Top right] Resident fitness evaluated at equilibrium as per (8). **(c)** [Bottom left] Invasion fitness gradient along r evaluated at equilibrium **- see** (42) **as well as** (38) **and** (39) **-** for incrementally less **(cyan curve)** and incrementally more conspicuous mutants **(orange curve)**. **(d)** [Bottom right] Average population traits plotted as trajectories with averaging frequency $g = 2,000$. Black markers represent the average traits of a single population after $10,000$ iterations.

The black markers in Figures 2(d) and 3(d) indicate that the majority of prey populations eventually converge close to the predicted equilibrium toxicity level of $1/f - 1/k$ as given in (30). We also remark that the lower the initial conspicuousness of the population the stronger the component of its associated trajectory toward crypsis. In Figure 2(d) this would be expected for initial conspicuousness values below the cut-off specified through (38) but we observe that even populations starting from evolutionarily stable strategies are invaded by less conspicuous types. In observing Figure 2 alone one could speculate that this is attributed to resident fitness being higher at crypsis. However, from Figure 3 we deduce that this is unlikely the case, since there populations evolve against increasing resident fitness and toward crypsis where the less conspicuous mutants are increasingly advantageous. **Presently, we make a number of important remarks about the invasion fitness gradient, which we use throughout to interpret the results of simulations.**

Following this, we discuss resident fitness and compare its impact on the evolution of prey traits alongside invasion fitness.

As it happens, plots of the invasion fitness gradient (orange and cyan curves in Figures 2c, 3c as well as in 4c and 5c) are consistent with the mutant landscape in the vicinity of the resident value along r. That is, when the cyan/orange curve is above the r-axis a resident population with that level of conspicuousness is predicted to be invasible by less/more conspicuous types (see cyan curves for $\bar{r}_1 < 1$ in Figures 2c as well as for $\bar{r}_1 < 0.5$ in 4c and the observed pull toward crypsis in 2d and 4d). In the majority of the cases we explore both the cyan and orange curves sit below the $r = 0$ axis (infinite population ESS analysis would deem such cases as evolutionarily stable along r) and the height below which they do so indicates how "worse-off" mutation in that direction is. An interesting effect of finiteness of the prey population is that strategies predicted as being evolutionarily stable (along r) still have a chance of being invaded if alternative mutations are not too worse-off. For instance, strategies with initial conspicuousness $1 < \bar{r}_1 < 3$ in Figure 2(d) and $\bar{r}_1 < 3$ in Figure 3 that are still invaded (through chance) by less conspicuous mutants.

For high enough levels of conspicuousness it is observed that the orange and cyan curves in Figures 2(c) and 3(c) converge (horizontally) to a common value. Technically, this can be attributed to our chosen forms for D (and therefore L - see (10)**), which exhibit a plateau for high enough levels of conspicuousness (already bright signals do not impact detection/recollection further). In such situations mutation in either direction leads to equally bad outcomes suggesting there is no directional selection associated with the invasion fitness gradient along r. The further below the r-axis the asymptote is reached the more worse-off mutations are predicted to be so that not only is invasion equally likely in either direction, the probability of this occurring shrinks. Indeed, from a quick reading at $\bar{r}_1 > 3$ it is clear that the trajectories in Figure 3(d) appear less incidental than in Figure 2(d), where the associated asymptote is above -0.5 (compared with -3 in Figure 3c). We conclude that the smaller the distance between the cyan and orange curves the smaller the difference in selection between either direction and the smaller the value that these converge to the more unlikely invasion (in either direction) is overall.**

While the simulations in Figures 2 and 3 fall under the same regime with respect to local clustering and background mortality ($a = 0$ and $\lambda = 0$) these **show** two principal differences, whose impact we explore further. **The first difference is with respect to the invasion fitness gradient: In Figure 2, overall selection for smaller conspicuousness is strongest (and manifest as a stronger pull toward crypsis) and when selection is absent (high $\bar{r}_1$) randomness (seen in the time evolution of trajectories) is higher because invasion is likelier (though equally so in either direction). In addition,** we have concluded that identifying strategies as "stable" or "unstable" is of limited use when studying prey populations that are finite, unless these are complemented with more precise statements describing "how stable"/"how unstable" those strategies are.

The second difference is with respect to the resident fitness at equilibrium. **Viewing Figure 2 it is difficult to set aside the impact of absolute resident fitness because this is highest for low $\bar{r}_1$ where (mutant fitness led) selection for less conspicuous types is also strongest. However, we observe that reversing the direction of increase of absolute resident fitness (Figure 3) does not significantly affect the outcome of the simulations. For sufficiently high values of $\bar{r}_1$ (where directional selection associated with the invasion fitness gradient is low) we could have expected prey trajectories in Figure 3(d) to evolve in the direction of increasing conspicuousness. Instead, these appear to evolve in mostly a random fashion and we conclude that this measure**

of fitness has little effect on the evolution of prey traits. This could be because under low local relatedness ($\boldsymbol{a = 0}$), resident fitness does not predict mutant fitness (which is the quantity determining the direction of evolution).

In particular, through the examples in Figures 2 and 3 we establish three important facts relating to the evolution of traits in finite prey populations: (i) ESS analysis provides accurate insight into the behaviour of finite populations even though notions of stability are less deterministic. (ii) Mutant fitness along r appears to be the stronger driver of changes in prey traits compared to the resident fitness. In fact, the probability of invasion along the r-direction depends continuously on how worse-off the mutant type is compared with the resident, as opposed to some absolute rule describing stability. (iii) In **the** absence of directional **selection** associated with incremental increases in mutant fitness (along r) and/or absolute resident fitness the evolution of traits is mostly random.

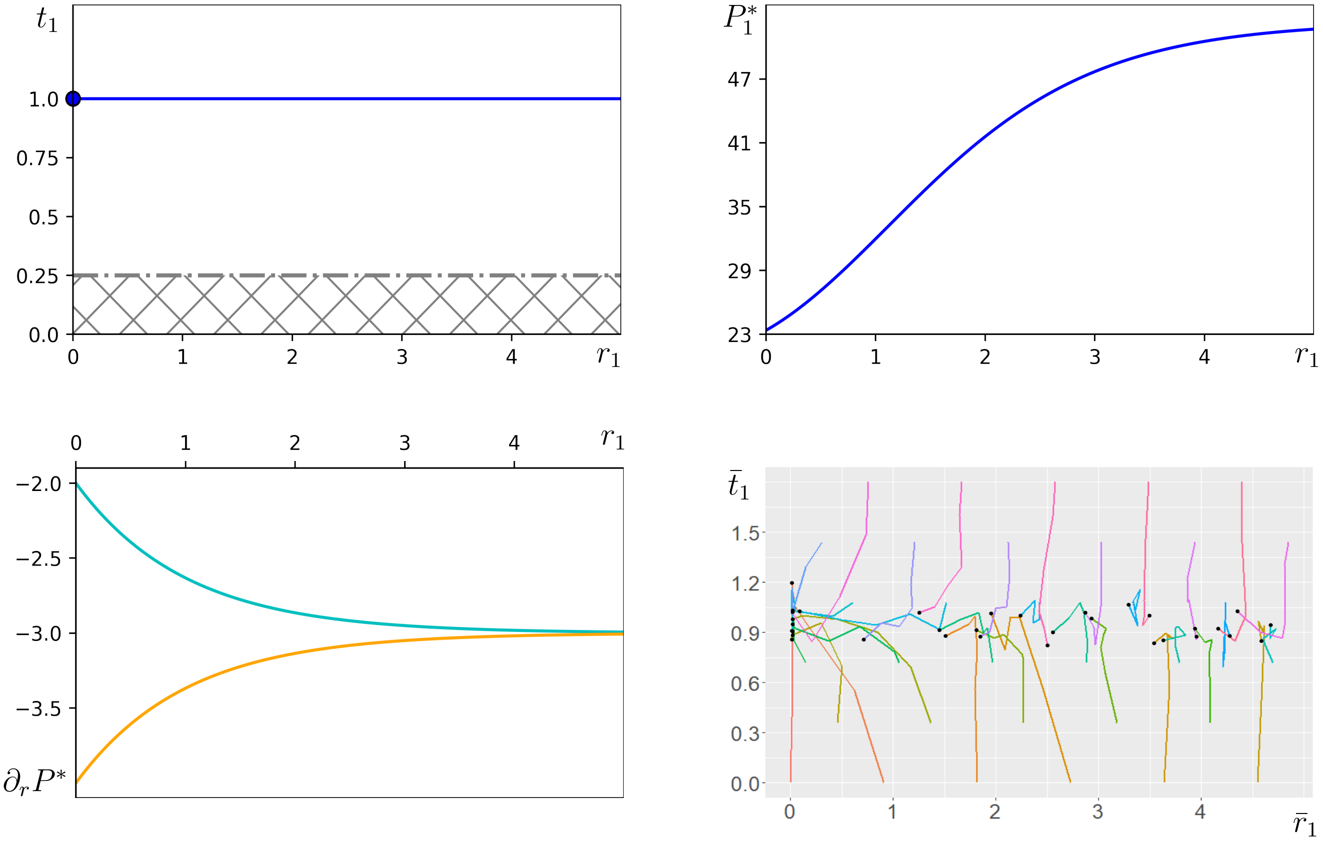

Figure 3: Parameter values $\lambda = 0, a = 0, f_0 = q_0 = k_0 = 1, f = 5/6, k = 5, t_c = 0.25, q = 0.4, N = 100, n = 10$ and $v = 1$ **(a)** [Top left] Grey-scale region $\{(r_1, t_1) : r_1 > 0, t_1 \leq 0.25\}$ contains unstable strategies. Solid blue marker at $(0, 1)$ is the unique cryptic ESS followed by a horizontal blue line of conspicuous ESSs. **(b)** [Top right] Plot of resident fitness at equilibrium. **(c)** [Bottom left] Invasion fitness gradient along r evaluated at equilibrium **- see** (42) **as well as** (38) **and** (39) **-** for incrementally less **(cyan curve)** and incrementally more conspicuous mutants **(orange curve)**. **(d)** [Bottom right]: Average population traits plotted as trajectories with averaging frequency $g = 2,000$. Black marks represent the average traits of a population after $10,000$ iterations.

Incorporating $\boldsymbol{a > 0}$

The black markers in Figure 4(d) suggest that **for** most populations **the average level of toxicity** converges to the predicted equilibrium **provided in** (31). **Populations starting with** low conspicuousness **risk being invaded by less conspicuous mutant types (the cyan curve in Figure 4(c) sits above the** $\boldsymbol{r}$**-axis for** $\boldsymbol{\bar{r}_1 < 0.5}$**)** and **the associated trajectories** quickly converge to crypsis, as expected. **For**

increasing levels of initial conspicuousness the leftwards component of the trajectories diminishes (more drastically than with $a = 0$ in Figures 2d and 3d) until it changes direction. **This change in direction is recorded at $\bar{r}_1 \approx 1.5$, beyond which directional selection associated with the invasion fitness gradient vanishes (the cyan and orange curves in Figure 4c converge) while the absolute resident fitness continues to increase. A considerable proportion of the trajectories with initial conspicuousness $\bar{r}_1 > 2$ in Figure 4(d) are observed to evolve toward higher conspicuousness and we speculate that this can be traced back to the resident fitness.**

Comparing these observations with those in Figures 2 and 3 we deduce that the impact of absolute resident fitness is more substantial when the size of the local relatedness parameter is greater. **Indeed, in Figure 4(d) we observe evolution** towards higher levels of resident fitness, especially in regions where there is no directional **selection** associated with mutant fitness and in which mutants that are less conspicuous are notably worse-off compared with the residents. Crypsis tends to be the default and preferred strategy for a multitude of chemically defended prey and it is of interest to determine how and why aposematic solutions with a strong signalling component could instead admit a more viable option.

A plausible explanation for the above results can be found by considering positive frequency-dependent selection. Consider a mutant invading a resident population whose r strategy is similar but distinct. For our model there is a continuum of r values that are stable against invaders playing different r (both smaller and larger; this is because there is an inherent disadvantage for looking different from everybody else). For a pair of such strategies, A and B, an A population is stable against B invaders and a B population is stable against A invaders. Mutants can appear with higher or lower r values, and there will be a small probability of successful invasion, which is amplified by the size of the local relatedness parameter a. If this parameter is large enough then due to positive frequency dependence on initial invasion and the finiteness of the population, invaders can quickly reach a sufficiently high overall frequency through a sequence of drift related invasions. It is likely that once a certain (threshold) frequency is reached selection turns positive for the mutant (as there is now an inherent disadvantage to the residents for looking unlike the invading mutant group) leading such an invader to go to fixation.

As we observe in the simulations of Figure 4 the type with the higher resident fitness generally has a higher probability to invade the type with lower resident fitness than for the reverse invasion. Thus a sequence of drift related invasions **of the kind discussed** will tend to move the population in the direction of higher resident fitness. The higher the value of parameter a the greater the local frequency of the mutant at the start, and so the lower the advantage to the resident. This increases the probability of any invasion in either direction, but the increase is more marked in the direction of higher resident fitness because of its relative stability, so that increasing a amplifies the above effect.

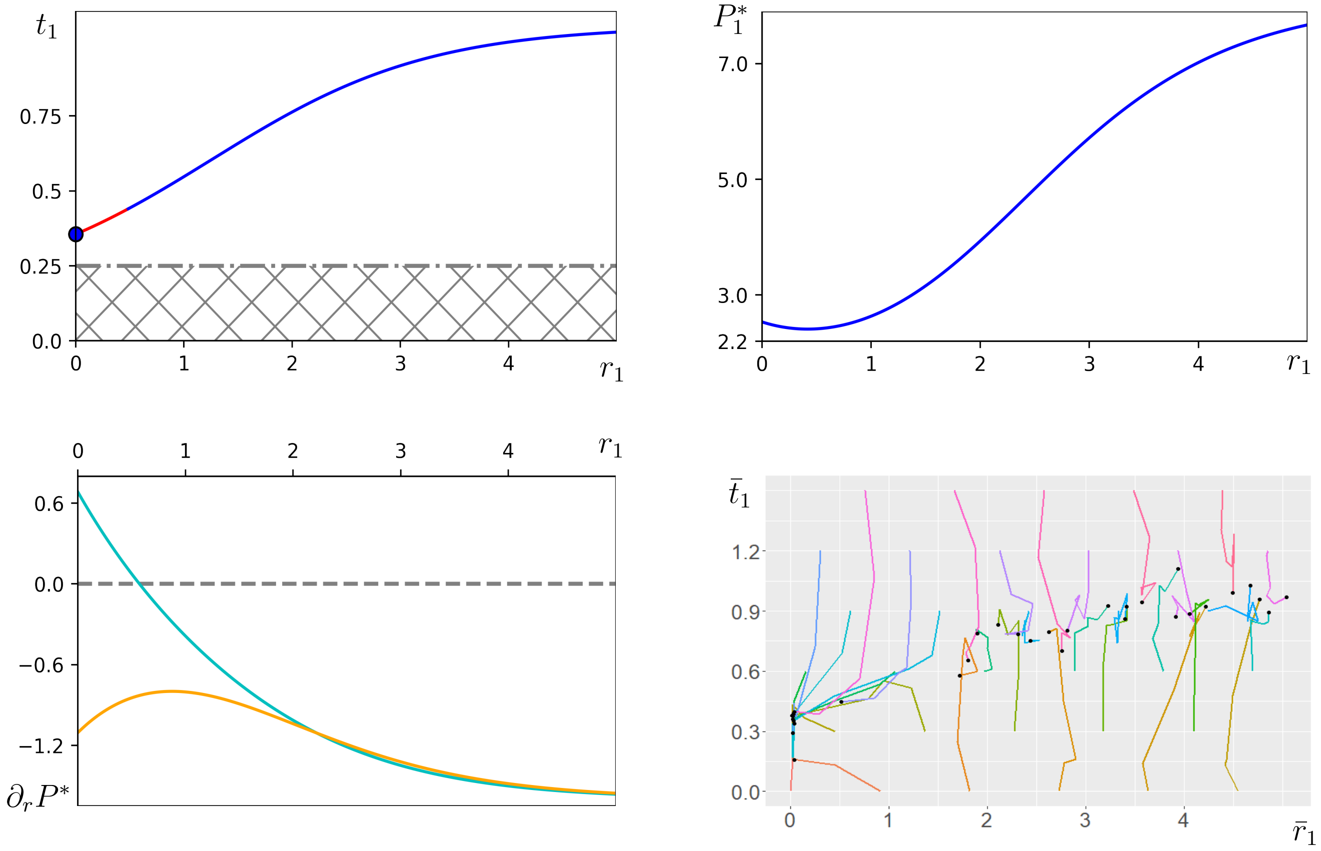

Figure 4: Parameter values $\lambda = 0, a = 0.5, f_0 = k_0 = q_0 = 1, f = 2.8, k = 5, t_c = 0.25, q = 0.4, N = 100, n = 10, v = 1$ **(a)** [Top left] Grey scaled region $\{(r_1, t_1) : r > 0, t_1 \leq 0.25\}$ violates (38). Blue marker at $(0, 0.356)$ represents the unique cryptic solution, which co-exists alongside a continuum of conspicuous unstable (red) and stable ESSs (blue) on the curve of (31). **(b)** [Top right] Resident fitness vs. conspicuousness evaluated at the equilibrium of Figure 4(a). Resident fitness is not impacted by parameter a and is therefore provided as in Figures 2(b) and 3(b) through (8). **(c)**[Bottom left] Invasion fitness gradient along r evaluated at equilibrium **- see** (42) **as well as** (38) **and** (39) **-** for incrementally less **(cyan curve)** and incrementally more conspicuous mutants **(orange curve)**. **(d)** [Bottom right] Average population traits plotted as trajectories with averaging frequency $g = 2,000$. Black markers represent traits averaged over the population after $10,000$ iterations and mostly converge to the equilibrium toxicity **levels** in 3(a).

3.2 Solutions with non-zero background mortality $\lambda > 0$

In the previous subsection we established a set of empirical rules that can serve as a guide in our understanding of how aposematic traits evolve in finite prey populations that are subject to random mutation. We observed that while in the t-direction populations mostly evolve toward the predicted equilibrium, **evolution along the *r*-axis is less straightforward**. That is, one first has to consider whether there is directional **selection** for less or more conspicuous mutant types (for less conspicuous **resident** populations it tends to be the former) and particularly how much better/worse the type in question is compared with the resident. Second to this, we gauge the size of the directional **selection** (at equilibrium) with regards to the absolute resident fitness. It appears that this secondary cause can select for conspicuous solutions provided the local relatedness parameter is large enough and **invasion** in either direction is unlikely from the mutant fitness perspective.

In this subsection we introduce non-zero rates of background mortality, initially in absence of local relatedness effects and finally including these. The presentation in this part places stronger emphasis on the outcomes of numerical simulation so as to showcase a larger breadth of examples within this less-explored regime and more effectively observe the impact of varying the background mortality rate on finite populations.

The $a \to 0$ limit

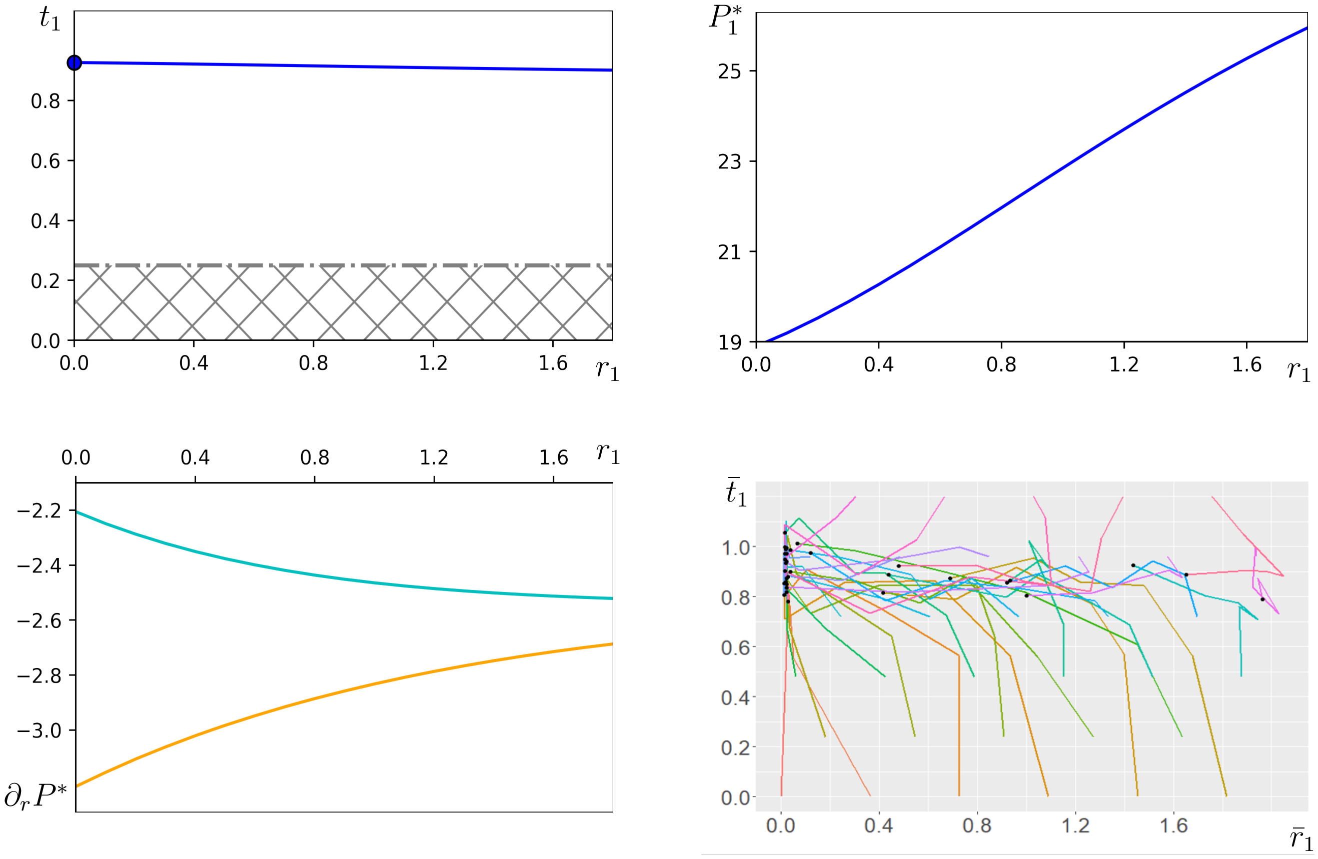

Figure 5: Parameter values $\lambda = 0.0015, a = 0, f = 5/6, k = 5, t_c = 0.25, q = 0.4, N = 100, n = 10, v = 1(a)$ [Top left] Unique cryptic ESS represented as a solid marker at $(0, 0.927)$ co-exists alongside a continuum of conspicuous ESSs drawn in blue and defined implicitly. The equilibrium toxicity is predicted to decrease (slowly) with increasing levels of conspicuousness, which is expected since it lies above the curve $c(r_1)$ in (51). (b) [Top right] Absolute resident fitness defined implicitly through (50) and plotted as a function of the conspicuousness. (c) [Bottom left] Invasion fitness gradient along r evaluated at equilibrium **- see** (42) **as well as** (38) **and** (39) **-** for incrementally less **(cyan curve)** and incrementally more conspicuous mutants **(orange curve)**. (d) [Bottom right] Average population traits plotted as trajectories with averaging frequency $g = 2,000$. Black markers show strong converge to crypsis, which is mostly supported from Figures 5(b) and 5(c)

Before discussing Figures 5 and 6 individually, we should remark that these relate to the same example but where in Figure 6 different sets of trajectories are plotted for different levels of background mortality (Figures 5d and 6b are identical). In Figure 5(d) prey traits are mostly observed to converge to the equilibrium level shown in Figure 5(a), which is determined implicitly through setting $a = 0$ in (26). The equilibrium level of defence is predicted to decrease with increasing levels of conspicuousness, although this effect is not captured in Figure 5(d) due to stochastic effects (for reasons discussed in due course these tend to be stronger when parameter a is small). The trajectories in 5(d) exhibit a strong pull toward crypsis and this is more pronounced for lower values of the conspicuousness, where the cyan curve is highest. Presently, we confirm existing intuition (drawn from our discussions of Figures 2 and 3), namely that absolute resident fitness has limited impact on trait evolution when the local relatedness parameter is small/vanishing. Indeed, even in absence of strong directional selection, resident conspicuousness evolves against the resident fitness and toward

lower values of $\bar{r}_1$.

These conclusions are valid for the remaining three plots in Figure 6, from which two additional conclusions can be drawn: As the background mortality increases the equilibrium level of defence decreases and its relationship to conspicuousness at switches from decreasing at equilibrium (Figures 6a and 6b) to increasing (Figures 6c and 6d). The simulated plots in Figure 6 (this includes Figure 5d) exhibit considerably more randomness than their counterparts in Figure 7, which is likely attributed to the larger value of the local relatedness parameter in the latter (and its impact on initial invasion and fixation/drift). For this reason we elaborate on (i) and (ii) in the context of Figure 7 below and compare these to the analytical predictions in Appendix II.

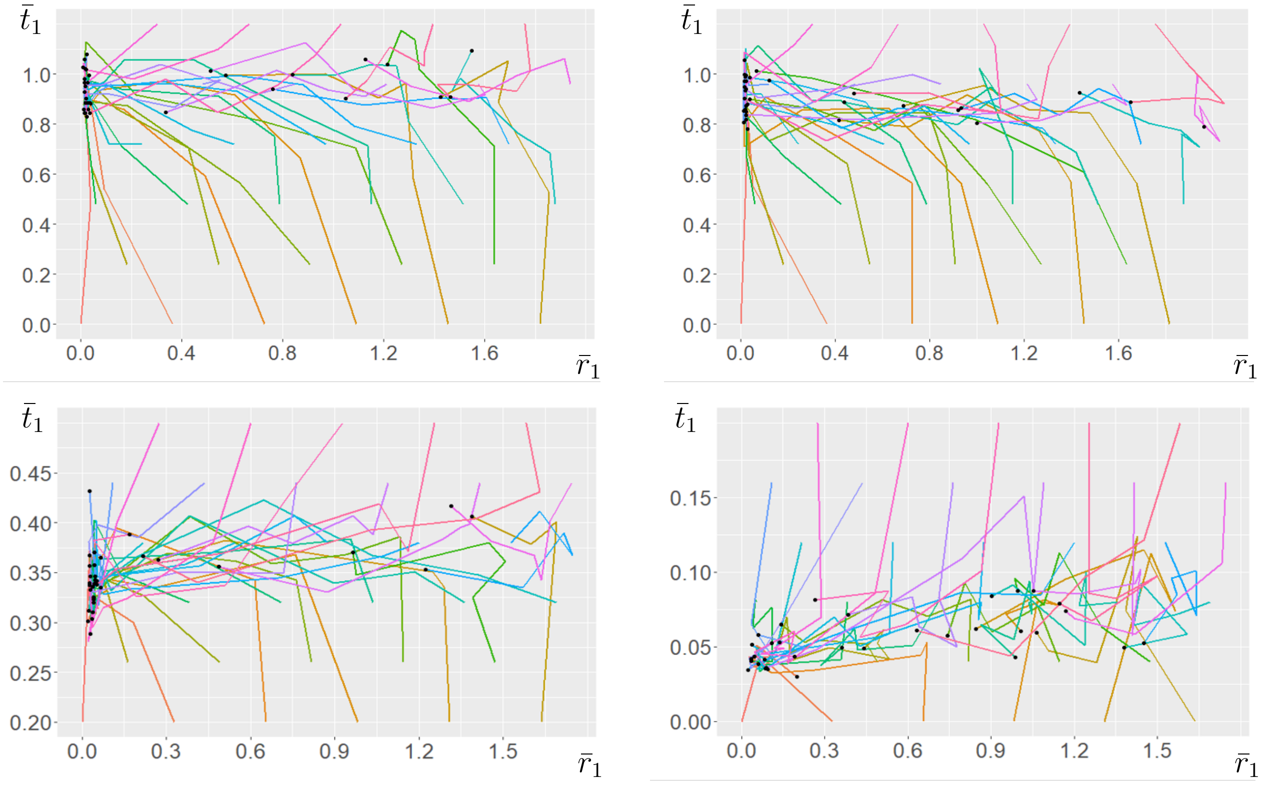

Figure 6: Parameters $a = 0, f = 5/6, k = 5, t_c = 0.25, q = 0.4, N = 100, n = 10, v = 1$. The plots are in increasing order of the parameter λ with (a) [Top left] $\lambda = 0.0001$; (b) [Top right] $\lambda = 0.0015$; (c) [Bottom left] $\lambda = 0.2$ and (d) [Bottom Right] $\lambda = 2$. Together the plots (mostly) confirm that increasing λ causes a decrease in the associated level of toxicity (for fixed conspicuousness) and that equilibrium toxicity switches from decreasing with the conspicuousness to increasing. The accumulation of black markers suggests strong selection for crypsis, likely driven by directional **selection** of mutant fitness in that direction. We also remark that trajectories in (a) convey a mostly flat equilibrium at $t \approx 1/f - 1/k$ as in (31) and that in (d) trajectories are traced out within the non-aversive region $t_1 \leq t_c = 0.25$.

Incorporating $a > 0$

In closing this section we consider non-zero rates of background mortality alongside non-zero levels of local relatedness (parameter a). The simulations in Figure 7 confirm the prediction (see related discussion in Appendix I) that for fixed conspicuousness the equilibrium level of defence decreases with increasing rates of background mortality. It is natural to expect that the value of anti-predatory defences diminishes when the level of threat outside of predation is significant. Since investment in chemical defences is costly, there is little benefit in investing in defences if this does not manifest as a notable increase in the life-span of prey (through a reduction in predation). Second and interrelated to the **above** is that the equilibrium level of defence can exhibit a **decreasing** relationship with respect to the conspicuousness (see Figures 6/7a and 6/7b) when the associated background mortality λ is low and that this relationship can switch to the converse when the background mortality increases (see Figures 6/7c and 6/7d). [1]

We have established - this is done analytically in Appendix I.B - that the level of defence at equilibrium decreases with increasing values of the parameter λ, such that for small values of λ, the associated level of defence is high enough that prey are highly aversive for predators. In such cases the plots in Figures 6/7 (a),(b) suggest that prey can increase aversiveness further by increasing conspicuousness while simultaneously (slightly) decreasing investment in defence. In contrast, if λ is high, the overall level of defence is low and prey are not very aversive so that larger conspicuousness selects for slightly more investment in defence (see Figures 6/7 c, d). The latter is likely because the gain in terms of (signalled) aversiveness outweighs the costs, which in turn can be traced back to with the choices of functional forms in (10).

The predicted slope of the equilibrium curve is provided by the *Implicit Function Theorem* in $\mathbb{R}^2$, which for the functions (10) used in the simulation takes the form shown in (29). As discussed in Appendix I.B the single term in the predicted equilibrium (26) that can accommodate changes in monotonicity is $(\lambda/DKQ) \times F'/F$ and describes the impact on fecundity (of increased defence) scaled as a proportion of background to predator-induced deaths. This quantity can be seen as an honest measure for the capacity of investment in aposematic defences to increase prey fitness (through favourable trade-off involving life-span and reproduction). When λ is low (and prey are aversive) it is optimal for prey to increase their reproductive success by reducing their toxicity in favour of slightly higher mortality (seen through increased conspicuousness). The functional forms are such that when λ is high (and prey are non-aversive) the optimal trade-off regime changes so that it is best for prey to reduce their reproductive success (by increasing toxicity) in favour of reduced predation (seen through an increased conspicuousness). The reader is strongly encouraged to compare the findings of Figures 6/7 with the analysis in Appendix I.B.

Third, we remark that when the invasion fitness gradient along r is flat enough in either direction and provided the local relatedness parameter is strong enough, absolute resident fitness can have a notable influence on the evolution of prey traits. This is likely attributed to an amplification of the group effect that larger values of the parameter a has (see earlier explanation about frequency-dependence) and could explain why strategies with considerable signalling component are selected for when λ is sufficiently small (see Figures 7a and 7b). The latter is rather clearly showcased in Figure 7, where trajectories evolve against the invasion fitness gradient and toward increasing levels of the absolute resident fitness.

[1] The correlation between aposematic traits at ESS (i.e. whether defence is positively or negatively related with respect to the conspicuousness) alludes to the notion of *honest signalling*, which remains crucial and yet unclear among theoreticians and empiricists - see Summers et al. (2015) and related comments in the discussion.

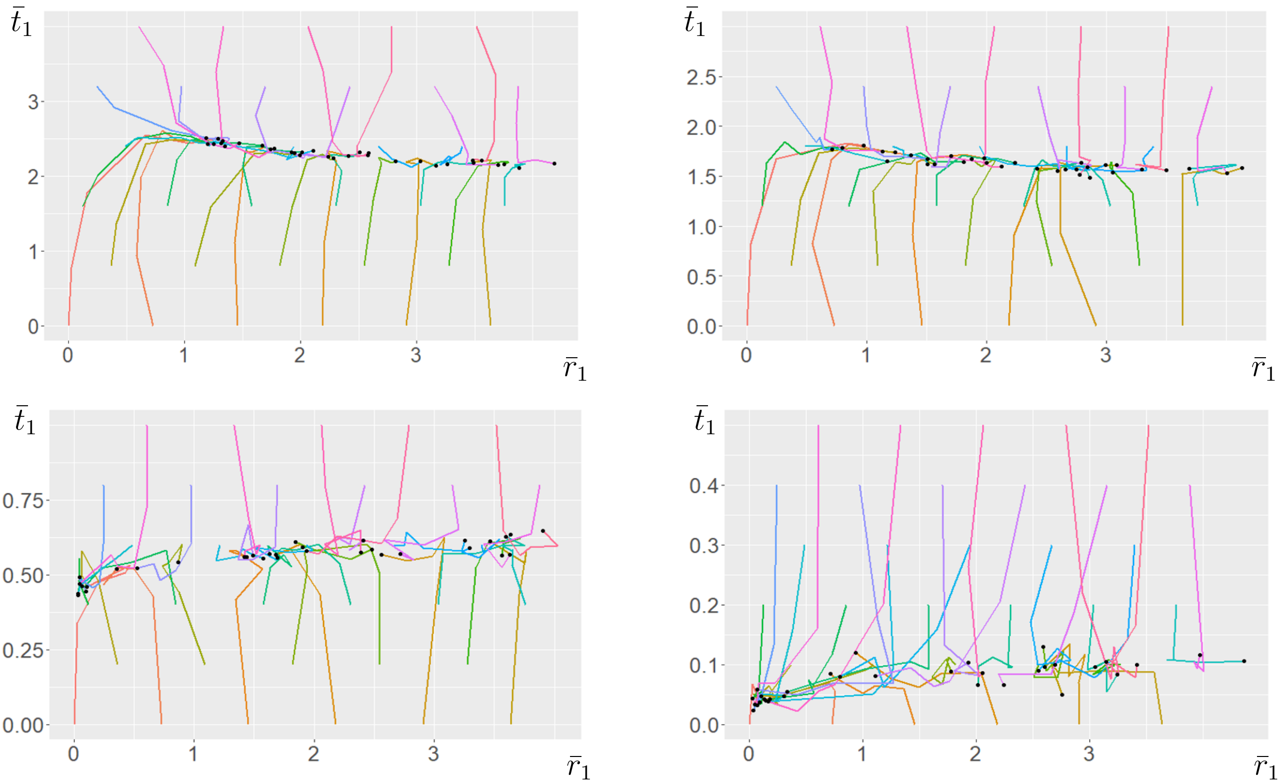

Figure 7: Parameter values $a = 0.5, f_0 = q_0 = k_0 = 1, f = 5/6, k = 5, t_c = 0.25, q = 0.4, N = 100, n = 10, v = 1$. Plots are positioned in increasing order of λ such that (a) [Top left] $\lambda = 0.0001$; (b) [Top right] $\lambda = 0.0015$; (c) [Bottom left] $\lambda = 0.2$ and (d) [Bottom Right] $\lambda = 2$. The plots confirm more certainly than Figure 6 that increasing λ is associated with a decrease in the associated level of toxicity and that the relationship between toxicity and conspicuousness switches from negative (in a and b) to positive (in c and d). In (a) and (b) there is strong selection for solutions with strong signalling component, likely on account of the absolute resident fitness being highest in that direction. In (c) and (d) it is clear that the resident fitness is not sufficient to counterbalance the impact of a strong invasion fitness gradient (from the left and along r). Large spaces between black markers (such as at)are likely due to a balancing effect of these opposite pulls.

4 Discussion

The results presented in the previous section have demonstrated both the strengths and limitations of applying **(infinite population)** ESS analysis within the broader mathematical development of Broom et al. (2006) to study the evolution of prey traits in finite populations. **To that end, had smaller populations been considered the outcomes of the simulations would have generally been driven by randomness.** We should remark that spatiotemporal variations in the various **environmental** factors (including territory quality) and in the predator's community structure are not explicitly accounted for **in our model**, even though we acknowledge their importance in the selection for/against aposematism in real populations. For instance, as discussed in Mappes et al. (2005) the genetic predisposition and cultural transmission of foraging strategies within families could lead to strongly localised selection for/against aposematism. In this closing section we call attention to these points and argue that the simulation model as described in Appendix II can be generalised to prey populations consisting of more than one species including Batesian mimicry complexes among others.

In an infinite population ESS analysis is all that matters, while in a very small population, stochastics dominates. For intermediate numbers, **stochastic mechanisms will eventually prevail in theory**, but this may take a really long time, so effectively the ESS analysis **is indeed** all that is needed. In the t-direction,

in any mixed population where all prey have similar conspicuousness, the optimal toxicity level is approximately the same, independent of the precise composition of the population, as long as the average conspicuousness does not change too much, or indeed often even if it does). In the r-direction, we **often** have a series of populations that are stable, but where the neighbouring mutants are not so much worse, so there is the prospect of invasion due to chance. **It is when we get to a substantial mutant sub-population that the mutants "resident fitness" (i.e., the mutants' fitness after fixation) comes into play**. Indeed full invasion is more likely to happen from the higher fitness side, so there will tend to be movement in that direction. The latter is manifest in Figure 7, where when the background mortality is sufficiently low aposematic strategies with considerable signalling component are selected (likely due to the higher associated fitness).

In this manuscript we have considered the functional forms of Broom et al. (2008) and compared regimes **with and without background mortality**. The conclusions drawn in previous works on aposematism have been constrained by the assumption that predation is the only source of prey death and the impact of varying regimes of background mortality has prior to now not been explored. In addition to accounting for sources of prey death outside of predation, we have explored the effect of local clustering through parameter a and utilised ESS (and fitness) analysis to draw conclusions about the evolution of prey traits in intermediate populations that are subject to stochasticity. While **simulation** models have been used in Ruxton and Beauchamp (2008) and in Zakharova et al. (2019) and elsewhere over the recent decades, these have never before been put to use to study aposematism. We have made contributions to the game-theoretical model of Broom et al. (2006) by broadening the scope of ESS analysis, by implementing it into a novel **simulation** model and by gauging the capacity of ESS (and fitness) analysis to predict the evolution of aposematism in **finite prey populations**.

In Broom et al., (2008) and previously in Broom et al. (2006) the use of simple functional forms and the suppression of background mortality had allowed to express the equilibrium toxicity explicitly in terms of the conspicuousness and to conjecture that more conspicuous appearances are associated with prey that are better-defended. This conjecture was disproved in Scaramangas and Broom (2022) and also presently, where a decreasing relationship between conspicuousness and defence was observed (see simulations in Figure 7) in regimes where prey death outside of predation is **rare**. In Scaramangas and Broom (2022) the justification of a decreasing relationship involved the implementation of a more elaborate (plausible nonetheless) association between the predator's propensity to attack based on its perception of prey aversiveness (through a modification of the form for Q). Interestingly, such a modification had also allowed us to disprove another conjecture of Broom et al. (2006) and to demonstrate that a certain level of signal strength may be associated **with** more than one ESS level of the defence.

Although the observation of a decreasing signalling-defence continuum is in this manuscript linked with functional forms that are different to those of Scaramangas and Broom (2022), perhaps the underlying mechanism is common. In the first simulations of Figure 7 (a and b) we observe that when the overall ESS level of defence is sufficiently high (background mortality is low) prey can afford to broadcast weaker defences through stronger signals because predator propensity to attack is already low and saturated. This is also complemented by the fact that a further investment in toxicity is costly to the fecundity and this is a cost worth bearing if it is manifest through reductions in predation, which in this case is not. In contrast, when background mortality is high and the associated toxicity is low (see simulations in Figures 7c and d) brighter appearances signal stronger defences because the reduction in fecundity is compensated with a reduction in predation and an increase in average prey life-span.

Indeed, of considerable importance to the theory of aposematic signalling is whether aposematic signals are *honest* (i.e. whether brighter prey are better defended) and the reader is encouraged to consult the review article by Summers et al. (2015) for a thorough account of this topic. While there is more empirical evidence reporting a positive relationship between conspicuousness and defence (Summers and Clough, 2001, Santos and Cannatella, 2011 and Maan and Cummings, 2012 are among several cited in Summers et al., 2015) there are noteworthy studies (including those of Wang, 2011 and Darst et al., 2006) suggesting that conspicuous signals could be dishonest. As argued in Scaramangas and Broom (2022) the model of Broom et al. (2006) is the only detailed exposition that can account for the full breadth of phenomena and this is observed presently.

The theory presented here makes clear predictions that would allow empirical testing. Perhaps our more interesting predictions stem from the comparison between the analytic theory and the simulations. It seems clear that when prey populations are large then the predictions of both modelling approaches converge, but for smaller population**s** the stochastic fluctuations captured in the simulation model should have a **strong** bearing. It would **be** valuable to explore **experimentally** with living prey how small a population has to be for these stochastic effects to have a strong **bearing** on evolutionary trajectories, how strong these effects are, and how exactly they alter the course of evolution. It seems more easy to imagine how such empirical explorations could be achieved in the laboratory than in natural populations. But even here there will be a challenge in finding a suitable prey type that can readily be kept in large numbers **and** shows the combination of appearance and toxicity characteristics of interest to us and that has a short enough generation time that meaningful evolutionary trajectories can be followed. A candidate here might be one of the stored-product beetles that are increasingly becoming model species for studies in evolution and population dynamics (of much relevance is the review article by Pointer et al., 2021). The most commonly-used species in such studies (Tribolium castaneum) is chemically defended and shows variation in coloration from red, through browns to black (see McLean, 2011).

We imagine that such experiments would involve not natural predators but artificial predation imposed by the experimenters – with different types of predation represented by removal of prey individuals from the population as defined by different sets of rules (mimicking the assumptions about predator behaviour in our theory). As well as exploring the consequences of prey population size (and indeed the size of the artificial predator population – as represented by the intensity of predator-mimicking mortality) on evolutionary trajectories – it would be straightforward to also explore our predictions about the effect of additional external non-predatory mortality in these experiments.

We also think that experiments with real predators would also be valuable in the context of testing our predictions. Well-developed systems for investigating why predators learn about aposematic prey and how this affects subsequent prey choice decisions already exist. These can use completely prey-naïve newborn domestic chicks (as in Rowland et al., 2013) or wild-caught insectivorous birds **temporarily** exposed to artificial prey in a laboratory setting (such as in Hämäläinen et al., 2020). Our model assumptions and predictions related to how predators respond in successive encounters with different types of prey items – particularly the assumptions about the spatial distribution of mutant types encapsulated in our parameter (a) could very naturally be explored empirically with such a system.

Furthermore, we see value in co-evolutionary experiments that allow us to explore whether the assumptions we make for predator behaviour in our models are likely a reasonable representation of those that evolve in real predators. For this, we might return to the evolutionary experiments with a simple laboratory prey organism like stored flour beetles discussed above, but rather than subjecting them to an unchanging predation regime, we allow the predatory regime to co-evolve with the prey. We have in mind here a population of artificial predators – each of which follows a set of rules about how it treats prey of different types, and thus imposes mortality on the prey population. However, variation in these rules will not only lead to variation in the form of mortality imposed on the prey but also on the fitness of the artificial predators – where a fitness score is awarded according to how well the predator exploits lower-defended prey and avoids higher-defended prey. If at each generation of the real prey the artificial predator population is changed such that more successful rule-systems become more prevalent in the artificial predator population, then we can effectively mimic predator-prey co-evolution – and most pertinently we can explore whether the predator population coalesces to rules that have commonality with those assumed in our theory. There is a collection of interesting studies examining the co-evolution of aposematic prey in a prey-predator complex including Teichmann et al. (2014) and Teichmann et al. (2015), whose results may be of particular insight to the experimenter.

In closing, we would like to highlight the success of the simulation in showcasing the evolution of aposematism in prey populations that are finite. We would also like to argue that it is possible to extend the game-theoretic treatment of Broom et al. (2006) to account for Batesian mimicry systems, which are arguably among the most important (and most studied) mimicry complexes encountered in nature. Work of this type could utilise the territorially-divided habitat structure referred to in Scaramangas and Broom (2022) and introduce on this a proportion of (beta-distributed) undefended mimics. Achieving stability of a model and a mimicking species in a certain habitat on the (longer) evolutionary time-scales requires that the individual sub-populations are stable on the (shorter) ecological time-scales and such a condition need be considered jointly with the ESS conditions detailed here. Research in this direction is promising and currently underway.

Appendices

Appendix I. The evolutionary stability of aposematic signalling

The purpose of this appendix is to summarise and expand on the mathematical modelling of aposematic signalling as this was originally formulated in Broom et al. (2006). The description we provide here accounts for later works on the same process, including Broom et al., (2008) and Scaramangas and Broom (2022).

In the first subsection of the appendix we discuss evolutionarily stable levels of aposematic signalling in generality (but without considering the *colouration* trait used in Broom et al., 2006). The discussion is particularised in the subsequent subsection, where special attention is payed to the functional forms of (10) that are employed in the **simulation** model. While the mentioned choices of functions have been utilised in previous works including Broom et al., (2008) these have not considered non-zero rates of background mortality. We should remark that the specific substitutions utilised here to evaluate the invasion fitness gradient at equilibrium and the absolute resident fitness at equilibrium are novel.

I.A General ESS analysis

Presently, we conduct ESS analysis in terms of the general functional forms of Table 1, following the presentation of Broom et al. (2006). Excluding colouration, the strategy space is identified with the boundary-inclusive, right upper-half plane and therefore there are four conditions for local ESS to consider. It is mentioned in (9) that in the interior subregion of the strategy space $\{(\rho,\tau):\rho>0,\tau>t_c\}\subset\{(\rho,\tau):\rho>0,\tau>0\}$ the conditions for local ESS read

$$\partial_t P(r_1,t_1)=0,\quad \partial_{tt}P(r_1,t_1)<0,\quad \overleftarrow{\partial}_r P(r_1,t_1)>0 \text{ and } \overrightarrow{\partial}_r P(r_1,t_1)<0.$$

The conditions for local ESS on the boundary $\{(\rho,\tau):\rho>0,\tau=0\}$ are

$$\overrightarrow{\partial}_t P(r_1,0)<0,\ \overleftarrow{\partial}_r P(r_1,t_1)>0 \text{ and } \overrightarrow{\partial}_r P(r_1,t_1)<0. \tag{11}$$

At the origin $\{(\rho,\tau):\rho=0,\tau=0\}$ these are

$$\overrightarrow{\partial}_t P(0,0)<0 \text{ and } \overrightarrow{\partial}_r P(0,0)<0, \tag{12}$$

and finally on the boundary $\{(\rho,\tau):\rho=0,\tau>0\}$, which describes prey that are cryptic but defended the conditions for ESS read

$$\partial_t P(r_1,0)=0,\quad \partial_{tt}P(r_1,0)>0,\ \text{ and } \overrightarrow{\partial}_r P(r_1,0)<0. \tag{13}$$

We now provide limit definitions for the partial derivatives that are mentioned in the conditions for local ESS above. Let h be positive and arbitrarily small **(i.e. $0<h\ll 1$). Quantity $\partial_t P(r_1,t_1)$, which features in** (9) **and** (13) **is shorthand for the partial derivative of the mutant payoff with respect to the mutant trait t evaluated at the resident value $(r,t)=(r_1,t_1)$ with remaining resident traits r_1 and t_1 held fixed. In particular**

$$\boldsymbol{\partial_t P(r_1,t_1)}=\partial_t P(r,t;r_1,t_1)|_{r=r_1,t=t_1}:=\lim_{h\to 0}\frac{P(r_1,t_1+h;r_1,t_1)-P(r_1,t_1;r_1,t_1)}{h}, \tag{14}$$

with higher order derivatives in (9) **defined in a similar way. If the resident value is drawn**

from the boundary $\{(\rho,\tau):\rho\geq 0,\tau=0\}$ as in (11) **and** (12) **mutations in** t **can only be positive so that quantity** $\overrightarrow{\partial}_t P(r_1,0)$ **describes the rate at which the mutant fitness changes in response to changes in the mutant trait for toxicity. We have**

$$\overrightarrow{\partial}_t P(r_1,0)=\overrightarrow{\partial}_t P(r,t;r_1,0)|_{r=r_1,t=0}=\lim_{h\to 0}\frac{P(r_1,h;r_1,0)-P(r_1,0;r_1,0)}{h}. \tag{15}$$

Mutant fitness is non-differentiable along the r-direction at $r=r_1$ and therefore we make use of half derivatives along the left and right directions. These are

$$\overleftarrow{\partial}_r P(r_1,t_1)=\overleftarrow{\partial}_r P(r,t;r_1,t_1)|_{r=r_1,t=t_1}:=\lim_{h\to 0}\frac{P(r_1-h,t_1;r_1,t_1)-P(r_1,t_1;r_1,t_1)}{h} \tag{16}$$

and

$$\overrightarrow{\partial}_r P(r_1,t_1)=\overrightarrow{\partial}_r P(r,t;r_1,t_1)|_{r=r_1,t=t_1}:=\lim_{h\to 0}\frac{P(r_1+h,t_1;r_1,t_1)-P(r_1,t_1;r_1,t_1)}{h}. \tag{17}$$

We remark that (16) **describes the rate with which the mutant fitness changes in response to a reduction in the level of conspicuousness. The change in fitness associated with a negative step** $-h$ **along** r **is given by** $P(r_1-h,t_1;r_1,t_1)-P(r_1,t_1;r_1,t_1)$ **- i.e. the value after mutation minus the resident value - and can be approximated (to first-order) by** $-\overleftarrow{\partial}_r P(r_1,t_1)\times h$**. For interpreting simulations it is more meaningful to use** $-\overleftarrow{\partial}_r P(r_1,t_1)$ **as the (left) invasion fitness gradient along** r **because this accounts for the sign of the mutation step (see also** (38)**,** (39) **and** (42) **for the explicit forms used in the simulation).**

Strategies satisfying the equilibrium condition $\partial_t P(r_1^*,t_1^*)=0$ are elements of the zero-level set of the map $(r,t)\mapsto\partial_t P(r_1,t_1)$. We should also clarify that the notation* is reserved for strategies defined on the curve

$$\{(r_1^*,t_1^*):\partial_t P(r_1=r_1^*,t_1=t_1^*)=0:r_1^*\geq 0,t_1^*>0\} \tag{18}$$

describing the equilibrium level of defence. Functions defined on the equilibrium curve are also denoted with a * **as are** mutant/resident quantities evaluated along this curve. For instance, the resident fitness along the equilibrium curve is commonly denoted P_1^* in Figures 2(b), 3(b), 4(b) and 5(b) and what is meant is $P_1^*:=P_1(r_1^*,t_1^*)$. **By construction, the mutant payoff in** (7) **is** $\mathcal{C}^l$ **with** $l\geq 2$ **sufficiently near the resident value. As Scaramangas and Broom (2022) discuss, condition** (18) **may not necessarily result in an expression for the equilibrium level of defence in terms of the conspicuousness that is explicit and in such cases the *Implicit Function Theorem* (IFT) in** $\mathbb{R}^2$ **can be used to better understand this relationship. The theorem states that if** (r_1^*,t_1^*) **is an equilibrium value defined through** (19) **then there exists a smooth function** g defined on the vicinity of r_1^* with $g(r_1^*)=t_1^*$, whose tangent has slope

$$-\frac{\partial_{r_1}\partial_t P(r_1^*,t_1^*)}{\partial_{t_1}\partial_t P(r_1^*,t_1^*)}. \tag{19}$$

In Broom et al. (2006) it is argued that for most choices of biologically feasible functions the numerator in (19) is strictly positive whenever $t_1>t_c$ and self-consistent reasoning was given to support this. Although this result holds true for the functions considered in Broom et al. (2008), it is not sufficient justification for ruling out the prospect of local ESSs that are decreasing with increasing conspicuousness as it does not account for the sign of the denominator.

Provided that $t_1>0$ substitution of (7) into the first equality in (9) suggests that a resident strategy (r_1^*,t_1^*) is an equilibrium strategy (in the t-direction) if it satisfies the level-set condition

$$\frac{\lambda}{D(r_1^*)K(t_1^*)Q(I_1)}\frac{F'(t_1^*)}{F(t_1^*)}+\frac{F'(t_1^*)}{F(t_1^*)}-\frac{K'(t_1^*)}{K(t_1^*)}-aI_1\frac{H'(t_1^*)}{H(t_1^*)}\frac{Q'(I_1)}{Q(I_1)}=0. \tag{20}$$

Aposematism is principally a form of anti-predatory defence and the value of λ as a proportion of the predator-induced mortality rate $D(r_1^*)K(t_1^*)Q(I_1^*)$ provides an honest measure of the capacity of aposematism to increase prey fitness (through positively influencing their life-span). If predation is the only source of prey death [2] the first term on the LHS of (20) vanishes (since $\lambda = 0$) and therefore the cost to fecundity payed by mutants with incrementally higher levels of internal defence (compared with the residents) is captured entirely by the second term, which depends only on the resident level of defence. If $\lambda > 0$ the first term on the LHS of (20) does not vanish, which suggests that the mentioned cost to fecundity must now account for the proportion of deaths attributed to outside sources compared with those attributed to predation $(\lambda/D(r_1^*)K(t_1^*)Q(I_1^*))$. From the model description in section 2, predator-induced mortality generally exhibits a complex dependence on prey traits, particularly because this depends on the predator's perception of prey aversiveness. It is therefore natural that the inclusion of non-vanishing rates of background mortality renders the relationship between conspicuousness and defence at ESS less straightforward and implicit. We return to this discussion in a more matter-of-fact manner in the section that follows, in which we consider the example functions of (10) that are used for the **simulation** model.

A strategy defined on the equilibrium curve (20) is stable along the t-direction if the associated mutant fitness is concave down

$$-\frac{\lambda}{D(r_1^*)K(t_1^*)Q(I_1)}\frac{F''(t_1^*)}{F(t_1^*)} - \frac{F''(t_1^*)}{F(t_1^*)} + \frac{K''(t_1^*)}{K(t_1^*)} + 2aI_1\frac{H'(t_1^*)}{H(t_1^*)}\frac{Q'(I_1)}{Q(I_1)}\frac{K'(t_1^*)}{K(t_1^*)}$$

$$+a^2\left(I_1\frac{H'(t_1^*)}{H(t_1^*)}\right)^2\frac{Q''(I_1)}{Q(I_1)} + aI_1\frac{H''(t_1^*)}{H(t_1^*)}\frac{Q'(I_1)}{Q(I_1)} > 0. \tag{21}$$

In the special case that the resident is totally undefended and plays $t_1 = 0$ stability in the t-direction is guaranteed provided this single inequality holds

$$\frac{\lambda}{D(r_1^*)K(0)Q(I_1)}\frac{F'(0)}{F(0)} + \frac{F'(0)}{F(0)} - \frac{K'(0)}{K(0)} - aI_1\frac{H'(0)}{H(0)}\frac{Q'(I_1)}{Q(I_1)} < 0. \tag{22}$$

A strategy (r_1, t_1) with non-zero signalling component $r_1 > 0$ is stable in the r-direction provided it can resist invasion from less and more conspicuous mutants. For the former, we require

$$-\frac{D'(r_1)}{D(r_1)} - aI_1\frac{D'(r_1)}{D(r_1)}\frac{Q'(I_1)}{Q(I_1)} + (1-a)I_1\frac{Q'(I_1)}{Q(I_1)}S'(0) > 0 \tag{23}$$

and for the latter we require

$$-\frac{D'(r_1)}{D(r_1)} - aI_1\frac{D'(r_1)}{D(r_1)}\frac{Q'(I_1)}{Q(I_1)} - (1-a)I_1\frac{Q'(I_1)}{Q(I_1)}S'(0) < 0. \tag{24}$$

We should remark that in the case of a cryptic strategy with $r_1 = 0$ only (24) is necessary

$$-\frac{D'(0)}{D(0)} - aI_1\frac{D'(0)}{D(0)}\frac{Q'(I_1)}{Q(I_1)} - (1-a)I_1\frac{Q'(I_1)}{Q(I_1)}S'(0) < 0 \tag{25}$$

as there are only the more conspicuous mutants to consider as potential invaders.

It is a general result of Broom et al. (2006) that the conspicuous signalling of strategies that are non-aversive is not locally evolutionarily stable. This is a direct consequence of (23) for which it holds that when a resident strategy is drawn from $\{(\rho, \tau) : \rho > 0;\ \ 0 \le \tau \le t_c\}$ the terms on the LHS are individually negative

[2] This is an assumption previous works including those of Broom et al., 2006, Broom et al., 2008 and Scaramangas and Broom, 2022 had relied on.

so that the reverse of (23) holds. A term-by-term interpretation of the mentioned inequality suggests that mutants playing a strategy that is (incrementally) less conspicuous accrue (small) advantages in fitness, which are associated with lower rates of predation (through reductions in detection, recollection and to an overall imperfect resemblance to a majority of residents that is perceived as attractive). The associated region $\{(\rho,\tau) : \rho > 0;\ \ 0 \leq \tau \leq t_c\}$ in the strategy space is shown in grey (see Figures 2a, 3a, 4a, 5a) to demonstrate that stability is generally ruled out.

While it is a known result of Broom et al. (2006) that non-aversive strategies fail (23), we remark that this only applies to strategies that are conspicuous. Since cryptic types cannot give rise to mutants that are less conspicuous these only risk being invaded by mutants that are more conspicuous and therefore stability along r is achieved solely through inequality (24), whose validity cannot be rejected a-priori **and makes crypsis the only possibility for non-aversive strategies to co-exist alongside aversive ones.** This forms the basis for the predictions in Broom et al. (2006) and in Scaramangas and Broom (2022) that cryptic ESSs can co-exist alongside conspicuous strategies.

I.B ESS analysis for the simulation model

In this subsection of the appendix we narrow our attention to the functional forms of (10) and draw analytical conclusions about the behaviour of the system at ESS. A strategy (r_1^*, t_1^*) is in equilibrium along the t-direction if it satisfies (20), which through the forms in (10) reads

$$-\frac{\lambda f}{q_0 k_0}(1+kt_1^*)(1+\exp(-r_1^*))\exp\left(q\frac{N}{n}\frac{t_1^*-t_c}{1+\exp(-r_1^*)}\right)-f+\frac{k}{1+kt_1^*}+\frac{aq\frac{N}{n}}{1+\exp(-r_1^*)}=0. \tag{26}$$

There is an immediate conclusion to be drawn from the above, which confirms our intuition that anti-predatory defences are of diminishing value in regimes of increasing non-predatory threat. Indeed, suppose that for some level of background mortality $\lambda = \lambda^*$, the strategy (r_1^*, t_1^*) is a solution to (26). Since the LHS of that equality decreases with increasing values of either t_1^* and/or λ^* it follows that an increase in λ would lead to a decrease in the equilibrium toxicity t_1^* associated with conspicuousness r_1^*. That is, for fixed conspicuousness, the equilibrium level of defence decrease with increasing levels of the background mortality. We would expect there to be little value in investing in defences that are costly to the fecundity in regimes where these have limited capacity to increase prey life-span and it is worth mentioning that this result is confirmed in the simulated plots of Figures **6**/7.

It is immediately clear from (26) that if $\lambda > 0$ and $a > 0$ it is not possible to obtain the ESS level of toxicity explicitly in terms of the conspicuousness. This is unlike the situations encountered previously in Broom et al. (2006) and Broom et al. (2008) and is indicative of a broader class of examples in which one trait can only be determined in terms of the other at ESS through a rule that is implicit. **Scaramangas and Broom (2022) discuss that in such cases the relationship between conspicuousness and defence at equilibrium can be better understood through the *Implicit Function Theorem* in** $\mathbb{R}^{\mathbf{2}}$**. Presently** we provide a derivation for the slope of the line tangent to the (implicitly defined) equilibrium curve given in (26) **by utilising** (19). The r_1-derivative of the LHS of (26) reads

$$-\frac{\lambda f}{q_0 k_0}(1+kt_1)\exp\left(\frac{q\frac{N}{n}(t_1-t_c)}{1+\exp(-r_1)}-r_1\right)\left[\frac{q\frac{N}{n}(t_1-t_c)}{1+\exp(-r_1)}-1\right]-\frac{aq\frac{N}{n}\exp(-r_1)}{(1+\exp(-r_1))^2}, \tag{27}$$

while the t_1-derivative reads

$$-\frac{\lambda f}{q_0 k_0}(1+\exp(-r_1))\exp\left(\frac{q\frac{N}{n}(t_1-t_c)}{1+\exp(-r_1)}\right)\left[k+\frac{q\frac{N}{n}\boldsymbol{(1+kt_1)}}{1+\exp(-r_1)}\right]-\frac{k^2}{(1+kt_1)^2}. \tag{28}$$

Evaluated at $(r_1, t_1) = (r_1^*, t_1^*)$ the slope of the line tangent to the equilibrium curve is given by

$$\frac{\dfrac{\lambda f}{q_0 k_0}(1 + kt_1)\exp\left(\dfrac{q\frac{N}{n}(t_1^* - t_c)}{1+\exp(-r_1^*)} - r_1^*\right)\left[\dfrac{q\frac{N}{n}(t_1^* - t_c)}{1+\exp(-r_1^*)} - 1\right] + \dfrac{aq\frac{N}{n}\exp(-r_1^*)}{(1+\exp(-r_1^*))^2}}{\dfrac{\lambda f}{q_0 k_0}(1+\exp(-r_1^*))\exp\left(\dfrac{q\frac{N}{n}(t_1^* - t_c)}{1+\exp(-r_1^*)}\right)\left[k + \dfrac{q\frac{N}{n}(1+kt_1^*)}{1+\exp(-r_1^*)}\right] + \dfrac{k^2}{(1+kt_1^*)^2}}. \tag{29}$$

It is immediately clear that the denominator in (29) is always positive so that the monotonicity of the equilibrium curve can change only through changes in the sign of the numerator. This is unlike the example discussed in Scaramangas and Broom (2022) where sign changes were attributed to the denominator and manifest as vertices at which the line tangent were vertical. Here, we observe that if the equilibrium level of defence is sufficiently low (this can be the case when λ is low) the term in square brackets can be made negative enough to make the numerator negative, such that the associated equilibrium level of defence decreases as the conspicuousness increases. Likewise, when the background mortality rate λ is high enough the associated term in square brackets is positive so that the numerator (and fraction) is positive overall and the equilibrium level of defence increases with increasing levels of conspicuousness. Changes in monotonicity are observed in Figures 6/7 and discussed therein.

From (28) it is clear that the terms on the LHS of (26) are decreasing with respect to t_1. Likewise, it is observed from (27) that when t_1 is sufficiently low/high (e.g. λ is high/low) the first term in (26) is increasing/decreasing with respect to r_1 while the fourth term is monotonically increasing with respect to r_1 (independent of t_1). Suppose that (r_1^*, t_1^*) satisfies the equilibrium condition (26) for some low enough value of λ that the overall sign of (27) is negative. In this case, a marginal increase in r_1^* will (by assumption) reduce the LHS of (26) which, on account of (28) being negative, must be compensated by a reduction in t_1^*. The latter suggests that when λ is sufficiently low the equilibrium level of defence (defined implicitly through (26)) is decreasing with respect to conspicuousness. Likewise, we can assume that (r_1^*, t_1^*) satisfies (26) for some value of λ that is sufficiently high that (27) is positive. In such a situation increasing r_1^* would cause the LHS of (26) to increase so that to restore equilibrium this must be compensated with an increase in t_1^*, suggesting that for high enough λ the equilibrium defence increases with conspicuousness.

Indeed, careful consideration of (26) leads us to the observation that there are four cases to consider: (i) $\lambda = 0, a = 0$; (ii) $\lambda = 0, a > 0$; (iii) $\lambda = 0, a > 0$ and (iv) $\lambda > 0, a > 0$. In case (i) it is immediately clear that setting $\lambda = 0$ and $a = 0$ **in** (26) eliminates the first and fourth terms on the LHS so that prey defence (at ESS) is not associated with the conspicuousness. That is

$$t_1^* = \frac{1}{f} - \frac{1}{k}, \tag{30}$$

for all $r_1^* \geq 0$. This suggests that mutants with incrementally higher levels of defence (compared with the residents) pay a price for reproducing at a slower rate, but are better defended against attacks that are potentially lethal so that at the (unique) ESS level of defence the two components balance as in (30). An important assumption of the model (see section 2) is that investment in defences (but not in bright colourations) is costly and this is reflected in the negative dependence on t of the fecundity function $F(t)$. Indeed, once the level of toxicity described in (30) is reached, resident strategies with different signalling component may have different overall levels of fitness, but cannot be invaded by mutants that are (incrementally) more/less defended (since the trade-off between F and K is exact). An alternative (but more equation-intensive) approach would be to impose that investment in bright colourations also impacts the fecundity negatively. Doing so would introduce a dependence on the conspicuousness of the ESS level of defence, even within the regime described by (i).

In case (ii) the level of defence satisfying (26) can be provided explicitly in terms of the conspicuousness as

$$t_1^*(r_1^*) = \frac{1}{f - \dfrac{aq\frac{N}{n}}{1+\exp(-r_1^*)}} - \frac{1}{k} \tag{31}$$

for all $r_1^* \geq 0$. The latter suggests that the ESS level of defence is increasing with increasing levels of the conspicuousness and that the increase is sharper for larger values of the parameter a; we direct the reader to Broom et al., (2008) for a more careful consideration of this example. The situation in (ii) is different to (i) in that mutation is now assumed to occur in clusters of size a, whose size influences their perceived aversiveness and the probability that predators visiting their site mount attacks on them. So while it is true that for mutants with incrementally larger levels of defence the cost to fecundity must be counterbalanced by the benefit of escaping potentially lethal attacks, there is in (ii) the effect of additional protection against predation accrued by the presence of better-defended mutants in a group that is sizeable. The relationship between conspicuousness and defence is more sharply increasing when the associated level of defence is smaller (see Fig. 4c) since prey must broadcast their aversiveness more strongly to reduce predation. Beyond a certain level of defence further increases in the conspicuousness have diminishing returns on the rate that they are attacked.

Stability in the t-direction

We should add that if (r_1^*, t_1^*) is an equilibrium strategy satisfying (26) then strategy (r_1^*, t_1) with $t_1 < t_1^*$, which includes the origin, is unstable along t. This is attributed to the fact that the LHS of (26) decreases with respect to positive changes in the argument t_1 and since (by assumption) the LHS is zero for $t_1 = t_1^*$ and $r_1 = r_1^*$ it follows that the LHS is positive for values $t_1 < t_1^*$. The argument could be repeated for choices of $t_1 > t_1^*$ in which case the LHS of (26) would be negative by continuity of the functional forms in (10) along t. The interpretation in either case suggests that levels of defence below the equilibrium are at risk of invasion against mutants that are more toxic - LHS of (26) is positive - while levels of defence above the equilibrium are at risk of invasion against less toxic mutants - LHS of (26) is negative. Through inspection of the equilibrium curve we have therefore arrived at the conclusion that the equilibrium curve is stable along t (since levels of defence below equilibrium are invaded by the more toxic types and levels beyond equilibrium are invaded by the less toxic types). We make this claim more formal in the lines that follow.

In Broom et al., (2008) it was shown that such strategies are stable in the t-direction in the sense of (21) for the case $\lambda = 0$. We extend the substitution method found therein in a straightforward manner to establish that it holds for all values of $\lambda \geq 0$. We proceed by considering the cases $t_1 > 0$ and $t_1 = 0$ separately.

A strategy with $t_1 > 0$ is stable along the t-direction if (32) holds in tandem with (26), which through (10) amounts to

$$\begin{aligned} -\frac{\lambda f^2}{q_0 k_0}(1+\exp(-r_1^*))(1+kt_1^*)\exp\left(q\frac{N}{n}\frac{t_1^*-t_c}{1+\exp(-r_1^*)}\right) - f^2 + \frac{2k^2}{(1+kt_1^*)^2} \\ + \frac{2aq\frac{N}{n}}{1+\exp(-r_1^*)}\frac{k}{1+kt_1^*} + \frac{a^2q^2\frac{N^2}{n^2}}{(1+\exp(-r_1^*))^2} > 0. \end{aligned} \tag{32}$$

We set

$$\alpha := \frac{\lambda f}{q_0 k_0}(1+kt_1^*)(1+\exp(-r_1^*))\exp\left(q\frac{N}{n}\frac{t_1^*-t_c}{1+\exp(-r_1^*)}\right) \tag{33}$$

and re-arrange (26) so that

$$\frac{aq\frac{N}{n}}{1+\exp(-r_1^*)} = \alpha + f - \frac{k}{1+kt_1^*}. \tag{34}$$

Condition (32) now amounts to

$$-\alpha f - f^2 + \frac{2k^2}{(1+kt_1^*)^2} + \frac{2k}{1+kt_1^*}\left(\alpha + f - \frac{k}{1+kt_1^*}\right) + \left(\alpha + f - \frac{k}{1+kt_1^*}\right)^2 > 0 \tag{35}$$

and simplifies to the trivial inequality

$$\alpha^2 + \alpha f + \frac{k^2}{(1+kt_1^*)^2} > 0. \tag{36}$$

We have therefore demonstrated that for all values of the parameter $\lambda \geq 0$ strategies on the equilibrium curve with $t_1 > 0$ defined through (26) are stable in the t-direction.

Strategies with $t_1 = 0$ are stable in the t-direction if the equality in (26) is replaced with inequality < 0. Furthermore, since from Broom et al. (2006) it is known that strategies of the form $\{(\rho,\tau) : \rho > 0 \;\; \tau = 0\}$ fail (22) it follows that the origin $\{(0,0)\}$ is the only possibility for a non-toxic strategy to be ESS. The strategy $(r_1, t_1) = (0,0)$ is stable in the t-direction if

$$-\frac{2\lambda f}{q_0 k_0}\exp\left(-q\frac{N}{2n}t_c\right) - f + k + aq\frac{N}{2n} < 0 \tag{37}$$

and it is clear that there is sufficient freedom on the parameters to either satisfy or fail to satisfy the above inequality.

Stability in the r-direction

From looking at conditions (23) and (24) it is clear that the conditions for stability along r are unaffected by the rate of background mortality λ. From a practical standpoint, this is the case because it is convenient for purposes of stability to consider the normalised gradient of the mutant fitness, which factors this dependence out. Going beyond this, we observe that differences in mutant fitness (along r) are associated with differences in the average life-span of prey through influencing the rates of predator detection, recollection and perceived aversiveness (by comparison with the resident appearance); it should be remarked that none of the above are affected by whether the threat of predation is large (i.e. by the value of λ) compared with threats outside of predation. Indeed, a given regime of background mortality applies to both the resident and the mutant and since incremental changes in the fitness of the latter (along r) are unaffected by the value of λ, the prospect of invasion by the latter is also unaffected by the value of λ.

While invasion along r does not depend on the parameter λ it does depend on the local clustering parameter a. This too comes from direct observation of (23) and (24) and admits a sensible remark; larger groups tend to be better recollected by predators that experience their type and further, the larger a group whose appearance deviates from (say, an aversive) resident majority the larger the fitness cost incurred to the mentioned group collectively.

A resident strategy with $r_1 > 0$ is stable in the r-direction if (23) and (24) both hold, which on account of (10) read

$$-\overleftarrow{\partial}_r P(r_1,t_1) \sim \exp(-r_1) - q\frac{N}{n}(t_1 - t_c)\left[\frac{a}{1+\exp(r_1)} + (1-a)v\right] < 0 \tag{38}$$

and

$$\overrightarrow{\partial}_r P(r_1, t_1) \sim -\exp(-r_1) + q\frac{N}{n}(t_1 - t_c)\left[\frac{a}{1+\exp(r_1)} - (1-a)v\right] < 0. \tag{39}$$

The normalised gradient of the mutant fitness along r (corresponding to quantities $-\overleftarrow{\partial}_r P$ and $\overrightarrow{\partial}_r P$ defined above) are referred to collectively as the *invasion fitness gradient* throughout the body of the text. Cryptic strategies are stable in r if (39) holds with $r_1 = 0$. The $\sim$ notation is used to remind readers that the quantities on the LHS of the inequalities are not equal to the derivatives $-\overleftarrow{\partial}_r P$ and $\overrightarrow{\partial}_r P$ but have been scaled by $(\lambda + DKQ)^2/FDKQ$. For purposes of notational convenience - specifically in (40) and (41) - we treat these as equal. **Once more, we emphasize that the incremental difference in fitness experienced by a mutant playing $r = r_1 - h$ with $0 < h \ll 1$ can be approximated by $-h \times \overleftarrow{\partial}_r P(r_1, t_1)$, which is why we retain the minus sign in (38) as well as in (40).**

A linear aversiveness function $H(t)$ as in (10) is both a technically sensible and a biologically plausible choice. As a consequence the LHS of the inequalities in (38) and (39) are linear in t_1, which allows us to express explicitly the toxicity in terms of the conspicuousness and $\overleftarrow{\partial}_r P$ or $\overrightarrow{\partial}_r P$. We have the useful substitutions

$$t_1 = \frac{\left[-(-\overleftarrow{\partial}_r P) + \exp(-r_1)\right](1+\exp(r_1))}{q\frac{N}{n}[a + v \times (1-a)(1+\exp(r_1))]} + t_c =: \boldsymbol{g}_-\big(r_1, -\overleftarrow{\partial}_r P\big) \tag{40}$$

and

$$t_1 = \frac{\left(\overrightarrow{\partial}_r P + \exp(-r_1)\right)(1+\exp(r_1))}{q\frac{N}{n}[a - v \times (1-a)(1+\exp(r_1))]} + t_c =: \boldsymbol{g}_+\big(r_1, \overrightarrow{\partial}_r P\big), \tag{41}$$

which we can utilise in (26) to obtain implicit expressions for the invasion gradient of the mutant along r $-\overleftarrow{\partial}_r P(r_1^*, t_1^*)$ and $\overrightarrow{\partial}_r P(r_1^*, t_1^*)$ at equilibrium. These are

$$\frac{\lambda f}{q_0 k_0}(1 + k\boldsymbol{g}_\mp)(1+\exp(-r_1))\exp\left(q\frac{N}{n}\frac{\boldsymbol{g}_\mp - t_c}{1+\exp(-r_1)}\right) + f - \frac{k}{1+k\boldsymbol{g}_\mp} - \frac{aq\frac{N}{n}}{1+\exp(-r_1)} = 0, \tag{42}$$

which we represent as orange and cyan curves in Figures 2(c), 4(c) and 5(c). The substitution method outlined above is general and especially useful in cases where the equilibrium toxicity cannot be expressed explicitly in terms of the conspicuousness (such as when $\lambda > 0$). However, in cases where $\lambda = 0$ we observe that $\overleftarrow{\partial}_r P^*$ and $\overrightarrow{\partial}_r P^*$ can be evaluated directly by setting t_1 equal to the equilibrium toxicity in the LHSs of (38) and (39), making the above method superfluous.

The parameter v defined in (10) and present in the r-stability conditions above can be understood as the predator's perception of small differences in the visual appearances of prey. We could for all intents and purposes think of this as the time a predator spends investigating a prey animal before deciding to mount an attack. The larger this quantity is the better the predators are at telling apart small differences in the conspicuousness of warning signals (they spend less time investigating it); the smaller this is the worse they are. As we detail, the significance of this term is different for attractive prey with $t_1 < t_c$ than it is for aversive prey with $t_1 > t_c$.

If $t_1 < t_c$ it is easy to observe that (38) cannot be solved for any sensible choice of v. This result is in line with the more general reasoning of Broom et al. (2006), in which it is argued that the conspicuous signalling of strategies that are non-aversive - i.e. drawn from $\{(\rho, \tau) : \rho > 0;\ 0 \leq \tau \leq t_c\}$ - risk invasion from mutations with incrementally smaller signalling component. The same is not true for (39) however, which can be solved for values of v below the threshold on the RHS of (43)

$$v < \frac{\exp(-r_1) + aq\dfrac{N}{n}\dfrac{|t_1 - t_c|}{1 + \exp(-r_1)}}{(1-a)q\dfrac{N}{n}(t_1 - t_c)}. \tag{43}$$

The direction of this inequality demonstrates that for an attractive resident strategy (cryptic) to successfully resist invasion of a more conspicuous mutant, the predator cannot be exceptionally observant, otherwise it would avoid attacking the mutant altogether making the latter comparatively fitter.

For strategies that are aversive i.e. $\{(\rho, \tau) : \rho \geq 0;\ \tau\, t_c\}$ the conditions for stability along r - see (38) and (39) - can be solved for values of v large enough that

$$v > \frac{\left|\exp(-r_1) - aq\frac{N}{n}\dfrac{t_1 - t_c}{1 + \exp(-r_1)}\right|}{(1-a)q\dfrac{N}{n}(t_1 - t_c)}. \tag{44}$$

The direction of the inequality is also justified in this instance, where we would expect an aversive majority of residents to withstand invasion provided the predator is sufficiently observant to detect incremental differences in the conspicuousness. Mutants that look different to a majority of prey that is perceived as aversive pay a price for this and the cost of that decision is magnified by the predator's ability to perceive such differences.

Resident fitness at equilibrium

In this part of the Appendix we discuss how and in which cases the resident fitness can be evaluated at equilibrium. From (10) the resident fitness is given as

$$P_1(r_1, t_1) = \frac{f_0 \exp(-ft_1)}{\lambda + \dfrac{q_0 k_0}{(1 + \exp(-r_1))(1 + kt_1)\exp\left(q\frac{N}{n}\dfrac{t_1 - t_c}{1 + \exp(-r_1)}\right)}}. \tag{45}$$

It is of interest to determine this quantity at equilibrium (26), particularly on account of its (assumed) influence on the direction of prey trajectories in simulated prey populations. There are four cases to consider: (i) $\lambda = 0, a = 0$; (ii) $\lambda = 0, a > 0$; (iii) $\lambda > 0, a = 0$ and (iv) $\lambda > 0, a > 0$.

The method for (i) and (ii) involves solving for the equilibrium toxicity explicitly in terms of the conspicuousness (which is possible) and replacing t_1 in (45) with the equilibrium value. For (i) the equilibrium condition we set $t_1 = 1/f - 1/k$ in (45) to obtain the required result. Likewise for (ii) we set (31) into (45).

For (iii) we proceed by re-arranging (45) so that

$$\frac{q_0 k_0}{(1 + \exp(-r_1))(1 + kt_1)\exp\left(q\dfrac{N}{n}\dfrac{t_1 - t_c}{1 + \exp(-r_1)}\right)} = \frac{f_0 \exp(-ft_1)}{P_1} - \lambda. \tag{46}$$

Substitution of this term into the equilibrium leads to the expression

$$-\frac{\lambda f P_1^*}{f_0 \exp(-f t_1^*) - \lambda P_1^*} - f + \frac{k}{1 + k t_1^*} = 0, \tag{47}$$

which is analogous to

$$t_1^* = \frac{1}{f} - \frac{1}{k} - \frac{\lambda P_1^*}{f f_0} \exp(f t_1^*). \tag{48}$$

The latter can be solved in terms of the principal branch of the *Lambert W-function* (this is such that $W_0(x)\exp(W_0(x)) = x$ provided $x \geq 0$ - the more mathematically-minded reader is encouraged to consult Corless et al., 1996 for an in-depth discussion of the properties and applications of this function). Using a known ansatz we arrive at an explicit expression for the toxicity in terms of the resident fitness

$$t_1^* = \frac{1}{f} - \frac{1}{k} - \frac{1}{f} W_0 \left(\frac{\lambda P_1^*}{f_0} \exp\left(1 - f/k\right) \right) =: G(P_1^*). \tag{49}$$

Swapping t_1^* for $G(P_1^*)$ in equality (26) leads to an implicit expression for the resident fitness and the conspicuousness at equilibrium

$$\frac{\lambda f}{q_0 k_0}(1 + \exp(-r_1^*))(1 + kG(P_1^*)) \exp\left(q \frac{N}{n} \frac{G(P_1^*) - t_c}{1 + \exp\left(-r_1^*\right)} \right) + f - \frac{k}{1 + kG(P_1^*)} = 0. \tag{50}$$

The numerator of (29) is zeroed when

$$t_1 = t_c + \frac{1 + \exp(-r_1)}{q \frac{N}{n}} =: c(r_1), \tag{51}$$

such that the slope of the tangent is positive for points with $t_1 < c(r_1)$ and negative for $t_1 > c(r_1)$. The equilibrium toxicity increases with increasing conspicuousness below the curve $c(r_1)$ and decreases with increasing conspicuousness beyond it. As discussed more extensively in Scaramangas and Broom (2022) and contrary to what prevailing theory contends the relationship of aposematic traits may not need not be an increasing one. As for case (iv) there is (to our knowledge) no way of determining the resident fitness at equilibrium and one may resort to numerical methods to achieve this.

I.C Aversiveness, relatedness & similarity

In Scaramangas and Broom (2022) **as well as in section 2 of this manuscript** it is explained how the description of Broom et al. (2006) can be extended to larger-scale *territorially-divided* habitat structures. The underlying assumption is that the habitat is subdivided into a large number of sites, each containing N prey (of a certain species) and such that each is visited by a group of n predators, who visit one site only. Maintaining the resident-mutant distribution of strategies, we imagine that almost all habitat sites are occupied by resident, except for a small number of sites that contain mutant colonies (making up an effectively negligible proportion of the total number of sites) playing a nearby mutant strategy $(r, t) \in (r_1 - \delta r, r_1 + \delta r) \times (t_1 - \delta t, t_1 + \delta t)$.

Cole (1946) had described *"the most persistent difficulties encountered in ecological field work"* as stemming from the fact that *"...populations of living organisms are very rarely distributed at random over the space available to them."* Cole proceeds to explaining that *" When plants reproduce either vegetatively or by means of seeds there is a tendency for the offspring to be concentrated in the neighbourhood of the parent plant. The same is true of animals which produce their young in litters and especially of the many forms which deposit masses*

of eggs thus temporarily leading to a heavy concentration of individuals within a small area. Most animals furthermore show some tendency toward active congregation." **Quoting the same manuscript we mention** ***"clumping of individuals into groups"*** **such that** ***"each group may be relatively or entirely independent of all similar groups and, therefore, that these distributional units may be randomly distributed"*****. Detailed discussions of the spatial distribution of insect populations can be found in Taylor (1984). It should also be mentioned that amphibian populations form colonies, including the frog species** ***Polypedates leucomystax*** **examined in Roy (1997).**

In one such site we enumerate prey as follows

$$j = \underbrace{1, 2, ..., \overbrace{\ i\ }^{\text{focal}}, ..., \text{nint}(aN)}_{\text{colony}}, \text{nint}(aN+1), ..., N \quad \text{for some} \quad i \in \{1, ..., aN\}, \tag{52}$$

where we clarify that we are implementing the *nearest integer function*, nint(x) described in WolframAlpha (2022), which is arguably beneficial compared with the floor and ceiling functions, as it prevents averaging biases. Individual $j = i$ is identified as the *focal* individual who plays strategy $(r_i, t_i) = (r, t)$ and of which there are aN clones who make up the colony. The remaining prey $j \in \{aN+1, ..., N\}$ are unrelated to the focal individual (i.e. they do not belong to the colony) and play arbitrary strategies $(r_j, t_j) = (r_1, t_1)$. The perceived aversiveness of individual i evaluated through (2) reads

$$\begin{aligned} I_i =& \frac{1}{n} \sum_{j=1, j\neq i}^{N} L(r_j)H(t_j)S(|r_i - r_j|) \\ =& \frac{1}{n} \sum_{j=1, j\neq i}^{\text{nint}(aN)} L(r_i)H(t_i)S(|r_i - r_i|) + \frac{1}{n} \sum_{j=\text{nint}(aN+1)}^{N} L(r_j)H(t_j)S(|r_i - r_j|) \\ \approx& \frac{aN-1}{n} L(r)H(t) + (1-a)\frac{N}{n} L(r_1)H(t_1)S(|r - r_1|). \end{aligned} \tag{53}$$

The contribution of the focal individual can be assumed to be negligible provided $aN \gg 1$, which is a realistic assumption provided the parameter a is not artificially small and the prey population size is large. For a focal individual chosen from a group of mutants this amounts to

$$I := a\frac{N}{n} L(r)H(t) + (1-a)\frac{N}{n} L(r_1)H(t_1)S(|r - r_1|), \tag{54}$$

which is introduced in (5) and is implemented thenceforth. We could repeat the process detailed above for a focal individual chosen from any other site. Since almost all of sites consist of prey playing the resident strategy and since the proportion of mutants in the overall prey population is negligible, we arrive at an approximate expression for the perceived aversiveness of a resident type

$$I_1 := \frac{N}{n} L(r_1)H(t_1). \tag{55}$$

For the remainder of this section we explain the meaning of properties iv) and v) of the similarity function (i.e. that $S'(x) \leq 0$ for all of $x > 0$ and $S'(0) < 0$) and discuss the assumptions about predator generalisation that are implicit in these. Indeed, while conditions i), ii) and iii) are a matter of definition, conditions iv) and v) depend on our interpretation of the underlying predator psychology. We deliberate on this last point presently. In order to do so, we may, for the time being (and without loss of generality in any of the claims that follow)

imagine that predators are (on average) used to encountering individuals resembling i, so that r_i represents some baseline level of appearances. In addition, we consider individual j that is more conspicuous than i (i.e. $0 < r_i < r_j$) such that $|r_i - r_j| =: x_*$, where x_* is sufficiently near the origin (so that condition S being $\mathcal{C}^l$ with $l \geq 2$ holds). Let h be positive and arbitrarily small such that $0 < h \ll 1$. The derivative with respect to x at x^* of the similarity function reads

$$S'(x_*) = \lim_{h \to 0} \frac{S(x_* + h) - S(x_*)}{h} = \lim_{h \to 0} \frac{S(|r_i - (r_j + h)|) - S(|r_i - r_j|)}{|r_i - (r_j + h)|}. \tag{56}$$

Quantity $S'(x_*)$ captures the rate at which the average predator perceives incremental variations in the visual appearance of prey at some "distance" x_* away from the baseline. A basic reading of condition iv) is that as this distance x away from the baseline increases the elevation of S does not increase. Namely that if i and j are assigned some level of similarity $S(x_*)$ then the associated level of S corresponding to an incrementally more conspicuous individual playing $r_j + h$ is not larger than the level associated with r_j. Indeed, condition iv) guarantees that a first-order expansion about x_* in this direction amounts to [3] $S(x_*) + h \times S'(x_*) \leq S(x_*)$, as required. It also follows that small differences from the baseline itself can be determined by evaluating the derivative $S'(x)$ at $x = 0$, which reads

$$S'(0) = \lim_{h \to 0} \frac{S(|(r_i + h) - r_i|) - S(|r_i - r_i|)}{h}. \tag{57}$$

Possible violations of condition v) would include cases with $S'(0) > 0$ or $S'(0) = 0$. The former is immediately rejected as it violates requirement i) - indeed in such a case the elevation of S would grow beyond unity since $S(|(r_i+h)-r_i|) \approx S(0)+h\times S'(0) > 1$. As for the possibility $S'(0) = 0$ - examples could include Gaussian forms on $x \geq 0$ - we remark the following. Since S is bounded from above and from below - condition i) - it is non-increasing within these bounds - condition iv) - and approaches the lower bound for large enough x - condition iii) - it follows that j could be chosen so that $S'(x_*) < 0$, once more maintaining the requirement that S is $\mathcal{C}^l$ with $l \geq 2$ at this value. From (56) it now follows that the change in elevation of S at such an $x = x_*$ can be approximated by $S(|r_i - (r_j + h)|) - S(|r_i - r_j|) \approx h \times S'(0) < 0$. From (57) and the assumption that $S'(0) = 0$ it also follows that the elevation of S does not change at $x = 0$ since $S(|(r_i + h) - r_i|) - S(|r_i - r_i|) \approx h \times S'(0) = 0$.

We have demonstrated that $S'(0) = 0$ is the only potential alternative to v) and that if this were to hold it would imply that predators are (on average) more sensitive to variations in appearance when these occur far from the baseline but not at the baseline itself. Such a conclusion seems to suggest that predators can distinguish small changes in the appearance of prey types that they are not used to encountering but not in the types that they are used to encountering. We might expect that such a result is less relevant for keen-sighted avian predators feeding on Poison dart frogs, to which the model of Broom et al. (2006) is adept (but not limited) to describing. Throughout this manuscript we insist on condition v) and exclude similarity functions that are flat-peaked at the origin from our discussions. The reader is encouraged to consult Balogh and Leimar (2005) for an illustration of the use of flat-peaked generalisation curves - this is done in the context of the evolution of mimicry - and a discussion of restrictions on the shapes that these can assume.

[3] Expansion in the opposite direction can be achieved by considering individual playing $r_j - h$ and yields the reverse inequality, namely $S(x_*) - h \times S'(x_*) \geq S(x_*)$

Appendix II. The simulation model

In this section of the appendix we provide a qualitative description of the simulation and a sample of the actual code for full transparency.

II.A A description of the simulation

Our simulations explicitly model all the individual members of a finite prey population. Individuals will potentially play different strategies, and the performance of individuals will depend on both their own strategy and the distribution of strategies of individuals that they interact with. A similar approach to addressing questions in the evolution of aposematism was taken by Speed and Ruxton (2005), and we further develop their approach. Here we represent evolution by selectively removing individuals from the population and replacing them with versions of other individuals. Prey phenotypes that perform well in the current population are more likely to contribute versions of themselves to the next generation. This mimics the effect of differential fitness in real populations, and is a common approach in evolutionary studies and beyond – often being labelled a genetic algorithm approach (see Ruxton and Beauchamp, 2008, **although we refer to the representation used here as a simulation model**). More generally, individual-based modelling is well established in the study of questions in evolutionary ecology (Zakharova et al., 2019).

The simulation assumes a population of N prey predated by n predators and playing strategies (r_i, t_i) with $i = 1, ..., N$ - to avoid confusion we restrict notations involving the iteration number only to where necessary (see the birth-death process detailed below). The specification of an individual's strategy directly determines the rate at which it reproduces (as $F_i = F(t_i) = f_0 \exp(-ft_i)$), the rate at which it is detected by predators ($D_i = D(r_i) = 1/(1+\exp(-r_i))$) and the **probability** at which a mounted attack results in death ($K_i = K(t_i) = k_0/(1+kt_i)$), as well as the aversiveness of the predator's experience (as $H_i = H(t_i) = t_i - t_c$) and the rate at which such experiences are recollected ($L_i = L(r_i) = 1/(1+\exp(-r_i))$). The specification of such quantities over the population is realised using lists ($1 \times N$ vectors). In contrast, the perceived visual similarity of prey is stored in the $N \times N$ symmetric and unit-diagonal matrix $\mathcal{S}$ defined as

$$(\mathcal{S})_{ij} := \boldsymbol{S}(|\boldsymbol{r_i} - \boldsymbol{r_j}|) = max\left(1 - v\left|r_i - r_j\right|, 0\right) \ \text{ for all } \ i, j = 1, ..., N. \tag{58}$$

The realisation of the i^{th} row of the matrix in (58) specifies the aversiveness $\mathcal{I}_i$ of that prey as perceived by the average predator through the rule

$$\mathcal{I}_i = \frac{1}{n}\left[(aN-1)L(r_i)H(t_i) + \frac{(1-a)N}{N-1}\sum_{j=1, j\neq i}^{N} L(r_j)H(t_j)\boldsymbol{S}(|\boldsymbol{r_i} - \boldsymbol{r_j}|)\right], \tag{59}$$

where the term

$$\frac{1}{N-1}\sum_{j=1, j\neq i}^{N} L(r_j)H(t_j)\boldsymbol{S}(|\boldsymbol{r_i} - \boldsymbol{r_j}|) \tag{60}$$

in (59) is the aversivess of the average prey (excluding the focal individual i). Implicit in (59) is the assumption that (a) **when encountering a prey and calculating its aversiveness $\boldsymbol{\mathcal{I}_i}$, a predators weighs the prey individual it is currently facing as a proportion $\boldsymbol{a}$ of the entire population (independent of phenotype, mutant-status, or even population size), and this happens with every prey that is encountered by a predator in the simulations. The implementation of the local relatedness in the infinite population ESS analysis is different ($\boldsymbol{a}$ is evaluated as a proportion of individuals in the site) and the reader is encouraged to consult our earlier**

discussion in Appendix I.C for a closer comparison of these. (b) Expression (59) represents an average (factor $1/n$) over the predator's experiences of prey and indeed an average over the prey that these encounter (factor $\mathbf{1/(N-1)}$ excludes the focal individual - see related explanation in section 2). The probability Q_i that an attack is mounted on i depends on (59) through $Q(\mathcal{I}_i) = q_0 \exp(-q\mathcal{I}_i)$ and its fitness hence - see (4) - is given as

$$P_i = P(r_i, t_i) = \frac{f_0 \exp(-ft_i)}{\lambda + \dfrac{k_0 q_0}{(1+\exp(-r_i))(1+kt_i)\exp(q\mathcal{I}_i)}}. \tag{61}$$

We should remark that parameter a plays a role in the calculation of the fitness of individuals in the simulation (through Q), which in turn affects the likelihoods of reproduction – however it plays no part in the nature of that reproduction (i.e. in the number of offspring, or the effect of mutation).

The simulation tracks the evolution of traits for a number of distinct prey populations in the following manner. It commences at $m = 0$ where the index $m = 0, 1, 2, ..., M$ specifies the *iteration number* and can be understood as the number of *birth-death* ***events*** that have preceded the population in question (the details of this processes are provided below). After a fixed number of iterations has passed, which is determined by the *averaging frequency g***,** the population traits are averaged and the averaged pair of values is represented as a point in the strategy space of averages. A straight line segment (starting at the initial strategy) is drawn between consecutive points, such that the union of segments forms a *trajectory* for that population. The number of segments making up a population's trajectory is given as M/g. Trajectories of this type are drawn for populations playing a number of distinct starting strategies.

Prey populations succeed one another by means of a birth-death process, whose details are as follows. A small sample of pN prey is selected at random to reproduce and their offspring replace an equally-sized sample. We remark that prey may be selected to reproduce more than once (i.e. give birth to more than one offspring) and are thus considered with multiplicity on the list consisting of parents. It is also possible for the same individual to reproduce and to be replaced (by its own offspring) at the end of the same iteration. The probability that an individual is selected to reproduce l times after pN trials (with replacement) is a binomially-distributed random variable

$$\mathfrak{P}(i \text{ becomes parent } l \text{ times}) = \binom{pN}{l} W_i^l (1 - W_i)^{pN-l}, \text{ for } l = 1, ..., pN \tag{62}$$

where W_i is a comparative measure of fitness defined as

$$W_i := \frac{P_i}{\sum_{j=1}^N P_j}. \tag{63}$$

We should add that for large enough populations we expect the comparative fitness of any one individual to be relatively small and therefore the distribution in (62) to be approximately Poisson distributed with parameter pNW_i. A *generation* can be understood as the average number of iterations (birth-death **events**) required for all the individuals in a population to be replaced. We stress that while alternative interpretations of a generation are possible, from the point of view of the simulation a generation is synonymous with the average number of birth-death processes required for the individuals comprising a certain population to be completely replaced.

Prey traits are subject to random mutation (in the sense that the offspring values can vary continuously within a small margin of error centred at the parent value) and this is encoded into the birth process. We remark that toxicity and conspicuousness are traits determined by common environmental factors (including predation threat and availability of food resources among others) and are likely polygenic, since few phenotypic traits have a **single-gene** origin. Aposematic traits exhibit notable differences depending on the species in question (the genetic origin of traits could provide a possible explanation for this). Furthermore, the specific mode of interaction of one trait with the other is (to our knowledge) mostly unknown. It is therefore the natural option for purposes of simulation to assume that mutation in one trait does not influence mutation in the other (**i.e.** mutation in either trait is independent) versus a more specific (and controversial) assumption about their mode of interaction. In the same spirit, we remark that it is possible for mutation to occur in both traits during a single birth process. To be specific we say that if the offspring of individual i replaces individual j in transitioning from the m^{th} to the $m+1^{st}$ iteration, the probability that either trait is carried through to the offspring is given as 95%. We write

$$\mathfrak{P}\left(r_j^{(m+1)} = r_i^{(m)}\right) = \mathfrak{P}\left(t_j^{(m+1)} = t_i^{(m)}\right) = 0.95, \tag{64}$$

while the probability that any of the traits change is given as

$$\begin{aligned} &\mathfrak{P}\left(r_j^{(m+1)} \in \left[r_i^{(m)} - \delta r, r_i^{(m)}\right) \sqcup \left(r_i^{(m)}, r_i^{(m)} + \delta r\right]\right) = \\ =&\mathfrak{P}\left(t_j^{(m+1)} \in \left[t_i^{(m)} - \delta t, t_i^{(m)}\right) \sqcup \left(t_i^{(m)}, t_i^{(m)} + \delta t\right]\right) = 0.05. \end{aligned} \tag{65}$$

From context it should be clear that the mutation range during the described birth process is precisely the closed rectangle with dimensions $2\delta r \times 2\delta t$ centred at the parent value. As a consequence of independence in trait mutations we also remark that the probability that both parent traits are carried through to the offspring is $0.95^2 \approx 0.9025$, while the probability that both traits change is $0.05^2 = 0.0025$. We should also remark that if a trait changes the step length is chosen uniformly from within the mutation range of the trait in question. For the first trait we write

$$\mathfrak{P}\left(r_j^{(m+1)} \in \delta x\right) = 0.05\frac{\delta x}{2\delta r} \tag{66}$$

to demonstrate the probability that if it increases (or decreases) its precise value is within the interval $\delta x \subset \left(r_i^{(m)}, r_i^{(m)} + \delta r\right]$ or within $\left[r_i^m - \delta r, r_i^{(m)}\right)$ if it decreases.

II.B Code

In this section of the appendix we include a sample of the code that was used in R to run the simulation in Figure 4(d). The plots for the remaining simulations were generated in a similar fashion.

```r
### Figure 4(d) ###


lamda <- 0
a <- 0.5


nprey <- 100
npred <- 10


frac_die <- 0.5 # The fraction of the population dying at the end of each iteration
ndie <- trunc(nprey*frac_die)
```

```
f0 <- 1
d0 <- 1
k0 <- 1
q0 <- 1
f <- 2.8
k <- 5
tc  <- 0.25
q <- 0.4
v <- 1

iterations <- 10000

rmin <- 0
rmax <- 5
rdiff <- rmax-rmin

tmin <- 0
tmax <- 1.5
tdiff <- tmax-tmin

listt <- c(rep(tmin,6), rep(tmin + tdiff/5,5), rep(tmin + 2*tdiff/5,6), rep(tmin + 3*tdiff/5,5),
rep(tmin + 4*tdiff/5,6), rep(tmax,5))

listr <- c(seq(rmin, rmin + 10*rdiff/11, rmin + 2*rdiff/11),
seq(rmin + rdiff/11, rmin + 9*rdiff/11, rmin + 2*rdiff/11),
seq(rmin + (1/3)*rdiff/11, rmin + (10+1/3)*rdiff/11, rmin + 2*rdiff/11),
seq(rmin + (1+1/3)*rdiff/11, rmin + (9+1/3)*rdiff/11, rmin + 2*rdiff/11),
seq(rmin + (2/3)*rdiff/11, rmin + (10+2/3)*rdiff/11, rmin + 2*rdiff/11),
seq(rmin + (1+2/3)*rdiff/11, rmin + (9+2/3)*rdiff/11, rmin + 2*rdiff/11))

progt <- c()
progr <- c()

for(d in 1:length(listt)){

  meant <- c(rep(0, iterations))
  meanr <- c(rep(0, iterations))
  iterrs <- c(rep(seq(1,iterations,1),2))

  progt <- append(progt, listt[d])
  progr <- append(progr, listr[d])

  # Initialise individual prey properties

  P <- c(rep(0,nprey))
  D <- c(rep(0,nprey))
```

```
K <- c(rep(0,nprey))
Q <- c(rep(0,nprey))
I <- c(rep(0,nprey))

t <- rep(listt[d], nprey)
r <- rep(listr[d], nprey)

tstart <- t
rstart <- r

rsum <- 0
tsum <- 0
iters <- 0

print(d)

for (iter in 1:iterations){

  meant[iter] <- mean(t)
  meanr[iter] <- mean(r)

  if(iter%%(iterations/5)<0.5){

    avt <- sum(meant[(iter-(iterations/5-1)):iter])/(iterations/5)
    avr <- sum(meanr[(iter-(iterations/5-1)):iter])/(iterations/5)

    progt <- append(progt, avt)
    progr <- append(progr, avr) }


  # Determine prey fitness

  F <- f0*exp(-(f*t))
  D <- 1/(1+exp(-r))
  K <- k0/(1+(k*t))

  for (x in 1:nprey){
    sum <- 0
    for (y in 1:nprey){
      if ((abs(x-y)) > 0.01){
        L <- D[y]
        H <- t[y] - tc
        S <- 1-(v*(abs(r[x]-r[y])))
        if (S<0){S<-0}
        sum <- sum +(L*H*S)}}
      L_self <- D[x]
      H_self <- t[x] - tc
      S_self <- 1

      I[x] = (((nprey*(1-a)/(nprey-1))*sum)+((a*nprey-1)*L_self*H_self*S_self))/npred}

    Q <- q0*exp(-q*I)
    for (z in 1:length(Q)){
      if (Q[z]>1){Q[z]<-1}}

    P <- F/(lamda+(D*K*Q))


    # Update population w.r.t. fitness

    wts_reproduce   <- P/sum(P)

    wts_die <- 1-wts_reproduce
    sumwts <- sum(wts_die)
    wts_die <- wts_die/sumwts

    parents <- sample(nprey, size = ndie, replace = TRUE, prob = wts_reproduce)
    replaced <- sample(nprey, size = ndie, replace = FALSE, prob = wts_die)

    copyr <- r
    copyt <- t

    for (x in 1:ndie){
      p1 <- parents[x]
      q1 <- replaced[x]
      mutr <- rbinom(1,1,0.05)
      if(mutr>0.5){
      r[q1] <- copyr[p1]+((-0.05)+(0.1*runif(1)))
        if(r[q1]<0){r[q1] <- 0}}
      else{r[q1] <- copyr[p1]}
      mutt <- rbinom(1,1,0.05)
      if(mutt>0.5){
        t[q1] <- copyt[p1]+((-0.05)+(0.1*runif(1)))
        if(t[q1]<0){t[q1] <- 0}}
      else{t[q1] <- copyt[p1]}} # End of change
      }
# End of iteration
}


# Produce plot for Figure 4(d)
```

```
length(progr)
length(progt)

lines <- rep(1:33, each=6)
lines <- as.factor(lines)

idealfig <- data.frame(progr, progt, lines)

markersr <- progr[seq(0, length(idealfig$progr), 6)]
markerst <- progt[seq(0, length(idealfig$progt), 6)]

markers <- data.frame(markersr, markerst)

plot4d <- ggplot(idealfig, aes(x=progr, y=progt)) +
  geom_path(aes(group=lines, col=lines), size=1.05) +
  theme(legend.position="none", axis.text.x = element_text(size=20),
  axis.text.y = element_text(size=20))
  + xlab("") + ylab("") +
  geom_point(data=markers, mapping=aes(x=markersr, y=markerst))+
  scale_x_continuous(breaks=seq(0,4,1))+
  scale_y_continuous(breaks=seq(0,1.2,0.3))

plot4d

# End of code
```

Acknowledgements

This research would not have been possible without the generous *EPSRC* stipend; we are very grateful for the awarding of this studentship to Alan Scaramangas. We would also like to thank Prof. Jeremy Van Cleve and two anonymous reviewers for reading through the manuscript and providing us with detailed insight that have helped us improve its content and presentation. We are also grateful to the award jury at the 19^{TH} *International Symposium on Dynamic Games and Applications* for awarding the paper with the Thomas L. Vincent Best Paper Award in Porto (July 2022).

References

Balogh, A. C., & Leimar, O. (2005). Müllerian mimicry: An examination of fisher's theory of gradual evolutionary change. *Proceedings of the Royal Society B: Biological Sciences*, *272*(1578), 2269–2275.

Broom, M., Ruxton, G. D., & Speed, M. P. (2008). Evolutionarily stable investment in anti-predatory defences and aposematic signalling. In *Mathematical modeling of biological systems, volume ii* (pp. 37–48). Springer.

Broom, M., Speed, M. P., & Ruxton, G. D. (2006). Evolutionarily stable defence and signalling of that defence. *Journal of theoretical biology*, *242*(1), 32–43.

Caro, T. (2005). *Antipredator defenses in birds and mammals*. University of Chicago Press.

Cole, L. C. (1946). A theory of analyzing contagiously distributed populations. *Ecology*, *27*(4), 329–341.

Corless, R. M., Gonnet, G. H., Hare, D. E., Jeffrey, D. J., & Knuth, D. E. (1996). On the lambertw function. *Advances in Computational mathematics*, *5*(1), 329–359.

Darst, C. R., Cummings, M. E., & Cannatella, D. C. (2006). A mechanism for diversity in warning signals: Conspicuousness versus toxicity in poison frogs. *Proceedings of the National Academy of Sciences*, *103*(15), 5852–5857.

Endler, J. A. (1978). A predator's view of animal color patterns. In *Evolutionary biology* (pp. 319–364). Springer.

Hämäläinen, L., Mappes, J., Thorogood, R., Valkonen, J. K., Karttunen, K., Salmi, T., & Rowland, H. M. (2020). Predators' consumption of unpalatable prey does not vary as a function of bitter taste perception. *Behavioral Ecology*, *31*(2), 383–392.

Leimar, O., Enquist, M., & Sillen-Tullberg, B. (1986). Evolutionary stability of aposematic coloration and prey unprofitability: A theoretical analysis. *The American Naturalist*, *128*(4), 469–490.

Maan, M. E., & Cummings, M. E. (2012). Poison frog colors are honest signals of toxicity, particularly for bird predators. *The American Naturalist*, *179*(1), E1–E14.

Mappes, J., Marples, N., & Endler, J. A. (2005). The complex business of survival by aposematism. *Trends in ecology & evolution*, *20*(11), 598–603.

McLean, I. (2011). *The relationship between chemical defence and death feigning in the red flour beetle (tribolium castaneum)* (Doctoral dissertation). Carleton University.

Pointer, M. D., Gage, M. J., & Spurgin, L. G. (2021). Tribolium beetles as a model system in evolution and ecology. *Heredity*, *126*(6), 869–883.

Poulton, E. B. (1890). *The colours of animals: Their meaning and use, especially considered in the case of insects*. D. Appleton.

Rojas, B., Nokelainen, O., & Valkonen, J. K. (2021). Aposematism. In *Encyclopedia of evolutionary psychological science* (pp. 345–349). Springer.

Rowland, H. M., Ruxton, G. D., & Skelhorn, J. (2013). Bitter taste enhances predatory biases against aggregations of prey with warning coloration. *Behavioral Ecology*, *24*(4), 942–948.

Roy, D. (1997). Communication signals and sexual selection in amphibians. *Current science*, 923–927.

Ruxton, G. D., Allen, W. L., Sherratt, T. N., & Speed, M. P. (2019). *Avoiding attack: The evolutionary ecology of crypsis, aposematism, and mimicry*. Oxford university press.

Ruxton, G. D., & Beauchamp, G. (2008). The application of genetic algorithms in behavioural ecology, illustrated with a model of anti-predator vigilance. *Journal of theoretical biology*, *250*(3), 435–448.

Ruxton, G. D., Speed, M. P., & Broom, M. (2009). Identifying the ecological conditions that select for intermediate levels of aposematic signalling. *Evolutionary ecology*, *23*(4), 491–501.

Santos, J. C., & Cannatella, D. C. (2011). Phenotypic integration emerges from aposematism and scale in poison frogs. *Proceedings of the national academy of sciences*, *108*(15), 6175–6180.

Santos, J. C., Coloma, L. A., & Cannatella, D. C. (2003). Multiple, recurring origins of aposematism and diet specialization in poison frogs. *Proceedings of the National Academy of Sciences*, *100*(22), 12792–12797.

Scaramangas, A., & Broom, M. (2022). Aposematic signalling in prey-predator systems: Determining evolutionary stability when prey populations consist of a single species. *Journal of Mathematical Biology*, *85*(2), 1–35.

Speed, M. P., & Ruxton, G. D. (2005). Warning displays in spiny animals: One (more) evolutionary route to aposematism. *Evolution*, *59*(12), 2499–2508.

Summers, K., Speed, M. P., Blount, J., & Stuckert, A. (2015). Are aposematic signals honest? a review. *Journal of evolutionary biology*, *28*(9), 1583–1599.

Summers, K., & Clough, M. E. (2001). The evolution of coloration and toxicity in the poison frog family (dendrobatidae). *Proceedings of the National Academy of Sciences*, *98*(11), 6227–6232.

Tarvin, R. D., Borghese, C. M., Sachs, W., Santos, J. C., Lu, Y., O'connell, L. A., Cannatella, D. C., Harris, R. A., & Zakon, H. H. (2017). Interacting amino acid replacements allow poison frogs to evolve epibatidine resistance. *Science*, *357*(6357), 1261–1266.

Taylor, L. (1984). Assessing and interpreting the spatial distributions of insect populations. *Annual review of entomology*, *29*(1), 321–357.

Teichmann, J., Broom, M., & Alonso, E. (2014). The evolutionary dynamics of aposematism: A numerical analysis of co-evolution in finite populations. *Mathematical Modelling of Natural Phenomena*, *9*(3), 148–164. https://doi.org/10.1051/mmnp/20149310

Teichmann, J., Alonso, E., & Broom, M. (2015). A reward-driven model of darwinian fitness. *2015 7th International Joint Conference on Computational Intelligence (IJCCI)*, *1*, 174–179.

Vences, M., Kosuch, J., Boistel, R., Haddad, C. F., La Marca, E., Lötters, S., & Veith, M. (2003). Convergent evolution of aposematic coloration in neotropical poison frogs: A molecular phylogenetic perspective. *Organisms Diversity & Evolution*, *3*(3), 215–226.

Wallace, A. R. (1877). The colors of animals and plants. *The American Naturalist*, *11*(11), 641–662.

Wang, I. J. (2011). Inversely related aposematic traits: Reduced conspicuousness evolves with increased toxicity in a polymorphic poison-dart frog. *Evolution: International Journal of Organic Evolution*, *65*(6), 1637–1649.

WolframAlpha. (2022). Nearest integer function. https://mathworld.wolfram.com/NearestIntegerFunction.html

Zakharova, L., Meyer, K., & Seifan, M. (2019). Trait-based modelling in ecology: A review of two decades of research. *Ecological Modelling*, *407*, 108703.