

Adaptive Processes in Biology and Culture

Models of Evolving Vaccine Resistance and the Record Statistics of Innovation

by

Simon Rella

September, 2025

*A thesis submitted to the
Graduate School
of the
Institute of Science and Technology Austria
in partial fulfillment of the requirements
for the degree of
Doctor of Philosophy*

Committee in charge:

Tim Browning, Chair

Gašper Tkačik

Fyodor Kondrashov

Björn Hof

Bernat Corominas-Murtra



The thesis of Simon Rella, titled *Adaptive Processes in Biology and Culture*, is approved by:

Supervisor: Gašper Tkačik, ISTA, Klosterneuburg, Austria

Signature: _____

Co-supervisor: Fyodor Kondrashov, Okinawa Institute of Science and Technology, Onna, Japan

Signature: _____

Committee Member: Björn Hof, ISTA, Klosterneuburg, Austria

Signature: _____

Committee Member: Bernat Corominas-Murtra, University of Graz, Graz, Austria

Signature: _____

Defense Chair: Tim Browning, ISTA, Klosterneuburg, Austria

Signature: _____

Signed page is on file

© by Simon Rella, September, 2025

CC BY-NC-ND 4.0 The copyright of this thesis rests with the author. Unless otherwise indicated, its contents are licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License. Under this license, you may copy and redistribute the material in any medium or format. This is on the condition that: you credit the author, do not use it for commercial purposes and do not distribute modified versions of the work.

ISTA Thesis, ISSN: 2663-337X

I hereby declare that this thesis is my own work and that it does not contain other people's work without this being so stated; this thesis does not contain my previous work without this being stated, and the bibliography contains all the literature that I used in writing the dissertation.

I accept full responsibility for the content and factual accuracy of this work, including the data and their analysis and presentation, and the text and citation of other work.

I declare that this is a true copy of my thesis, including any final revisions, as approved by my thesis committee, and that this thesis has not been submitted for a higher degree to any other university or institution.

I certify that any republication of materials presented in this thesis has been approved by the relevant publishers and co-authors.

Signature: _____

Simon Rella
September, 2025

Signed page is on file

Abstract

This thesis is organized into two parts, each comprising two chapters: Chapter 1&2 offer models for the evolution of vaccine resistance in response to diverse vaccination strategies. Chapter 3&4 review the statistics of records, their connection to models of innovation and an application to the cultural evolution of sports.

In chapter 1 we present a modelling study from 2021 on the evolution of SARS-CoV-2. At that time the vaccine-resistant Omicron variant had not yet evolved. In our model we consider a population that is becoming vaccinated over time, while a pathogen is spreading in the population and eventually becoming resistant to the vaccine. We explore effective pharmaceutical and non-pharmaceutical interventions to prevent the emergence of vaccine resistance.

In chapter 2 we model a particular set of complex vaccination strategies, mosaic and pyramid vaccination, where an immunologically diverse portfolio of vaccines is considered. We find that a bet-hatching strategy, in which vaccine types are distributed in the population, is effective at hindering the evolution of vaccine resistance if mutation rates are high.

In chapter 3 we switch gears and present a review on the statistics of records. We highlight similarities and analogies to other models in the fields of statistical physics, evolution and innovation. This offers interesting complimentary perspectives on well-known models.

In chapter 4 we apply models of record statistics and innovation to study cultural evolution in sport. We propose a model of sport evolution that combines deterministic improvements in performance and stochastic bursts of improvements due to innovation.

Acknowledgements

I want to express my gratitude to Gašper and Fedya for being both fantastic supervisors and empathetic colleagues. I would also like to thank my colleagues at ISTA, many of who have become friends over these years, as well as the THINK conference community. I am grateful to my partner, Dorina, and I thank Rax for always being there.

About the Author

Simon Rella completed his BSc in Astrophysics and Physics, followed by a master in Physics at the University of Vienna. In September 2019 he joined ISTA. His research revolves around evolution in complex systems. Initially coming to ISTA with the idea of study ants, he affiliated with the Kondrashov and Tkačik group to investigate bridges between evolutionary biology and statistical physics. During the unfortunate Covid19 pandemic Simon actively participated in the scientific discourse by developing models of evolving vaccine resistance. After this he focused on developing models that combine the mathematical field of record statistics with adaptive evolution in biology and culture. Simon presented his work at public health and complex systems conferences. He organized a puzzle hunt for the campus community and the yearly THINK conference for interdisciplinary science.

List of Collaborators and Publications

The following publications are featured in full in this thesis (chapter 1&2).

1. Simon A Rella, Yuliya A Kulikova, Emmanouil T Dermitzakis, and Fyodor A Kondrashov. Rates of sars-cov-2 transmission and vaccination impact the fate of vaccine-resistant strains. *Scientific Reports*, 11(1):15729, 2021
2. Simon A Rella, Yuliya A Kulikova, Aygul R Minnegalieva, and Fyodor A Kondrashov. Complex vaccination strategies prevent the emergence of vaccine resistance. *Evolution*, 78(10):1722–1738, 2024

Table of Contents

Abstract	vii
Acknowledgements	viii
About the Author	ix
List of Collaborators and Publications	x
Table of Contents	xi
1 Introduction	1
2 Transmission and Vaccination Impact the Fate of Vaccine-resistant Strains	3
2.1 Introduction	3
2.2 Results	4
2.3 Discussion	8
2.4 Methods	10
3 Complex Vaccination Strategies & the Emergence of Vaccine Resistance	15
3.1 Introduction	15
3.2 Methods	16
3.3 Results	23
3.4 Discussion	32
4 Records, Evolution and Innovation	37
4.1 Introduction	37
4.2 Record Statistics	39
4.3 Models of Evolution and Innovation	45
4.4 Generalized Models of Innovation	54
4.5 Applications and Outlook	59
5 Sport as a Model of Cultural Evolution	63
5.1 Introduction	63
5.2 Results	64
5.3 Discussion	72
5.4 Methods	75
Bibliography	81

CHAPTER 1

Introduction

Understanding how systems evolve under constraints of chance and selection is a central challenge across biology, culture, and society. This thesis examines two domains where evolutionary principles govern the dynamics of change: the emergence of vaccine resistance and the statistics of records. At first glance these appear to be disparate fields – epidemiology and innovation – yet they are connected by a shared aim: to quantify how the past constrains and informs the future of evolving systems.

In the epidemiological setting, the focus is on the co-evolution of pathogens and human societies. Pathogens adapt on time scales that overlap with those of social responses, such as vaccination campaigns, non-pharmaceutical interventions, or behavioral shifts. The SARS-CoV-2 pandemic highlighted how this interplay unfolds in real time: new variants with increased transmissibility or partial vaccine escape appear in bursts, while societies counter with changing policies and novel vaccines [KCGP06]. This coupling of biological and cultural evolution drives a dynamic arms race, where predicting outcomes requires combining models of infection, mutation, and human adaptation.

In record statistics and models of innovation, the emphasis lies on cultural and technological domains where progress is punctuated by breakthroughs. Records, whether in sports, technology, or science, embody innovations that reset the baseline of performance. The process is evolutionary in nature: long periods of incremental change are disrupted by sudden advances, akin to the emergence of highly fit pathogen variants. These bursts generate heavy-tailed distributions of waiting times and step sizes, reminiscent of the power laws observed in pathogen phylogenies [NRS14] and epidemic sizes [MKPP21].

Both fields grapple with systems in which rare but consequential innovations determine future trajectories. Both must contend with stochasticity at the level of emergence and determinism at the level of establishment. And both are shaped by cultural evolution: in epidemiology, the co-evolution of human responses with viral adaptation; in records, the viral diffusion of ideas, technologies, and training methods across society.

Short summary of the following chapters

Chapter 2. Transmission and vaccination impact the fate of vaccine-resistant strains

This chapter develops a model to study how resistant variants of a pathogen can emerge during a vaccination campaign. By combining stochastic introductions of new variants with

large-scale epidemic dynamics, the analysis shows that the timing of emergence is critical: the risk of resistant strains becoming established peaks when vaccination is already well underway but not yet complete. The results highlight how continued interventions late in a rollout, even when much of the population is protected, can strongly reduce the chances of resistant strains taking hold.

Chapter 3. Complex vaccination strategies and the emergence of resistance The third chapter explores how different ways of designing and distributing vaccines influence the evolutionary response of pathogens. Beyond uniform vaccination, strategies that diversify which variants or epitopes are targeted can change the likelihood that resistant strains evolve and spread. The findings suggest that “mosaic” vaccination strategies – where populations receive slightly different vaccines – may provide an evolutionary advantage by making it harder for resistant strains to gain a foothold. This work emphasizes that vaccination policy is not only about protecting individuals, but also about shaping long-term evolutionary outcomes.

Chapter 4. Records, evolution, and innovation Here, the focus shifts from epidemiology to the study of records as a general model for innovation. We review existing literature on the statistics of records, the sample space reducing process, non-parametric Bayesian methods and innovation models. We show that these distinct fields bear rich connections, through the concept of the adjacent possible, and that they can sometimes be seen as two sides of the same coin. We argue that record statistics may serve as a simple but powerful lens through which to understand how innovation unfolds across different cultural and biological systems.

Chapter 5. Sports as a model of cultural evolution The final chapter turns to the empirical case of sports as a window into cultural evolution. Athletic performances, techniques, and technologies evolve through a process of competition, imitation, and innovation. By analyzing records and performance trajectories, this chapter identifies how breakthroughs spread across disciplines and how the “heritability” of performance reflects shared cultural practices. We apply methods developed in chapter 4 to study statistical signatures of innovation in world record data. Sports provide a rich dataset for studying cultural evolution, where performance can be treated as a measurable form of fitness and innovation can be tracked across time.

Transmission and Vaccination Impact the Fate of Vaccine-resistant Strains

Published Work This chapter appeared in full as Simon A Rella, Yuliya A Kulikova, Emmanouil T Dermitzakis, and Fyodor A Kondrashov. Rates of sars-cov-2 transmission and vaccination impact the fate of vaccine-resistant strains. *Scientific Reports*, 11(1):15729, 2021. At the time of writing the SARS-CoV-2 pandemic was far from being under control and the text should be read with this context in mind.

Author Contributions E.T.D. and F.A.K. formulated the question. S.A.R. and Y.A.K. formulated the model. S.A.R. programmed and ran the model. All authors jointly analyzed the results and wrote the manuscript.

Abstract Vaccines are thought to be the best available solution for controlling the ongoing SARS-CoV-2 pandemic. However, the emergence of vaccine-resistant strains may come too rapidly for current vaccine developments to alleviate the health, economic and social consequences of the pandemic. To quantify and characterize the risk of such a scenario, we created a SIR-derived model with initial stochastic dynamics of the vaccine-resistant strain to study the probability of its emergence and establishment. Using parameters realistically resembling SARS-CoV-2 transmission, we model a wave-like pattern of the pandemic and consider the impact of the rate of vaccination and the strength of non-pharmaceutical intervention measures on the probability of emergence of a resistant strain. As expected, we found that a fast rate of vaccination decreases the probability of emergence of a resistant strain. Counterintuitively, when a relaxation of non-pharmaceutical interventions happened at a time when most individuals of the population have already been vaccinated the probability of emergence of a resistant strain was greatly increased. Consequently, we show that a period of transmission reduction close to the end of the vaccination campaign can substantially reduce the probability of resistant strain establishment. Our results suggest that policymakers and individuals should consider maintaining non-pharmaceutical interventions and transmission-reducing behaviours throughout the entire vaccination period.

2.1 Introduction

Vaccines are among the most effective public health measures against infectious disease [A⁺08]. Their track record brings hope that SARS-CoV-2 may soon be under control [AVTC20] as a

consequence of a plethora of vaccine development efforts [P⁺20, L⁺21, B⁺21, D⁺20b, Gra20, M⁺20b]. A potential cause of concern is the low rate of vaccine production and administration [Our21] coupled with reports of new strains with higher transmission rates [D⁺21, T⁺21, F⁺21] and even with potential for some degree of vaccine resistance [C⁺21, A⁺20b, WB21, B⁺20]. A number of models considered the dynamics of the spread of a vaccine-resistant strain in the population [Luo19, HK13, FH20, XM13]. However, to our knowledge, the interplay of the population vaccination rate with the stochastic dynamics of emergence of a resistant strain has been discussed [VD18], but not formally modeled. Specifically, a concern is whether a combination of vaccination and transmission rates can create positive selection pressure on the emergence and establishment of resistant strains [RSA16, G⁺04]. To address this issue, we implemented a model to simulate the probability of emergence of a resistant strain as a function of vaccination rates and changes in the rate of virus transmission, resembling those caused by non-pharmaceutical interventions and behavioural changes. We then performed a number of simulations based on realistic parameters to study the likelihood and pattern of the emergence of a resistant strain. Finally, we considered possible countermeasures to reduce the probability of the establishment of the resistant strain in the population.

2.2 Results

We implemented a modification of a SIR model [HK13, BSKA20] that included additional states to study the interplay of the rate of vaccination, rate of transmission and the likelihood of emergence of resistant strains (Fig. 2.1a). In addition to other states, individuals could be vaccinated (V), infected by the resistant strain (I_r), or simultaneously be vaccinated and infected with the resistant strain ($I_r V$). The model was run to simulate a population of 10,000,000 individuals over 3 years with the vaccination starting after the first year. In the model, the susceptible individuals (S) are infected by the wildtype, or the original, strain at a rate of β and infected individuals recover at a rate of γ or die at a rate of δ . At each time step, a fraction θ of all non-infected individuals is vaccinated and with some fixed probability p , an infected individual becomes infected with a resistant strain. Conversely, any individual infected with the resistant strain can revert back to the wildtype strain population with the same probability, p . Immunity acquired after infection decayed at a rate of μ . Overall, our model included eight character states and six transition parameters between them (Fig. 2.1a, Table 2.1).

The rate of transmission in the course of a pandemic is typically cyclical [BV16, Mar18, KTG⁺20] due to government interventions [KTG⁺20, H⁺20], behavioural changes [AHD14, WMRA20, VB⁺20], and environmental [MBVN20, CCH⁺21] and other factors [R⁺21, A⁺20a]. Generally, the number of infected individuals is wave-like, guided by periods of high rate of transmission, followed by periods of a low rate of transmission [C⁺20b, BV16, Mar18, CCS20]. We thus varied β , the rate of virus transmission to reflect this cyclical behavior (Fig. 2.1b,c). A high rate of transmission ($\beta_h = 0.18$, equivalent to the effective reproduction number of $R_h = 2.52$) was alternated with a low rate ($\beta_l = 0.055$ or $R_l = 0.77$), which broadly reflected the observed rates of transmission in various countries affected by the SARS-CoV-19 pandemic with and without lockdown measures, respectively [C⁺20b, HK20, ATS20]. The low rate of transmission was triggered in the model when the number of individuals infected with any strain reached a high threshold $F_h = (I_{wt} + I_r + I_r V)$. We considered two modes of transition from a low rate of transmission back to a high rate, when the number of infected individuals reached a low threshold F_l at a constant value of $F_l = (I_{wt} + I_r + I_r V) = 1000$, which was used to generate the main figures, and at a relative value of $F_l = F_h/8$, which are available as

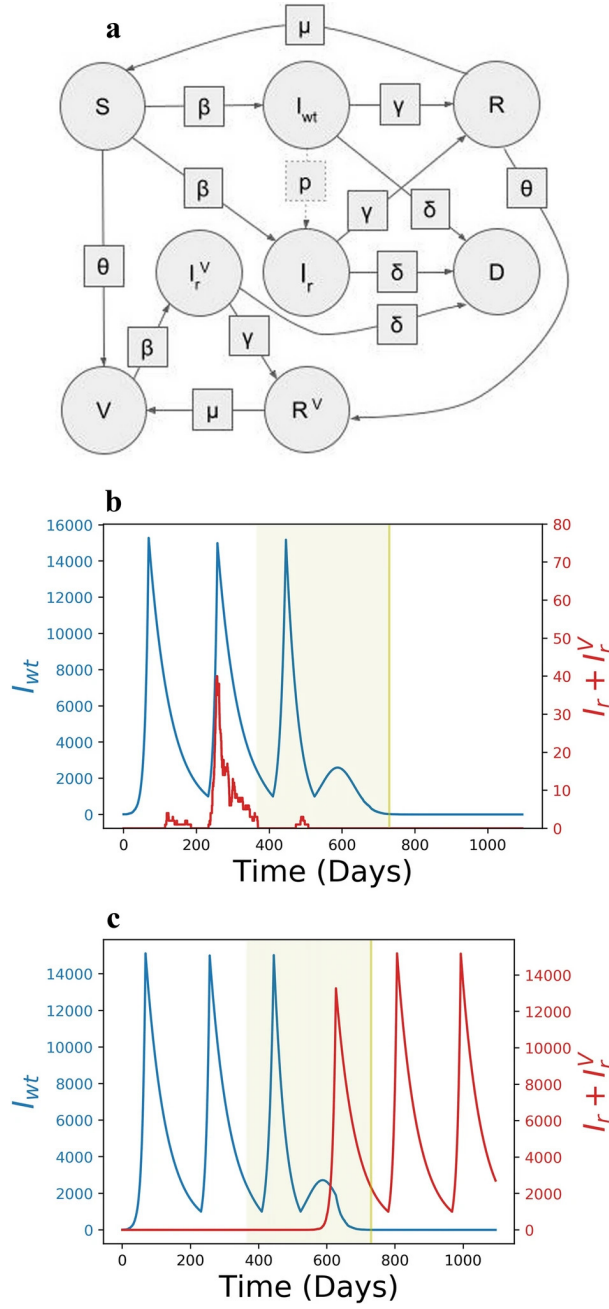


Figure 2.1: The states, transition parameters and dynamics. (a) States are shown in circles and transition parameters in squares. The transition parameter, μ , is the rate at which individuals lose natural immunity and p , is the probability that an individual infected with the wildtype strain transmits a resistant strain, so it is not a deterministic parameter. Example dynamics of the number of individuals infected with the wildtype (blue) and resistant strains (red) for $p = 10^{-6}$, $\theta_0 = 1/365$ and $F_h = 15,000$. The period of vaccination is highlighted (green). Under the same parameters the resistant strain may emerge and go extinct, (b), or become established, (c).

“Supplementary materials S1” in [RKDK21]. The explored values of F_h and F_l were selected such that the number of infection waves during the first year of the model roughly coincided with the number of waves of SARS-CoV-2 infection observed in most countries during the first year of the pandemic.

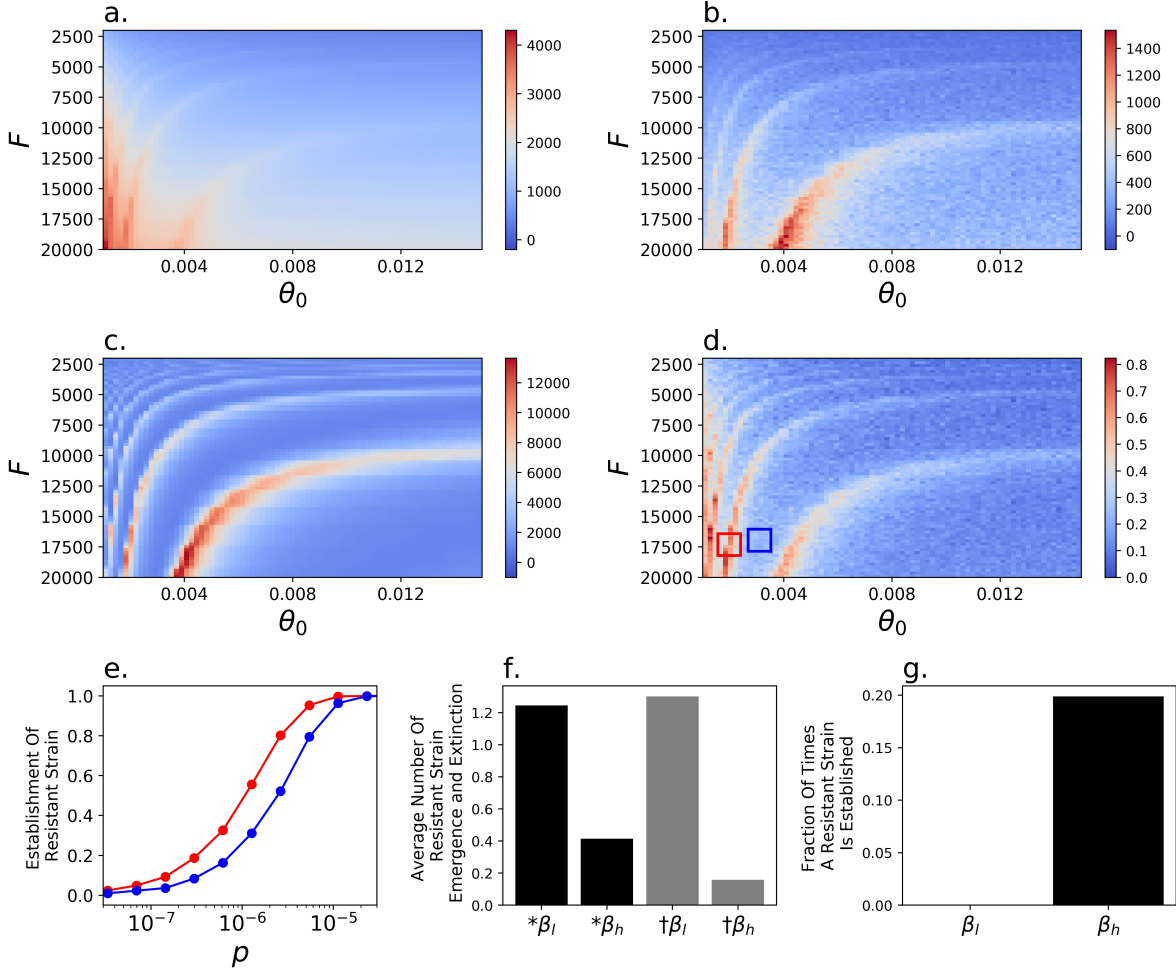


Figure 2.2: Impact of the rate of vaccination, θ , and the initiation of low rate of transmission, F_h , on model dynamics. The cumulative death rate from the (a) wildtype and (b) resistant strains, (c) the number of wildtype- strain infected individuals at t_{V60} , the point in time when 60% of the population is vaccinated and (d) the probability of resistant strain establishment, for $p = 10^{-6}$. (e) The probability of emergence of the resistant strain as a function of the probability of emergence, p shown for the parameter ranges of θ and F_h in the corresponding red and blue boxes from figure (d). (f) The average number of times of 8×10^6 simulation runs during which a resistant strain emerges (black) or goes extinct (grey) during periods of low (β_l) or high (β_h) transmission for $p = 10^{-6}$. (g) A resistant strain was never observed to establish during periods of low transmission (β_l) for $p = 10^{-6}$.

SIR-like models frequently consider only deterministic dynamics [HK13]. However, the emergence of a new strain is an inherently stochastic process under extensive influence of genetic drift [CC10, Hal27]. Therefore, we incorporated a stochastic stage into our model to allow for genetic drift in the early phases of population dynamics of the resistant strains. The growth rate of the number of individuals infected with the wildtype strain at time t was determined deterministically by $(\beta S/N - \gamma - \delta)I_{wt}$. By contrast, when the frequency of the resistant strain in the population is low, the number of transmissions of the resistant strain was drawn from a Poisson distribution with a mean of $(\beta I_r(S + V)/N)dt$ [Gil01]. However, when the frequency of the resistant strain is greater than 1000 individuals (0.01% of the population), making it highly unlikely to disappear by stochastic forces, the dynamics are treated deterministically in

the same manner as the wildtype strain.

We define three stages of vaccine-resistant strain propagation, including emergence, the appearance of the first individual with the infected strain, establishment, the time point when the number of infected individuals reached 1000, and extinction, when the number of resistant strain infected individuals returned to zero. The impact of three parameters on the resistant strain propagation were explored: the probability of the emergence of the resistant strain (p), the speed of vaccination (θ) and the initiation of periods of lower rate of transmission (F_h). All other parameters were constants and their values were chosen to be broadly reflecting a realistic set of parameters that approximate the available data for the SARS-CoV-2 pandemic, $\mu = 1/180$, $\gamma = 0.99 * 1/14$ and $\delta = 0.01 * 1/14$ (see Table 2.1).

Higher probability of the initial emergence of a resistant strain emerging in a single individual, p , had a predictably [ES05] positive effect on the probability of the establishment of the resistant strain, but depended on the rate of vaccination (θ) and low transmission initiation (F_h) in a complex manner (Fig. 2.2). The dependency was periodic and the probability of establishment of the resistant strain was different by a factor of two even for some similar values of θ and F_h (Fig. 2.2d,e).

The behaviour of the emergence, establishment and extinction of the resistant strain in the population bears striking resemblance to the population genetics problem addressing the survival of a beneficial allele in a growing population [UH11, Esh84]. To understand the stochastic dynamics of the resistant strain in the model, it is therefore instructive to consider the underlying mechanism in population genetics terms. Unless the rate of mutation is zero, or infinitesimally low, new variants will emerge in the population at a rate of $p \cdot I_{wt}$. When the rates of transmission of the wildtype and the resistant strain are equal, the probability that the resistant strain will go extinct, $1 - Q_t$, can be approximated by

$$1 - Q_t = e^{-Q_{t+1}(1+s)R_t^r} \quad (2.1)$$

where s is the selection advantage of the resistant strain [CC10] [OW97] and $R_t^r = (S + V)\beta/(N(\gamma + \delta))$, or the rate of population growth. Therefore, even when there is no selective advantage of the resistant strain ($s = 0$) over the wildtype strain but the rate of transmission is high ($R_t^r > 1$), the likelihood that a new mutation is lost from the population is small (roughly 10% for $\beta_h = 0.18$, or $R_0 = 2.52$). By contrast, when the rate of transmission is low ($R_t^r < 1$, which is the case during the low transmission periods in the model), the probability of extinction of the resistant strain by genetic drift is substantial [CC10, Hal27]. The results of our model are consistent with theory, such that the resistant strain emerges during periods of high and low transmission rates, but goes extinct with higher probability during periods of low transmission (Fig. 2.2f). Furthermore, under the parameters of our model the resistant strain becomes established only when the rate of transmission is high (Fig. 2.2f).

The complex influence of the speed of vaccination, θ , and initiation of low transmission period, F_h , on the dynamics of establishment of the resistant strain (Fig. 2.2d) is, therefore, likely driven by the overlap of the vaccination period and the periodicity of the cycles of the number of infected individuals driven by the interaction of F_h and θ (Fig. 2.1b,c). The coincidence of a high number of vaccinated individuals and a high rate of transmission has two effects on the resistant strain. First, as mentioned previously, because the rate of transmission is high, the emerging resistant strain is not lost through genetic drift (see also [UH11, OU08]). Second, a high number of vaccinations creates a selective advantage of the resistant strain over the

wildtype strain [G⁺04]. The effective reproductive number of the wildtype versus the resistant strains, R_t^{wt}/R_t^r , is $(V + S)/S$, which is the selective advantage $1 + \sin$ Eq. 2.1. Thus, when V is large the resistant strain has a growth advantage over the wildtype strain, contributing to its establishment in the population towards the end of the vaccination campaign. Taken together, the highest probability for establishment of the resistant strain for a given p is reached when V , I_{wt} and β (and the corresponding R_t^r) are large (Fig. 2.2c, Eq. 2.1).

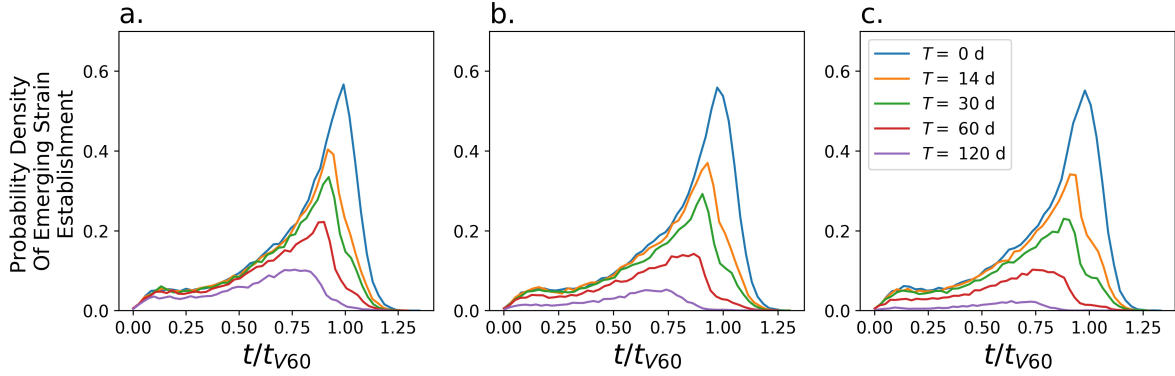


Figure 2.3: Impact of the rate of vaccination, θ , and the initiation of low rate of transmission, F_h , on model dynamics. The cumulative death rate from the (a) wildtype and (b) resistant strains, (c) the number of wildtype- strain infected individuals at t_{V60} , the point in time when 60% of the population is vaccinated and (d) the probability of resistant strain establishment, for $p = 10^{-6}$. (e) The probability of emergence of the resistant strain as a function of the probability of emergence, p shown for the parameter ranges of θ and F_h in the corresponding red and blue boxes from figure (d). (f) The average number of times of 8×10^6 simulation runs during which a resistant strain emerges (black) or goes extinct (grey) during periods of low (β_l) or high (β_h) transmission for $p = 10^{-6}$. (g) A resistant strain was never observed to establish during periods of low transmission (β_l) for $p = 10^{-6}$.

Indeed, when $p = 10^{-6}$, in those cases when the resistant strain becomes established, its initial time of emergence frequently occurs at around the time when 60% of the population is vaccinated (Fig. 2.3). Therefore, we then tested the influence of a single intervention triggering at a single extraordinary period of low transmission centred around 60% of vaccinated individuals in the population (Fig. 2.3). We varied the duration of this intervention, T , ranging from 1 week to 120 days and considered three rates of transmission, $\beta_l = 0.055$ ($R_0 = 0.77$), 0.03 ($R_0 = 0.42$) and 0.01 ($R_0 = 0.14$). Both parameters decrease the probability of establishment of the resistant strain with the length of the intervention having a relatively stronger effect (Fig. 2.3).

2.3 Discussion

Our model suggests three specific risk factors that favour the emergence and establishment of a vaccine-resistant strain that are intuitively obvious: high probability of initial emergence of the resistant strain, high number of infected individuals [BGAC04] and low rate of vaccination [WMC17]. By contrast, a counterintuitive result of our analysis is that the highest risk of resistant strain establishment occurs when a large fraction of the population has already been vaccinated but the transmission is not controlled. Similar conclusions have been reached in a SIR model of the ongoing pandemic [RWK21] and a model of pathogen escape from host

immunity [C⁺18]. Furthermore, empirical data consistent with this result has been reported for influenza [Bon08]. Indeed, it seems likely that when a large fraction of the population is vaccinated, especially the high-risk fraction of the population (aged individuals and those with specific underlying conditions) policy makers and individuals will be driven to return to pre-pandemic guidelines [ERT20] and behaviours conducive to a high rate of virus transmission [F⁺20, D⁺20a]. However, the establishment of a resistant strain at that time may lead to serial rounds of resistant strain evolution with vaccine development playing catch up in the evolutionary arms race against novel strains.

Prior to discussion of the implications of our model we reflect on several properties of the assumptions and implementation of our model. In classical SIR-like deterministic models even a single individual infected with a vaccine resistant strain with reproduction number $R_t > 1$ will lead to automatic establishment of the strain in the population. In an analytical solution, a SIR-like model, even for $R_t < 1$ for the vaccine resistant strain, the number of infected individuals will tend to 0 but only as time tends to infinity. In actual populations, a single individual infected with a vaccine resistant strain still has a non-negligible chance not to infect anyone causing the variant to go extinct due to random stochastic forces [Hal27]. Therefore, the implementation of stochastic dynamics [AMB21, Kha20] of the vaccine resistant strain at low frequency in our model, considers the impact of random drift on its dynamics, which lies at the heart of extinction of rare strains.

We considered the dynamics of a single vaccine resistant strain, however, there may be different mutations that can lead to vaccine resistance. The emergence of different genotypes causing the same phenotype is analogous to a distinction in population genetics between alleles identical by state and by descent [Gil04a]. In our model, the treatment of independent emergence of different mutations as a single entity does not influence the dynamics under the following two assumptions. First, that different mutations lead to exactly the same phenotype, which is vaccine resistance, and, second, that there is no recombination. However, the reported dynamics may be quantitatively different if either of the two assumptions do not hold.

We have not explored the parameter ranges of β_h , β_l , the high and low rates of transmission, respectively, and F_l the threshold between low and high rate of transmission. We selected the β_h and β_l , to represent the known transmission values at the start of the pandemic [C⁺20b, HK20, ATS20]. However, evolving strains are reported to have a higher rate of transmission [D⁺21] leading to higher β_h and, possibly, β_l values than we used. An increase in the rate of transmission is not expected to qualitatively influence the reported dynamics, but would shift the probability density of establishment of a resistant strain (Fig. 2.3). Indeed, the peak probability at 60% vaccinated individuals roughly corresponds to the point at which for the given β_h , R_t , the average number of transmissions for one infected individual, becomes less than 1. Because the reproduction number for the vaccine resistant strain, R_t , is equal to $(S + V)\beta/N(\gamma + \delta)$, the perk of the risk of establishment of the vaccine resistant strain would increase proportional to an increase of β . An increase in either F_h or F_l would lead to more individuals becoming infected and a proportionally higher rate of emergence of the vaccine related strain, but would not change the qualitative behaviour of the model. Furthermore, an increase of F_l would lead to reversion to a high transmission rate with higher number of infected individuals, leading to shorter periods of low transmission and decreased probability of extinction of the vaccine resistant strains.

The results of our model provide several qualitative implications for the strategy forward in the months of vaccination. In our model, the probability of emergence of a resistant strain in one individual per day was in the range of 10^{-5} to 10^{-8} for a population of 10^7 individuals.

For the entire human population of roughly 10^{10} that probability would be 10^{-8} to 10^{-11} , which does not seem improbably large. As of February 2021, about 10^9 individuals have been infected by SARS-CoV-2 [DDG20] with an average 14 days of sickness per individual [L⁺20], so $> 10^{10}$ number of total days of infected individuals. Furthermore, highly mutated strains may emerge as a result of long shedding in immunocompromised individuals, a rare but realistic scenario [A⁺20c, C⁺20a, BSD⁺21, RKMK24]. Taken together, the emergence of a partially or fully vaccine-resistant strain and its eventual establishment appears inevitable. However, as vaccination needs to be ahead of the spread of such strains in similar ways to influenza [G⁺04], it is necessary to reduce the probability of establishment by a targeted effort to reduce the virus transmission rate towards the end of the vaccination period before the current vaccines become ineffective. Conversely, lack of non-pharmaceutical interventions at that time can increase the probability of establishment of vaccine-resistant strains. For example, plans to vaccinate individuals with a high risk of a fatal disease outcome followed by a drive to reach herd immunity while in uncontrolled transmission among the rest of the population is likely to greatly increase the probability that a resistant strain is established, annulling the initial vaccination effort. Another potential risk factor may be the reversion of vaccinated individuals to pre-pandemic behaviours that can drive the initial spread of the resistant strain.

One simple specific recommendation is to keep transmission low even when a large fraction of the population has been vaccinated by implementing acute non-pharmaceutical interventions (i.e. strict adherence to social distancing) for a reasonable period of time, to allow emergent lineages of resistant strains to go extinct through stochastic genetic drift. The implementation of non-pharmaceutical measures at a time of high vaccination can also help reduce infectivity when the efficacy of vaccines is not perfect [M⁺20a]. Additional factors that may make these measures even more effective are: (1) increased and widespread testing, (2) rigorous contact tracing, (3) high rate of viral sequencing of positive cases [Bon08, PR18] and (4) travel restrictions. Finally, while our model formally considers only one homogenous population, our data also suggest that delays in vaccination in some countries relative to others will make the global emergence of a vaccine-resistant strain more likely. Without global coordination, vaccine resistant strains may be eliminated in some populations but could persist in others. Thus, a truly global vaccination effort may be necessary to reduce the chances of a global spread of a resistant strain.

2.4 Methods

Our extension of the SIR Model features a modular number distinct states. Susceptible, S , and recovered, R , individuals are vaccinated over time to become vaccinated, V , or recovered vaccinated, R^V . Susceptible individuals can become infected with the wildtype, I_0 , or a mutant strain I_j with $j \neq 0$. After a while any infected individual recovers or dies, D . Finally, we assume that the recovered population retains natural immunity towards both strains, but becomes susceptible again with some small rate, μ . In our model, immunity against the wildtype strain gained through vaccination is not lost during the entire model period of 3 years. The total number of individuals, N , in the population remains constant at 10,000,000.

$$S + I_{wt} + I_r + I_r^V + R + R^V + D + V = N \quad (2.2)$$

In the limit of large population sizes, the full dynamics without mutations can be described by the following set of differential equations.

$$\dot{S} = \mu R - \theta(t)S - \beta(t)(I_{wt} + I_r + I_r^V)S \quad (2.3)$$

$$\dot{I}_{wt} = -(\gamma + \delta)I_{wt} + \beta(t)SI_{wt} \quad (2.4)$$

$$\dot{I}_r = -(\gamma + \delta)I_r + \beta(t)S(I_r + I_r^V) \quad (2.5)$$

$$\dot{I}_r^V = -(\gamma + \delta)I_r^V + \beta(t)V(I_r + I_r^V) \quad (2.6)$$

$$\dot{R} = -\mu R - \theta(t)R + \gamma(I_{wt} + I_r) \quad (2.7)$$

$$\dot{R}^V = -\mu R^V + \theta(t)R + \gamma I_r^V \quad (2.8)$$

$$\dot{D} = \delta(I_{wt} + I_r + I_r^V) \quad (2.9)$$

$$\dot{V} = \mu R^V + \theta(t)S - \beta(t)I_r V \quad (2.10)$$

Or $\dot{X} = \hat{P}X$, with $X = (S, I_{wt}, I_r, I_r^V, R, R^V, D, V)^T$

$$\hat{P} = \begin{pmatrix} -\theta(t) & -\beta(t)S & -\beta(t)S & -\beta(t)S & \mu & 0 & 0 & 0 \\ \beta(t)I_{wt} & -(\gamma + \delta) & 0 & 0 & 0 & 0 & 0 & 0 \\ \beta(t)I_r & 0 & -(\gamma + \delta) & \beta(t)S & 0 & 0 & 0 & 0 \\ 0 & 0 & \beta(t)V & -(\gamma + \delta) & 0 & 0 & 0 & \beta(t)I_r^V \\ 0 & \gamma & \gamma & 0 & -(\mu + \theta(t)) & 0 & 0 & 0 \\ 0 & 0 & 0 & \gamma & \theta(t) & -\mu & 0 & 0 \\ 0 & \delta & \delta & \delta & 0 & 0 & 0 & 0 \\ \theta(t) & 0 & -\beta(t)V & 0 & 0 & \mu & 0 & 0 \end{pmatrix} \quad (2.11)$$

The dynamics are influenced by the following constant parameters: the recovery rate, γ , the death rate, δ , and the rate at which natural immunity is lost, μ . Additionally we introduce a time dependent transmission rate $\beta(t)$ and a function $\theta(t)$, which controls the speed of vaccination.

Time Dependent Transmission Rate $\beta(t)$ switches between high and low transmission rates [C⁺20b]. The model begins with a period of a high rate of transmission, β_h/N . A low transmission rate, β_l/N , is initiated when the fraction of individuals infected with any strain, $I = (I_{wt} + I_r + I_r^V)$, reaches the value of F . Transition from a period of low to high transmission occurs when $I = F_l = 1000$.

Vaccination Vaccination is modelled as a mostly linear function with saturation. h denotes the number of individuals in the population that are never vaccinated. A maximum of $N - h$ individuals can be vaccinated at the end of the vaccination program. The constant k controls the saturation of the vaccination speed once the number of susceptible individuals is significantly depleted. The state dependent vaccination speed $\theta(t)$ is given as:

$$\theta(t) = \left(1 - \frac{h}{S(t) + R(t) + I_{wt}(t) + I_r(t)}\right) \cdot \frac{\theta_0}{S(t) + R(t) + k}, \quad (2.12)$$

where θ_0 can take different values and h and k are chosen to be small (see Extended Data Table 2.1).

Integration Method The deterministic differential equations, Eq. 2.11, were numerically solved using an Euler Forward Integration Scheme, with time step Δt , measured in days.

Resistant Strain Each day and for every I_{wt} individual, there is a small probability p , that a vaccine-resistant strain emerges in that individual. Then this individual switches from state I_{wt} to state I_r . Conversely, any I_r individual can revert back to the wildtype strain population I_{wt} with the same probability p . Each time step the number of mutations is drawn from a Poisson distribution with mean $\Delta t p I_{wt}$, for mutations to the resistant strain, or $\Delta t p I_r$, for mutations to the wildtype strain.

Stochastic and Deterministic Regimes The population dynamics of a rare variant is an inherently stochastic process [CC10, Hal27]. We can formally treat the spread of a disease in our model as a stochastic birth-death process. In the following we illustrate this with the number of wildtype infections I_{wt} as an example. In each infinitesimally small time step dt , there is a probability $\beta S I_{wt} dt$, that the wildtype population I_{wt} grows by 1, $I_{wt} \rightarrow I_{wt} + 1$, while the susceptible population is decreased, $S \rightarrow S - 1$. Similarly, with probability $(\gamma + \delta) I_{wt} dt$, I_{wt} is reduced by 1, $I_{wt} \rightarrow I_{wt} - 1$, while the number of recovered or dead grows by 1. We carefully model small populations, $I_{wt} < N^*$, using a stochastic Tau-Leaping Algorithm [CGP05]. We choose a fixed time step size τ , that is equal to the time step of the Euler Integrator Δt . For very small I_{wt} Tau Leaping Algorithm can produce negative population [CGP05]. This stems from the fact, that the number of events K that occur in time τ is drawn from a Poisson Distribution, that always assigns a non-zero probability for any $K > I_{wt}$. We reduce the chances of such a scenario, by solving the exact SSA Gillespie Algorithm for very small I_{wt} that is below a critical size N_c .

For large I_{wt} , larger than some N^* , this stochastic process can be approximated with the limiting differential 2.4 and an Euler Integration Scheme. Once $I_{wt} \geq N^*$, we consider that the resistant strain of the virus is established in the population and we continue modelling it using the deterministic equation Eq. 2.11].

Parallel Evaluation Of Deterministic And Stochastic Variables In our model we can then evaluate deterministic and the stochastic dynamics in parallel. While small populations of infected individuals are treated as stochastic, other variables, such as the number of susceptible individuals, S , are evaluated within the deterministic regime. While the infection numbers of the wildtype or the emergent strain are in the stochastic regime ($< N^*$), the corresponding terms that contain the wildtype infections I_{wt} or the emergent infections I_r & I_r^V are removed from the deterministic rate equations. When I_{wt} or $I_r + I_r^V$ grows above the threshold value N^* , the corresponding population of infected individuals is treated as deterministic.

Sources of Errors Finally, we discuss some sources of errors in our simulation: (1) Depending on the time step Δt the Euler Integration Scheme is not exact. In most of our simulations, we choose a time step of one day, $\Delta t = 1d$. (2) Using the deterministic rate equations for the infection numbers in 2.11 is an approximation to the exact stochastic dynamics given by a birth-death model. The quality of this approximation is given by the threshold value N^* . In our model we use a threshold of $N^* = 1000$. (3) When I_{wt} or $I_r + I_r^V$ trespasses the threshold N^* from above, the populations of infected individuals changes from being treated as a real number (the mean field average) to being treated as a natural number. We truncate the mean field average with a floor function and treat the remainder as part of the recovered population, $I_{wt} \rightarrow \lfloor I_{wt} \rfloor$ & $R \rightarrow R + I_{wt} - \lfloor I_{wt} \rfloor$. (4) The Tau Leaping algorithm is an approximation to the exact SSA Gillespie Algorithm, that allows faster evaluation with a constant step size τ . Increasing the threshold value N_c increases the accuracy of the model (Supplementary Figure 2.4). (5) As discussed above, it can rarely happen that a population of infected individuals

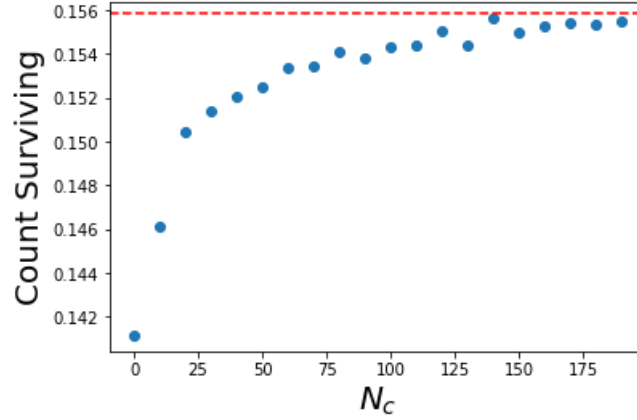


Figure 2.4: Shown is the fraction of surviving strains after $T = 200$ days in 10^7 runs, first initialized with $I_{wt} = 200$ infected individuals. The red dashed line shows the expected fraction of surviving strains, as computed with 2.13. The stochastic algorithm becomes exact, if no Tau Leaping is employed and instead the whole simulation is evaluated using the Gillespie SSA scheme.

drops below 0 in one leap. If this happens we disregard the draw and redraw from the same Poisson distribution. (6) Finally, while the time step of the deterministic model Δt and τ are chosen to be equal, for population sizes above N_c , the SSA algorithm acts on exponentially distributed waiting times τ_{SSA} between reactions [Gil01]. This introduces errors, if $\tau_{SSA} \gg \Delta t$, because the interaction rates may change, while we wait for the next reaction to occur.

In order to determine a range of acceptable values of N_c , we ran our simulation for a period of $T = 200$ days, initially loading the system with $I_{wt} = 200$ wildtype carriers. For multiple values of N_c and no mutations, we compared the results of disease survival with the analytical solution derived for the birth death process,

$$P(I_{wt}(T) = 0) = \left(\frac{(\delta + \gamma)e^{(\delta + \gamma - \beta)T} - \gamma - \delta}{(\delta + \gamma)e^{(\delta + \gamma - \beta)T} - \beta} \right)^{I_{wt}(0)}. \quad (2.13)$$

We pick $N_c = 100$ days, which gives us a reasonably small error.

Assumptions and Choice of Parameters The model is run for a total time of three years, with vaccination starting one year into the model. We assume that the wildtype and emergent strains have the same infectivity (β is the same for both strains). We assume that infection by any one strain provides immunity to both, reflecting that many vaccines carry only the Spike protein of the SARS-CoV-2 virus and it may be easier to escape immunity provided by the vaccine than the immunity provided by infection. We also assume that the immune response provided by the vaccine is more permanent and that immunity provided by infection, μ , decays in 0.5 years [K⁺18] after recovery. Both of these assumptions influence the model when the number of infected individuals becomes large, which is unlikely for realistic average rates of transmission across the simulated time.

We assume that susceptible and recovered individuals have an equal chance to be vaccinated, θ_0 . We also assume that the infection-recovery rate, γ , and infection-fatality rate, δ are the same for the wildtype and mutated strains.

2. TRANSMISSION AND VACCINATION IMPACT THE FATE OF VACCINE-RESISTANT STRAINS

We regulate the rates of transmission exogenously in the model, with the rate of transmission (β) switching between a high rate, β_h and a low rate, β_l , when the total number of individuals infected with either strain reaches a threshold parameter. This threshold parameter, F , simultaneously reflects the impact of all non-pharmaceutical interventions and behavioural changes of individuals. We operate under the assumption that vaccine efficacy not only impacts disease manifestation but also blocks transmission at the same rate, which is a reasonable assumption based on previous vaccine performance but has not yet been demonstrated.

In Extended Data Table 2.1 we present the choice of parameters for the model, the ones that were constant and those that were varied, including their boundaries.

Parameter	Value	Comments
Fixed parameters		
Population size, N	10,000,000 individuals	
Cut-off for stochastic mode, N^*	1000	
Cut-off for Gillespie algorithm, N_c	100	
Recovery rate, γ	$0.99 \cdot 1/14$	Average disease duration 14 days
Death rate, δ	$0.01 \cdot 1/14$	Infection-fatality rate is 1%
Loss of immunity rate, μ	$1/180$	On average in 180 days
Share of non-vaccinated, h	0.01	1% of population
Saturation parameter, k	0.01	
Transmission rate, $\beta = \{\beta_l, \beta_h\}$	$\{0.055, 0.18\}$	$R_0 = \{0.77, 2.52\}$
Bound for initiation of high transmission	1,000 individuals	
Varying parameters		
Bound for initiation of low transmission, F	2,000 to 20,000 individuals	
Probability of emergence of resistant strain, p	1e-8 to 1e-5	Daily for every infected
Vaccination speed, θ_0	0.001 to 0.015	0.1% to 1.5% daily

Table 2.1: Model Parameters

Complex Vaccination Strategies & the Emergence of Vaccine Resistance

Published Work This chapter appeared in full as Simon A Rella, Yuliya A Kulikova, Aygul R Minnegalieva, and Fyodor A Kondrashov. Complex vaccination strategies prevent the emergence of vaccine resistance. *Evolution*, 78(10):1722–1738, 2024.

Author Contributions S.A.R. and F.A.K. formulated the question. S.A.R., F.A.K, Y.A.K. and A.R.M. formulated the model. S.A.R. programmed and ran the model. All authors jointly analyzed the results and wrote the manuscript.

Abstract Vaccination is the most effective tool to control infectious diseases. However, the evolution of vaccine resistance, exemplified by vaccine-resistance in SARS-CoV-2, remains a concern. Here, we model complex vaccination strategies against a pathogen with multiple epitopes - molecules targeted by the vaccine. We found that a vaccine targeting one epitope was ineffective in preventing vaccine escape. Vaccine resistance in highly infectious pathogens was prevented by the full-epitope vaccine, that is, one targeting all available epitopes, but only when the rate of pathogen evolution was low. Strikingly, a bet-hedging strategy of random administration of vaccines targeting different epitopes was the most effective in preventing vaccine resistance in pathogens with low rate of infection and high rate of evolution. Thus, complex vaccination strategies, when biologically feasible, may be preferable to the currently used single-vaccine approaches for long-term control of disease outbreaks, especially when applied to livestock with near 100% vaccination rates.

3.1 Introduction

The COVID-19 pandemic raised public awareness to the dangers and epidemiological characteristics of infectious disease. The obvious danger is that of a new disease emerging in the human population, livestock or crops impacting public health and the global food chain supply. Our experience with COVID-19 also demonstrated the danger of the emergent disease to evolve, changing our ability to contain the spread of the disease through an increase of infectivity [WLZ⁺22] and evolving vaccine resistance [WGM⁺22, GBLD⁺21]. An interesting aspect of SARS-CoV-2 is its propensity for rapid evolution driven by the host immune response in immunocompromised individuals over the course of very long infection periods

[CCR⁺20, KCD⁺21, SPS⁺22]. In such cases, the SARS-CoV-2 virus has the time to adapt to the relatively weak pressure of the compromised immune system with the resulting adapted variants posing a greater threat to the general population [LWC⁺22, ABB⁺23].

The ideal goal when dealing with an emergent disease is eradication, as has been achieved in some cases [OR18, The19, BA80, Nor10]. The second best option is to control the pathogen's spread and evolution that would allow it to avoid these control mechanisms. Vaccination of the entire population should reduce the rate of spread of the virus. However, the COVID-19 pandemic demonstrated that even the goal of containment may not be easily achieved [ZRZ⁺22]. SARS-CoV-2 rapidly evolved variants with a much higher infectivity [SKK⁺21] and showed a tendency to evade the vaccine-induced immune response [PSM⁺22, MBM⁺21]. Both of these factors may have been driven by evolution of SARS-CoV-2 in immunocompromised patients [SPS⁺22].

A potential solution for the prevention of pathogen evolution is a multi-epitope or mosaic vaccine [KR17, BTW⁺18, CM10, HSF⁺19, Suh97]. Such a vaccine causes the immune system to develop antibodies against different epitopes, which are molecular targets for the immune system, frequently a part of a protein displayed on the surface of the pathogen. Their application to SARS-CoV-2 by selection of several epitopes in the Spike protein has been considered [KNB⁺20, ZHL⁺22] but not implemented. In theory, a vaccine that targets several epitopes at once reduces the probability of evolution of vaccine resistance [KR17] thereby allowing us to achieve the second best outcome — long-term control of the pathogen spread in the population. Furthermore, a more complex, mosaic strategy was proposed by McLeod et al. [MWM21], whereby a combination of different vaccines targeting a different set of epitopes can be used in the population reducing the rate of spread of vaccine resistance.

Barring issues with the host immune response, the hypothesis that a vaccine that simultaneously targets several epitopes is better than a vaccine that targets just one seems logical. Specifically, when a multi-epitope vaccine is used, more mutations have to occur in the pathogen to evolve vaccine escape, reducing the probability of such evolution [REX13, MWM21]. On the other hand, simultaneous introduction of a vaccine against all epitopes may have a potential weakness. In the arms race between the pathogen and the immune system, massive vaccination of all individuals with a multi-epitope vaccine may be equivalent to showing all of the cards to the pathogen allowing it to simultaneously adapt to all of the epitopes presented in the vaccine and winning in the evolutionary arms race. This process may be particularly relevant if the evolution of the pathogen is enhanced by selection due to the immune system in the same way as SARS-CoV-2 evolution is enhanced in immunocompromised individuals [GD08, SPS⁺22, LWC⁺22, ABB⁺23].

Here, we consider a hypothetical model of pathogen evolution in a vaccinated population. We study the conditions under which the pathogen does not evolve vaccine resistance, in other words when the pathogen remains controlled or, in rare cases, is eradicated from the population. We specifically focus on the efficacy of complex vaccination strategies with a combination of different vaccines targeting different epitopes to control the spread of the pathogen and to reduce the probability of evolving vaccine resistance.

3.2 Methods

In our model, a schematic of which is shown in Fig. 3.1, we have an infinite population size with different states assigned to individuals. Given a current snapshot of the population, the

state of an individual can be unvaccinated (S), unvaccinated and infected by a pathogen φ (SP_φ) and vaccinated by a specific vaccine of type σ (V_σ). Further, some fraction of the population, hereby referred to as immunocompromised (I), experiences prolonged disease when infected. The immunocompromised can be vaccinated (IV_σ), vaccinated and infected ($IV_\sigma P_\varphi$) or unvaccinated and infected (IP_φ). In the immunocompromised vaccinated and infected individuals ($IV_\sigma P_\varphi$) the immune system can select for mutations ($\varphi \rightarrow \varphi'$) in epitopes that have been displayed by the vaccine σ to the immune system [ABB⁺23]. These mutations render the vaccine ineffective against the respective epitopes. Such selection is hindered in healthy individuals as vaccines effectively suppress vaccinated and infected states ($V_\sigma P_\varphi$). We later consider a variant of the model, which relaxes this assumption: Immunocompromised individuals are excluded but regular vaccinated individuals experience breakthrough infections and in these individuals the pathogen can evolve. In that case, the individual states are, as before, (S), (SP_φ), (V_σ) and with an additional state of vaccinated and infected ($V_\sigma P_\varphi$). It is possible to extrapolate from one model to the other by tuning the parameters ρ (fraction of breakthrough infections) and τ (length of disease), discussed below.

For simplicity, we base our model on an epidemiological SIS-model with vaccination [KR08] without a compartment of recovered individuals, therefore only vaccines provide immune protection against the pathogen. We review this assumption in the discussion.

Using equations derived from stochastic epidemiological models we consider the probability of fixation at a given point in time of such vaccine-resistant variants in an immunocompromised individual and the probability of establishment of these variants in the entire population. Note, that we are not solving for the expected fraction of infected and susceptible, as is usually done with SIS-models, but rather we compute the probability of emergence and establishment of resistant variants for given model variables. We use the term establishment when a pathogen spreads to many individuals in the population and we talk about fixation of a variant within a host.

Epitopes We consider the epidemiological and evolutionary dynamics of a pathogen with n epitopes, each epitope we denote as e_k , where $k = \{1, \dots, n\}$. We introduce vaccines that induce antibodies against these epitopes and we distinguish several different types of vaccines. A single-epitope vaccine creates antibodies against one epitope while multi-epitope vaccines create antibodies against several epitopes. A multi-epitope vaccine does not necessarily induce antibodies against all possible epitopes: this is achieved by a full-epitope vaccine, which is a special case of the multi-epitope vaccine. We denote the number of epitopes per multi-epitope vaccine as its valence, m , where m is a discrete number ranging between $\{1, \dots, n\}$. The binomial coefficient $C(n, m)$ gives the number of possible types of multi-epitope vaccines with valence m . The broadest immune response will be induced by the full-epitope vaccine with valence $m = n$.

Pathogen Variants In the model there are several pathogen variants, φ , where each variant carries a unique set of epitope states. We denote a variant as the set of its epitopes in the wildtype state. The initial condition starts with the wild type pathogen that carries all unmutated epitopes $\varphi = \{e_1, \dots, e_n\}$, $|\varphi| = n$, where all e_k can be targeted by existing vaccines. New variants can emerge that carry mutations in some epitopes, rendering them undetectable by the memory immune response induced by the corresponding vaccines. The number of epitopes that acquire resistant mutations is denoted as i . A super resistant pathogen acquires mutations in all n epitopes (Fig. 3.2), such that $\varphi = \emptyset$ and $|\varphi| = 0$. There exist different pathogens with i mutated epitopes, $C(n, i)$.

Immunity Ideal vaccines prevent all vaccinated individuals from becoming infected [GMNR01].

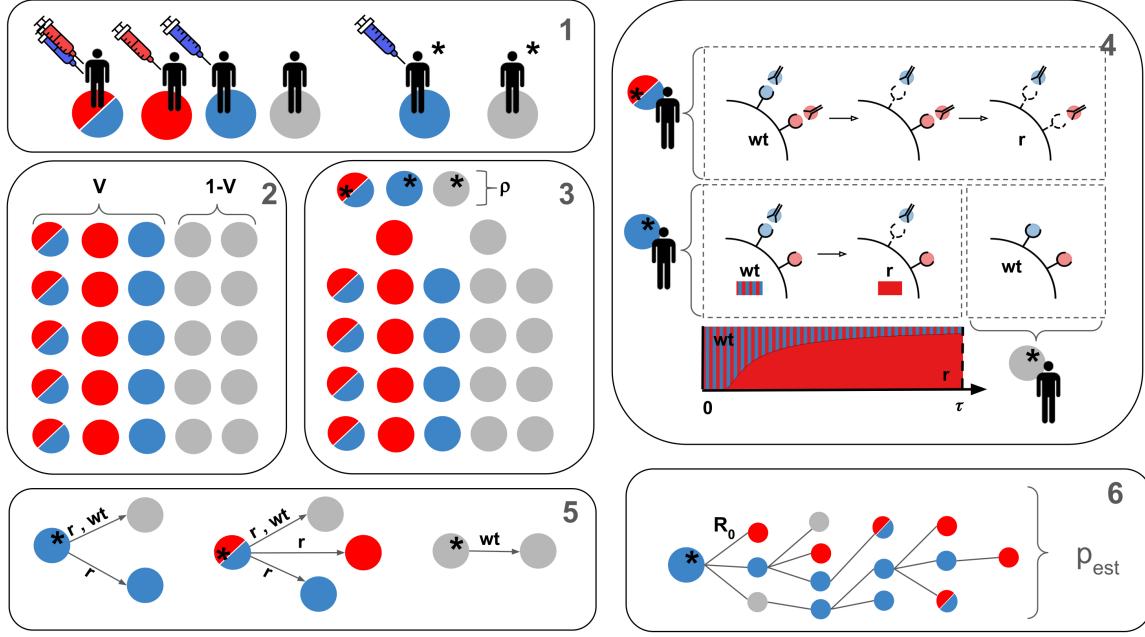


Figure 3.1: Overview of the model for the case of two epitopes in vaccine design. (1) The model features several states of an individual: healthy or vaccinated. Individuals can be vaccinated with a 2-epitope vaccine or one of two single-epitope vaccines (red and blue). Unvaccinated individuals are shown in gray. Further, each individual can be in a diseased state (marked by an asterisk, *). (2) All N individuals are vaccinated in accordance with the chosen vaccination strategy. (here $N = 25$ for visualization purposes, the actual model operates in the limit of large N). (3) Immune-driven selection in infected individuals may lead to the emergence of vaccine-resistant strains with some small probability. (4) Selection pressure drives the wild type population of pathogens (wt) to evolve resistance towards administered vaccines (r). For instance, an individual that received the blue vaccine, will only develop an immune response towards the blue epitope. Over time the pathogen population in the host will become a vaccine-resistant population from a wild type population. We assume that vaccine resistance does not emerge in unvaccinated individuals. (5) Once evolved, mutants will spread within the population of hosts. However, due to different host immune types, some variants have reduced transmissibility. (6) Using a branching process we evaluate the probability that the vaccine resistant strain infects a certain fraction of the population, not vanishing in the large N limit, which we identify as the probability of establishment.

However, no vaccine is perfect, and a small fraction of the vaccinated population (ρ) can get infected and infect others. Further, as we will detail below, these individuals can still produce some immune response which can drive the selection of the pathogen inside the body. The vaccinated infected individuals experience the disease for a time period τ . Under model conditions with immunocompromised individuals, ρ is small and τ is large. When we model breakout infections with imperfect vaccines, ρ is larger and τ is small. The immunocompromised model and the model with imperfect vaccines are similar. In the former few individuals have a high probability of causing the evolution of a vaccine-resistant strain while in the latter, many individuals have a small probability of doing the same.

In our model an individual that received a vaccine containing the antigen e_k will be immune against every pathogen that carries the wild-type version of the epitope e_k . The immune state

induced by a vaccine can be represented as the set of epitopes, against which immunity is generated $\sigma = \{\dots, e_k, \dots\}$. In the baseline model described here, the size of σ , $|\sigma|$, equals the vaccine valence m .

Our model relies on the simplifying assumption that each recognized epitope contributes equally to the overall rate of pathogen clearance. Vaccines with more epitopes do not clear a recognized pathogen faster than a single-epitope vaccine, because both are assumed to induce the same total number of antibodies, but the multi-epitope vaccine induces multiple available antibody types. This symmetry assumption reduces the relevant model parameters from the full sets σ and φ to their integer set sizes, m and i .

Immunodynamics The immune system will react differently towards specific pathogen variants. Drawing inspiration from basal formulations of immunodynamics [NM00] we denote the growth rates of the infected cell population x_φ with pathogen φ by the differential equation

$$\frac{dx_\varphi}{dt} = rx_\varphi - (a + z_\sigma(x_\varphi))x_\varphi \quad (3.1)$$

where $z_\sigma(x_\varphi)$ corresponds to the strength of the immune response towards pathogen x_φ if the immune type of the patient is σ , a is the baseline decay rate and r the growth rate of infected cells. If $r/(a + z_\sigma(x_\varphi)) < 1$, the population of infected cells cannot grow, ultimately leading to the clearance of the pathogen. While the pathogen population might initially grow, a healthy vaccinated individual will soon activate its immune response $z_\sigma(x_\varphi)$, such that $r/(a + z_\sigma(x_\varphi)) < 1$. Individuals with prolonged diseased state will have $r/a > r/(a + z_\sigma(x_\varphi)) > 1$ for a longer period of time and the pathogen can not be cleared from the patient's system right away. This provides the pathogen with an environment that can select for resistance mutations. Note, that the absence of an immune response $z_\sigma(x_\varphi) = 0$ as well as a strong immune system with $r < (a + z_\sigma(x_\varphi))$ will not provide a strong selective environment [KR17, BMSN97].

Evolution In each infected and vaccinated individual ($IV_\sigma P_\varphi$) we model evolution as a Bernoulli process, later referred to as the Bernoulli Model. Mutation, selection and fixation occur at once and immediately change the state of the whole pathogen population in a patient from one variant to another. A pathogen population with i mutations, in a host vaccinated against m epitopes, becomes a pathogen population with $i + 1$ mutations with a probability p per day of infection. Always starting with the wildtype, the probability to find a pathogen with i mutated epitopes by time t , given a vaccine of valence m was administered, can be approximated with a Poisson distribution,

$$p_m(i) = \frac{(pt)^i e^{-pt}}{i!}, \quad \text{for } i < m. \quad (3.2)$$

If we ignore any adaptive immune response that goes beyond the epitopes that were targeted by the vaccine, we further set

$$p_m(i) = \sum_{j=m}^{\infty} \frac{(pt)^j e^{-pt}}{j!}, \quad \text{for } i = m, \quad (3.3)$$

$$p_m(i) = 0, \quad \text{for } i > m. \quad (3.4)$$

Note that this simple treatment of the evolutionary process ignores any differential fitness effects between variants. Further, it requires that evolutionary trajectories always start with the

wildtype. Employing more complicated evolutionary algorithms, such as the infinite population model of population genetics or the Wright-Fisher model with mutation and selection [HC06] in an immunodynamical setting, generates similar qualitative results (see supplementary sections 1-2 in [RKMK24]). More generally speaking, the presented algorithm, as well as its more complicated variants, do not exclude the possibility that pathogen diversity has existed even prior to vaccination, which would correspond to a different initial condition.

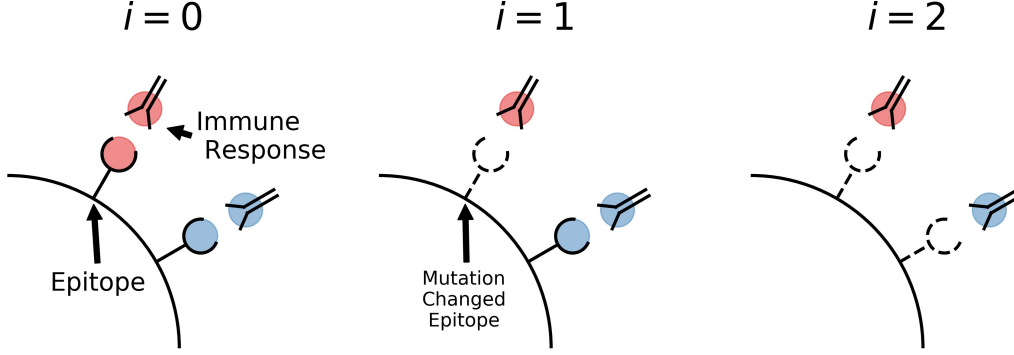


Figure 3.2: An example of evolution of vaccine resistance in a pathogen with two epitopes. Starting with a pathogen displaying two epitopes ($n = 2$, coloured spheres), two of which are recognized by antibodies induced by a vaccine with a valence of 2 ($m = 2$, corresponding coloured antibodies). Mutations may render the epitopes unrecognizable by the vaccine-induced antibodies. If both of the epitopes acquire such mutations (farthest right drawing) the pathogen is no longer impaired by the vaccine.

Vaccination Strategies When multiple epitopes can be targeted, vaccines with different valences (number of targeted epitopes) can be created (Fig. 3.3). If a combination of different vaccines can be used, four different vaccination strategies become possible: (1) a full-epitope vaccination strategy, in which a fraction V_n of the population receives a vaccine of valence $m = n$, (2) a single-epitope vaccination strategy, in which a fraction V_1 of the total population receives one of n different vaccines with valence $m = 1$ at random (thus, a fraction V_1/n is vaccinated with one of the n vaccines), (3) an m -epitope vaccination scheme, in which each vaccinated individual in a fraction V_m of the total population receives one of $C(n, m)$ possible combinations of epitopes at random, and (4) a mixture of the strategies 1-3, constrained by the normalization condition

$$S + \sum_{m=1}^n V_m = 1 \quad . \quad (3.5)$$

Under all strategies, a fraction S of the population remains unvaccinated, which ultimately is an important parameter for the observed dynamics in our model.

Transmission We assume a finite number of infected individuals N , out of which a fraction ρ , are infectious. Given one of $N\rho$ individuals, in which the pathogen can evolve if vaccinated, ($IV_\sigma P_\varphi$) or ($V_\sigma P_\varphi$), the probability $p_{trans}(i)$ that it will transmit a pathogen variant with i mutated epitopes is

$$p_{trans}(i, \tau) = \sum_{m=1}^n V_m \int_0^\tau \frac{p_m(i)}{\tau} dt \quad (3.6)$$

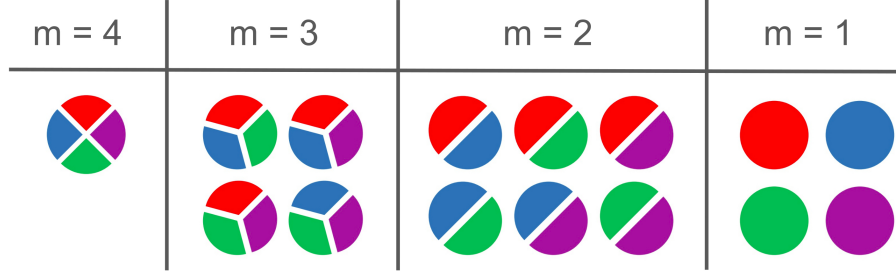


Figure 3.3: Available multi-epitope vaccines for a vaccination procedure targeting n different epitopes. Different possible vaccines for the number of epitopes $n = 4$. Full-epitope strategy (left panel) with $m = 4$, where the population is inoculated with the n -epitope vaccine. A single-epitope strategy (right panel) with $m = 1$, where the population is inoculated with 4 different vaccines, each of them against a single epitope. The intermediate, m -epitope strategy (two middle panels) with inoculation by multiple vaccines, against 2 or 3 different epitopes. A mixed strategy can have vaccines containing a combination of all 4 different vaccine types (n -epitope, m -epitope and single-epitope).

where p_m is a Poisson distribution as defined before and V_m is the fraction of the population vaccinated with any m -epitope vaccine. The integral corresponds to the probability of a variant i being transmitted at any time t from the beginning of the disease to its end at $t = \tau$.

Establishment within the Population To understand whether a pathogen variant is fixed in the population, we first derive the basic reproductive number for a pathogen with i mutations, R_i ,

$$R_i = R_0 \left(1 - (1 - \rho) \left(\sum_{m=1}^i V_m \left(1 - \frac{C(i, m)}{C(n, m)} \right) + \sum_{m=i+1}^n V_m \right) \right). \quad (3.7)$$

R_0 denotes the basic reproductive number, that is, the expected number of secondary infections induced by the wildtype pathogen in an unvaccinated population, and ρ is the fraction of the vaccinated population that is infectious, thereby contributing to population-level transmissibility. Finally, establishment in a population with a large population size and random interactions can be approximated by a branching process as $1 - 1/R_i$, where $R_i \geq 1$ [PW08], [LHN05]).

We now derive Eq. 3.7. Let us begin with the case of a strategy employing equal proportions of m -epitope vaccines (the total number of epitopes is n) with perfect efficacy ($\rho = 0$). A fraction V_m of the population is vaccinated with one of $C(n, m)$ m -epitope vaccines, while the remaining fraction is unvaccinated, S .

$$S + V_m = 1 \quad , \quad V_k = 0 \quad \text{for} \quad k \neq m \quad . \quad (3.8)$$

We may also denote the vector $V = (S, 0, \dots, V_m, \dots, 0)^T$ as the strategy. Then, the basic reproductive rate of the wildtype is given by,

$$R_{wt}[V] = R_0(1 - V_m) = R_0 S \quad . \quad (3.9)$$

A mutant that is immune to the vaccine, will again have the reproductive success of a wildtype in an unvaccinated population, R_0 . A mutant with i mutations will be able to infect

$C(i, m)$ types of vaccinated individuals, that is a fraction $C(i, m)/C(n, m)$ of all $C(n, m)$ vaccines. Therefore

$$R_i[V] = R_0 \left(S + V_m \frac{C(i, m)}{C(n, m)} \right) . \quad (3.10)$$

In case of a general strategy we can write

$$R_i[V] = R_0 \left(S + \sum_{m=1}^i V_m \frac{C(i, m)}{C(n, m)} \right) = R_0 \left(1 - \sum_{m=1}^i V_m \left(1 - \frac{C(i, m)}{C(n, m)} \right) \right) - \sum_{m=i+1}^n V_m . \quad (3.11)$$

If vaccines are not perfect, a fraction ρ experiences wildtype transmission R_0 , while the remainder $(1 - \rho)$ has transmission Eq. 3.11, therefore

$$R_i[V] = R_0 \left(\rho + (1 - \rho) \left(S + \sum_{m=1}^i V_m \frac{C(i, m)}{C(n, m)} \right) \right) . \quad (3.12)$$

After simple reformulations we reach Eq.3.7.

Overall Probability of Pathogen Evolution A new variant can potentially emerge in one of the $N\rho$ individuals, be transmitted and ultimately be established in the population. The probability for the combined outcome is given by

$$p_{\text{est}}(i, \tau) = p_{\text{trans}}(i, \tau) \left(1 - \frac{1}{R_i} \right), \quad \text{with } R_i \geq 1. \quad (3.13)$$

We are in particular interested in the probability p_{est} that any variant will establish in the population, as this signifies failure of the vaccination campaign. As a result of ρN initial transmission events (the average number of infected and immunocompromised individuals), this probability of establishment is

$$p_{\text{est}} = 1 - \left(\prod_{i=1}^n (1 - p_{\text{est}}(i, \tau)) \right)^{\rho N} = 1 - \exp \left(\rho N \sum_{i=1}^n \log(1 - p_{\text{est}}(i, \tau)) \right), \quad (3.14)$$

which, among various life-history attributes of the pathogen, will also depend on the employed vaccination strategy. Note that the presented algorithm only considers establishment as resulting from two distinct and sequential processes: (1) Mutation and (2) Transmission. Any additional mutations, which might occur within the transmission chain, are neglected, as they are not expected to contribute heavily to the probability that any variant will establish.

Evaluation of the Epidemic Burden Establishing variants will trigger an epidemic and a considerable fraction of the population will subsequently become infected. Basic epidemiology can provide estimates for the speed, size and length of an epidemic as a function of the reproductive number of a mutant. In fact, pathogens which have a high probability of establishing in a population cause relatively more secondary infections and ultimately larger epidemics. In the classical SIR model, the size of the pandemic $s(i)$, which is the total fraction of the population which becomes infected by an establishing variant i , can be determined by the inverse relation,

$$1 - s(i) = \exp(-R_0 s(i)). \quad (3.15)$$

It can be shown that this measure for the fraction of infected is strongly correlated with the probability of establishment of variant i with transmissibility R_0 (See supplementary section 5 in [RKMK24]). In this manuscript we present the probability of establishment, both as a measure for the risk of pathogen evolution, as well as a measure for the epidemic burden. In the case of Fig. 4 and Fig. 9 we also explicitly present the expected size of the epidemic resulting from vaccine resistant variants (see Supp. Res. Sec. 4 [RKMK24]).

3.3 Results

General description of the model We considered a hypothetical pathogen with n epitopes, which are targets of vaccines. Thus, a vaccine may target any number of $m = 1, \dots, n$ epitopes, with the full-epitope vaccine causing an effective immune response against all n epitopes. The effect of the vaccine epitopes is not cumulative, in other words a single-epitope vaccine is just as effective as the full-epitope vaccine against a pathogen with all n epitopes. The pathogen evolves by accumulating mutations that change the epitope in a way that renders the antibody against this epitope produced by any vaccine ineffective. For example, an individual vaccinated by a single-epitope vaccine with epitope e_3 can still be infected by a pathogen in which the epitope e_3 has been mutated. In the extreme case, a pathogen variant in which all epitopes were mutated can infect an individual vaccinated by any vaccine, including the full-epitope vaccine. The population has a fraction of individuals (V) that are vaccinated by a random vaccine from the pool of vaccines that is being used. Initially we consider vaccines to be perfect so that all vaccinated non-immunocompromised individuals cannot be infected and transmit the pathogen. However, a small fraction of immunocompromised individuals (ρ) can get infected even though they are vaccinated. When that happens, the pathogen in the body of the vaccinated and infected individual is subjected to mutation and selection (Eqs. 3.2-3.4) driven by the vaccine-induced immune response. The new pathogen variant created by these intra-host processes has some probability to be transmitted to other susceptible individuals. The transmitted pathogen variant may become extinct or may spread in the population and eventually become fixed.

The ultimate goal of a successful vaccination campaign is to prevent the spread of disease and reduce the associated disease burden in the form of the number of infected, diseased or disabled individuals. In this paper, we are not concerned with the complex dynamics of the disease burden, rather we study the possibility that a complex vaccination strategy can prevent the spread of vaccine-resistant variants in the first place. Therefore, we use the probability of establishment of the vaccine-resistant variant as a proxy for the success of a specific vaccination strategy, based on the following logic. If the vaccination strategy fails to prevent a vaccine-resistant variant from spreading in the population then the subsequent disease burden that will be caused by the spread of this variant will be determined by SIR-like dynamics and are not of further concern to the question whether the said strategy is successful. Furthermore, we show that the probability of establishment of a vaccine-resistant strain is correlated with the disease burden (see Supp. Res. Sec. 5 [RKMK24]).

Model with 2 epitopes We first consider the case of a pathogen carrying only two epitopes ($n = 2$). The population may be vaccinated by the full-epitope vaccine ($m = 2$) or each individual may receive one of the two single-epitope vaccines ($m = 1$). The complete set of strategies is determined by the parameter α , such that some proportion, α , of the vaccinated individuals receive one of the single-epitope vaccines, while the rest of the vaccinated individuals, $1 - \alpha$, receive the full-epitope vaccine. When $\alpha = 0$ all administered vaccinations are full-

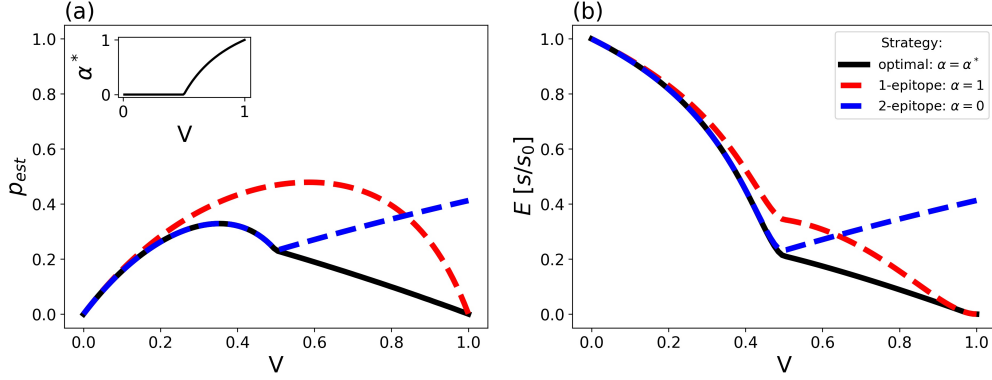


Figure 3.4: Optimal vaccination strategy to prevent evolution of vaccine resistance in the 2 epitope scenario. a) The probability of establishment of a vaccine resistant variant (p_{est}) as a function of V for three different vaccination strategies, full-epitope vaccination (blue line), single-epitope vaccination (red line) and for an optimal mixed vaccination strategy (black line). The black line of the inlay plot shows the optimal α^* , share of individuals vaccinated with a single-epitope vaccine, as a function of the fraction of the vaccinated population V . The full-epitope vaccination strategy matches the mixed optimal strategy below herd immunity ($V = 1 - 1/R_0 = 0.5$), after which the mixed strategy is best (black line). b) The expected epidemic burden, $E[s/s_0]$, which is the average fraction of affected individuals in the population compared to the unvaccinated scenario, as a function of V for the same three strategies as in (a). Other parameters for this figure were $R_0 = 2$, $\rho N = 10$, $\tau = 200$, $p = 10^{-3}$.

epitope, while when $\alpha = 1$ all administered vaccinations are single-epitope. It is straightforward to evaluate the optimal strategy, α , that minimizes the establishment of vaccine resistance for a given set of parameters using Eq. 3.14 (see the Model section). Similarly, we can compute the expected burden induced by evolving vaccine resistance for different values of α using the methods outlined in Supp. Res. Sec. 4 [RKMK24]. Fig. 4 shows the results of the optimal strategy as a function of the fraction of the vaccinated population, V . In a pure $m = 1$ strategy ($\alpha = 1$) the probability of establishment is largest at intermediate levels of vaccination (blue line in Fig. 3.4a). However, under this strategy the probability of establishment and the epidemic burden of a resistant variant is much lower for higher values of V when most individuals are vaccinated with one of the two single-epitope vaccines because the rate of transmission is lower ([RKDK21, CLN⁺18]). The full-epitope vaccine strategy ($\alpha = 0$) corresponds to the optimal strategy when vaccination rates are below the herd immunity threshold ($V_H = 1 - 1/R_0$). For vaccination rates higher than the herd immunity threshold the full-epitope vaccine strategy loses its efficacy and becomes less effective than a mixed strategy due to the small probability of establishment of super-resistant variants. Thus, for $V > V_H$, the optimal mixing strategy α^* outperforms the strategy relying on a single full-epitope vaccine (Fig. 3.4a,b).

The optimal strategy does not only depend on the vaccination rates, V , but is also strongly affected by the infectivity of the pathogen (R_0) and its persistence and evolution in the immunocompromised individual (τ , p , ρ). In Fig. 3.5 we show the probability of establishment of a vaccine resistant variant as a function of R_0 and the length of the disease duration τ of the immunocompromised individuals (we obtained similar results for p as we do for τ). For an optimal vaccine strategy (α^*), when R_0 and τ are small, vaccine resistance does not evolve. When both R_0 and τ are large, a pathogen is highly infectious and spends a long time in

an immunocompromised individual making the establishment of the vaccine-resistant variant inevitable. For the intermediate ranges of R_0 and τ , an optimal vaccination strategy mostly does not depend on τ (Fig. 3.5a). The pure full-epitope strategy is only optimal for high levels of R_0 . However, even the straightforward single-epitope strategy outperforms the full-epitope strategy when τ is large and R_0 is small. Across a large range of parameters a mixed optimal strategy outperforms both the straightforward single-epitope and the full-epitope strategies (Fig. 3.5b).

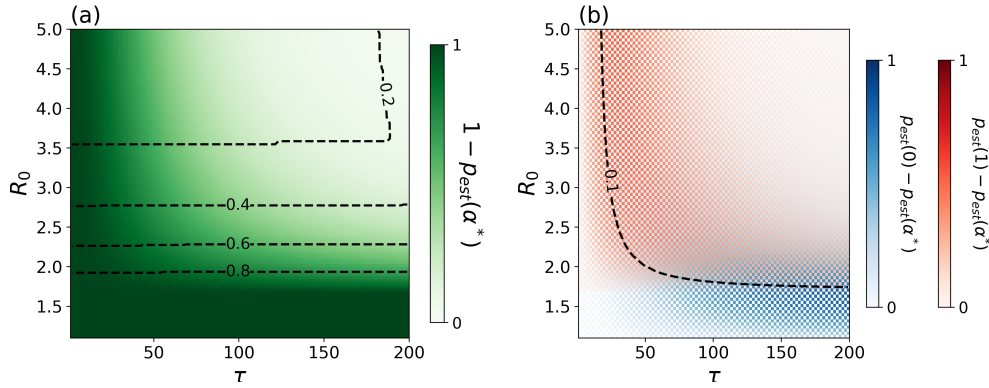


Figure 3.5: Optimal vaccination strategy as a function of pathogen infectivity and disease duration in an immunocompromised individual. Share of vaccinated individuals in population, $V = 0.8$, initial number of immunocompromised individuals, $N = 1$, τ – disease duration of immunocompromised individuals, R_0 – transmissibility of the pathogen. (a) The probability of preventing the establishment of a resistant variant (green) for optimal vaccination strategy (α^*) is shown. Dotted contour lines show the fraction of single-epitope vaccines (0.8, 0.6, 0.4 and 0.2) in the optimal solution α^* for different levels of R_0 and τ . (b) Difference in the probability of establishment of a vaccine resistant strain for the optimal strategy α^* and the single-epitope vaccination strategy (blue) and the full-epitope strategy (red). Parameter space above the contour line corresponds to the probability of establishment of a vaccine resistant variant above 10% for the optimal strategy α^* (see a).

Model with n epitopes The case with a large number of epitopes is more complex due to the combinatorially large number of mixed vaccination strategies. An optimal vaccination strategy may be similar to the one we observed in the 2-epitope case: a combination of multi-epitope vaccines with different valences m . Such a mixed use of multi-epitope vaccines may simultaneously reduce transmission at the population level and reduce the probability of fixation of vaccine-resistant variants in immunocompromised individuals. However, for a case with several epitopes such a strategy may be unrealistic to implement in practice and we show later that considering a mix of vaccines of different valences in the same vaccination campaign does not provide substantial advantages (SFig. 3). We therefore primarily consider a simpler set m -epitope mixed vaccination strategy, whereby all individuals are vaccinated by different vaccines with the same valence, m . Thus, individuals receive a vaccine against a different set of epitopes but always the same number of epitopes. Under such strategy the herd immunity threshold in the population is achieved through the vaccination of individuals by vaccines protecting against different pathogen variants (related to the concept of the diversity threshold introduced by [KL12]). The n epitopes can be distributed to a total of $C(n, m)$ groups of m epitopes per vaccine, thereby generating a high diversity of individuals inoculated by different vaccines, with different individuals protected against and susceptible to infection by different variants. We can define the m -epitope herd immunity threshold towards a variant

with i mutant epitopes, by calculating the threshold for which $R_i \geq 1$ (Eq. 3.7):

$$V_H(m, i) = \frac{1 - 1/R_0}{1 - C(i, m)/C(n, m)}. \quad (3.16)$$

In an infinite population and at the limit of strong selection, in the host vaccinated with an m -epitope, the establishment of a variant that carries all m resistance mutations, $m = i$, is guaranteed. Therefore, we first looked at the efficacy of different vaccination strategies against vaccine-resistant variants when mutant epitopes were already present in the population. The variants spread faster in populations vaccinated with low-epitope vaccines, however, all vaccines were equally good at preventing the spread of the wildtype pathogen and all were equally ineffective against the super-resistant variant (Fig. 3.6a). Similarly, high-epitope vaccines lead to a lower herd immunity threshold when the population is infected with variants with an intermediate number of mutated epitopes (Fig. 3.6b).

If evolution within the immunocompromised individual is rapid, and consequently the probability of establishment of the super-resistant variant is high, the best vaccination strategy is the one that diversifies the immune response types in the population and uses the m -epitope vaccine strategy with optimal valence $m = \lfloor \frac{n+1}{2} \rfloor$, where $\lfloor \frac{n+1}{2} \rfloor$ denotes the floor function (Fig. 3.6c, probability of establishment derived from Eq. 3.13 in the limit of $m = i$ and conditioned on transmission $p_{\text{trans}}(m) = 1$). At the extreme cases of $m = 0$ no vaccines are administered, and when the vaccine is completely ineffective ($m = n$), the probability of establishment of the strain will be driven purely by genetic drift, as follows from Eq. 3.13, independent of V . For intermediate values of m , higher vaccination rates can greatly reduce, and even eliminate (when $V = 1$) the probability of establishment of the resistant variant (Fig. 3.6c). This is due to the different ways intermediate valence vaccines with $1 < m < n$ can be combined (Fig. 3.3), whereby the evolution of resistance towards one such combination of epitopes does not imply resistance to all combinations. In the same limit, increasing the total number of epitopes n , the probability of establishment decreases until it saturates at the level of $1 - 1/(R_0(1 - V))$, which is the probability of establishment of a wildtype in a vaccinated population. This threshold is approached fastest when $m = \lfloor \frac{n+1}{2} \rfloor$ (Fig. 3.6d).

We then considered the model when only the wildtype pathogen was present initially in the population. As in the case of 2-epitopes, the optimal vaccination strategy in the case of n -epitopes depends on the life-history attributes of the pathogen and its evolution in the immunocompromised host. Having more than 2 epitopes has two effects on pathogen establishment as a function of population level transmissibility and within host evolution: (1) it is harder to evolve resistance to an n -epitope vaccine as it will take longer for variants with a mutation in all epitopes to fix and (2) m -epitope vaccines can be combined to create versatile combinations that reduce the probability of establishment in the population by diversifying the immune response of the individuals in the population. Thus, pathogens with faster rates of evolution in the immunocompromised individuals and higher rates of transmission may be contained by vaccines targeting m out of n epitopes. The best protection against pathogens with a fast rate of evolution in the immunocompromised individuals is achieved when $m_{n/2} = \lfloor \frac{n+1}{2} \rfloor$, while highly transmissible diseases with low rates of evolution are best counteracted with a full-epitope vaccine (Fig. 3.7).

When R_0 and τ are high, no vaccination strategy prevents the establishment of a resistant variant (Fig. 3.8a). The benefits of optimal vaccination strategies are apparent only within certain ranges of R_0 and τ parameters. Specifically, the optimal strategy is substantially better

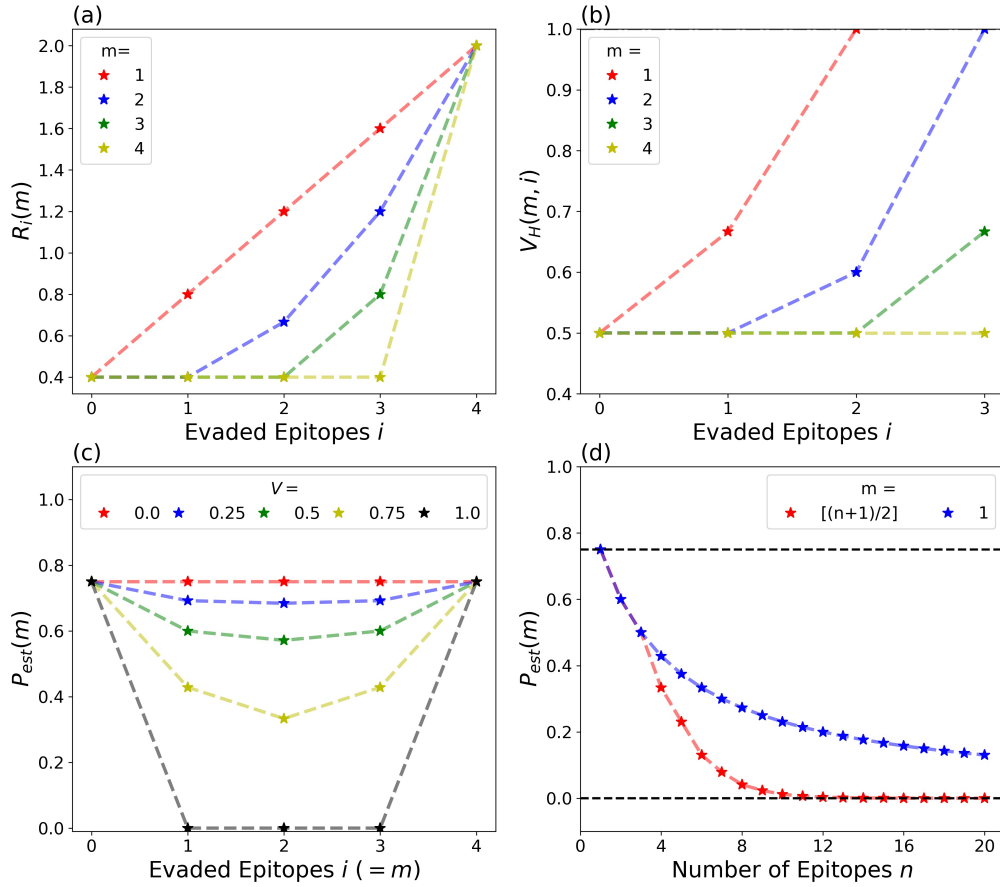


Figure 3.6: Effect of different vaccine strategies in preventing the establishment of vaccination-resistant variants. (a) The rate of spread of a pathogen with i mutations, indicated by the reproductive number $R_i(m)$, of variants carrying i mutant epitopes (X-axis) (Eq. 3.7) for different valencies m of the vaccination strategy (in color). $R_0 = 2$ for wild-type. (b) The herd immunity threshold $V_H(m, i)$ (Eq. 3.16) for the number of mutant epitopes in the variant present in the population (X-axis). $R_0 = 2$ for wild type. (c) The probability of establishment at the population level, $p_{est}(m)$, of a pathogen resistant to one of the vaccines of valence m (X-axis), for different levels of vaccination in population (in color). (d) Probability of establishment at the population level, $P_{est}(m)$, of a pathogen that developed full resistance to one of the vaccines of valence m as a function of number of epitopes n : single-epitope vaccine, $m = 1$ (blue) and optimal vaccine valence, $m = \lfloor (n + 1)/2 \rfloor$ (red). For panels (a) and (d) $V = 0.75$.

than the single-epitope vaccine when R_0 is high and τ is low and better than the full-epitope vaccine when R_0 is low and τ is high (Fig. 3.8b). Any vaccination strategy contains pathogen spread when both R_0 and τ are low (Fig. 3.8b).

For many parameter combinations the optimal vaccine strategy may not be very different from other strategies. Therefore, we determined the near-optimal scenarios with the fewest epitopes per vaccine (the minimal valence strategy) and the near-optimal scenario with the most epitopes per vaccine (the maximum valence strategy), that result in the probability of

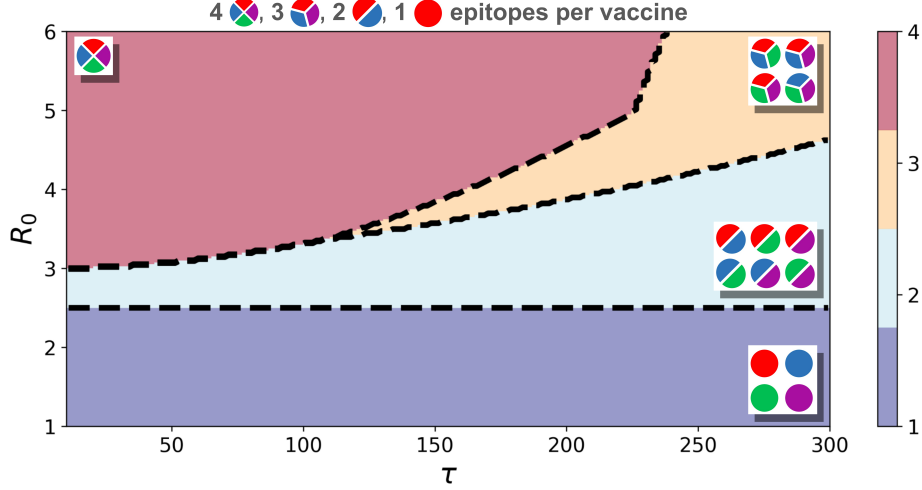


Figure 3.7: The optimal choice of epitopes per vaccine (m , vaccine valence) as a function of disease duration τ and transmissibility R_0 . The graph illustrates the split of the hyper-parameter space for an optimal vaccine strategy for vaccines with $m = 4$ epitopes. Optimal choice of m^* minimizes $p_{\text{est}}(m^*)$, the probability of a pathogen establishment at the population level. For R_0 below ~ 2.5 , single epitopes are sufficient, for low τ and high R_0 full-epitope vaccines provide the lowest probability of mutant establishment within the population. Intermediate, m -valence vaccine strategies are best for other values of τ and R_0 . Importantly, there are areas in which all vaccine strategies have approximately equal efficacy, either equally high for low τ and R_0 or equally low for high τ and R_0 (see Fig. 3.8).

establishment of a mutated pathogen within 10% of the optimal vaccination strategy m^* , $p_{\text{est}}(m) < p_{\text{est}}(m^*) + 0.1$. For the minimal strategy a single-epitope vaccine was efficient for low values of R_0 (Fig. 3.8d) and a higher valence vaccine was always as good as a single-epitope vaccine (Fig. 3.8c). For the maximal valence strategy, the full-epitope approach was efficient when R_0 was high (Fig. 3.8d, 3.8c). For a very specific range of R_0 values, around 2.5, a 2-valent vaccine strategy was always the optimal one (Fig. 3.8d, 3.8c). The efficacy of the optimal vaccination strategy was higher when the number of epitopes in the pathogen was high (Fig. 3.8e) and when the population had more vaccinated individuals (Fig. 3.8f).

Imperfect vaccine model Not all pathogens lead to the same type of evolution in immunocompromised individuals as we see in SARS-CoV-2. Similarly, as observed in SARS-CoV-2, vaccines may not be 100% effective in preventing infection, so they can be imperfect [GMNR03, GMNR01, KFM⁺21]. Thus, we used our model to study the optimal vaccination strategy without immunocompromised individuals, but with imperfect vaccines. Under this scenario vaccinated individuals have a chance (ρ) to get infected and have a short period of infection during which they can transmit the virus to others. The same principle of evolution in the vaccinated but infected individuals as in the immunocompromised individuals was applied. Broadly speaking, only a few immunocompromised individuals ($\rho N \sim 1$) were in the population and the pathogen had a substantial time (τ) in these immunocompromised hosts. With imperfect vaccines the time the pathogen spends evolving in a vaccinated individual is short, however, depending on the degree of imperfection of the vaccine, the number of infected vaccinated individuals in which the pathogen evolves can be large.

Considering the 2-epitope scenario, when the vaccine is ineffective and herd immunity can

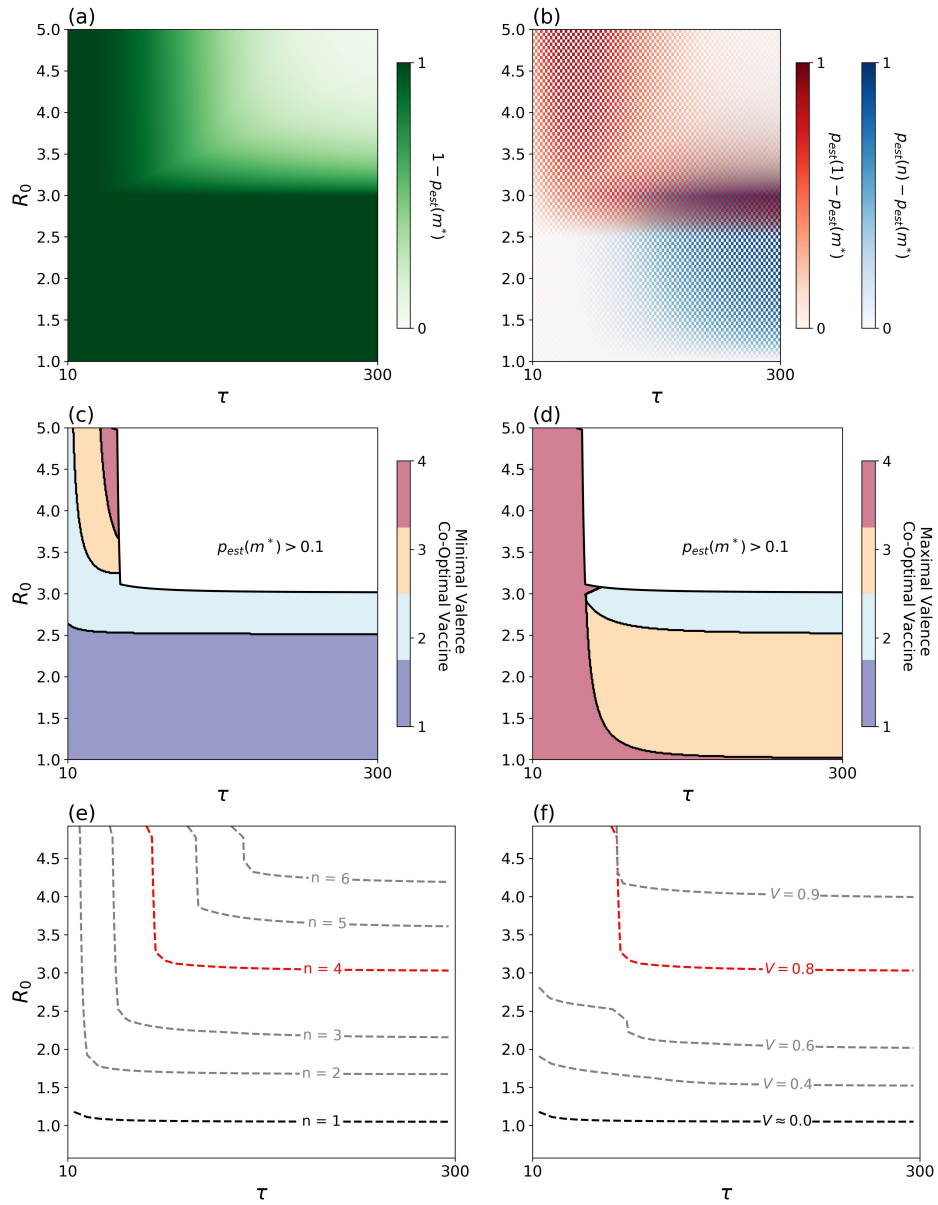


Figure 3.8: Optimal vaccination strategies countering a 4-epitope pathogen. (a) The probability of preventing the establishment of a resistant variant (green) when the optimal vaccination strategy was applied, for different disease duration, τ , and rate of spread, R_0 . (b) The difference between the optimal vaccination strategy and the single-valence vaccine strategy (red) and the full-epitope vaccine strategy (blue). In the overlap region the optimal strategy outperforms both the single-valence and the full-valence vaccine strategies. The near-optimal minimal (c) and near-optimal maximal (d) vaccination strategy. Coloured regions define areas where the different strategies were minimally (c) or maximally (d) near-optimal. The area in white shows the region where the optimal vaccine strategy has a lower than 90% chance to prevent the establishment of a resistant strain ($p_{\text{est}}(m^*) > 0.1$). The same threshold, $p_{\text{est}}(m^*) = 0.1$, is shown as a function of n , the number of epitopes (e), for $V = 0.8$, and (f) as a function of V , with $n = 4$. For all figures, $p = 0.02$.

not be reached, $\rho > 0.5$ for $R_0 = 2$, it does not matter much which vaccine strategy is

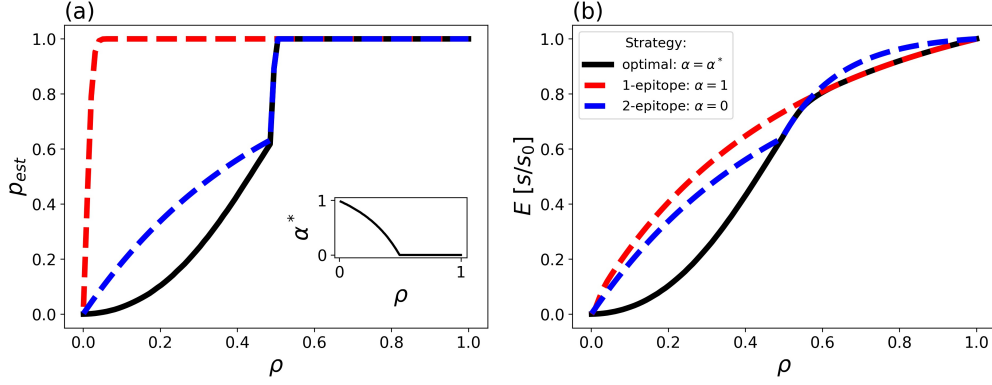


Figure 3.9: Optimal vaccination strategy to reduce evolution of vaccine resistance in the 2-epitope scenario and imperfect vaccination. (a) The probability of establishment of a vaccine resistant variant as a function of ρ for 2-epitope vaccination (blue line), single-epitope vaccination (red line) and optimal mixed vaccination (black line) strategies. The inset graph represents the optimal vaccination strategy, α^* , for which p_{est} is minimal, characterized by α , share of single-epitope vaccines, as a function of vaccine imperfection ρ (fraction of breakthrough infections). (b) The expected epidemic burden, $E[s/s_0]$, which is the average fraction of affected individuals in the population compared to the unvaccinated scenario, as a function of ρ for the same three strategies as in (a). For small values of ρ admixture can considerably reduce the burden. Fixed parameters in these plots: Initially infected population $N = 10^5$, full vaccine rollout ($V = 1$), $R_0 = 2$, $p = 10^{-3}$, $\tau = 5$.

employed because the probability of establishment of a resistant variant will be close to 1. By contrast, when the imperfection of the vaccines is small the optimal strategy greatly reduces the probability of establishment of resistant variants (Fig. 3.8a) and the expected size of the epidemic (Fig. 3.9b). The admixture of full-epitope (2-epitope in this case) vaccines has a strong effect even for very small values of ρ (Fig. 3.9b) even though for these small values of ρ the share of 2-epitope vaccine admixture into the optimal vaccine mix is low (Fig. 3.9a). Thus, when breakout infections are rare (low values of ρ) the best strategy is close to a single-epitope strategy and requires administering only a small number of full-epitope vaccines in the population but following a pure single-epitope strategy even when breakout infections are rare almost certainly will lead to the establishment of a resistant strain (Fig. 3.9b). The reason for this is because mixing a small fraction of a full-epitope into a population otherwise vaccinated with single-epitopes diversifies the population and pushes the herd immunity threshold below 1. This radically decreases the probability of establishment if the number of transmission attempts is high, as in this case ($N = 10^5$). When ρ is higher, the best strategy requires a higher dose of administered full-epitope vaccines.

In the n -epitope scenario we derive the optimal vaccination strategy (m^*) where all individuals in the population receive different vaccines with the same valence. We study the probability of establishment of a vaccine-resistant variant for the optimal vaccination strategy as a function of the degree of imperfection of the vaccine (ρ) and the number of infected individuals in the population (N). When the number of infected individuals and vaccine imperfection are high the optimal vaccination strategy does not prevent the establishment of a vaccine-resistant strain (Fig. 3.10a). Similarly, when N and ρ are low, any vaccine strategy is effective (Fig. 10b). The optimal strategy outperforms the single-epitope vaccine strategy when N is high (red in Fig. 3.10b) and outperforms the full-epitope vaccine strategy when ρ is low (blue in Fig. 3.10b).

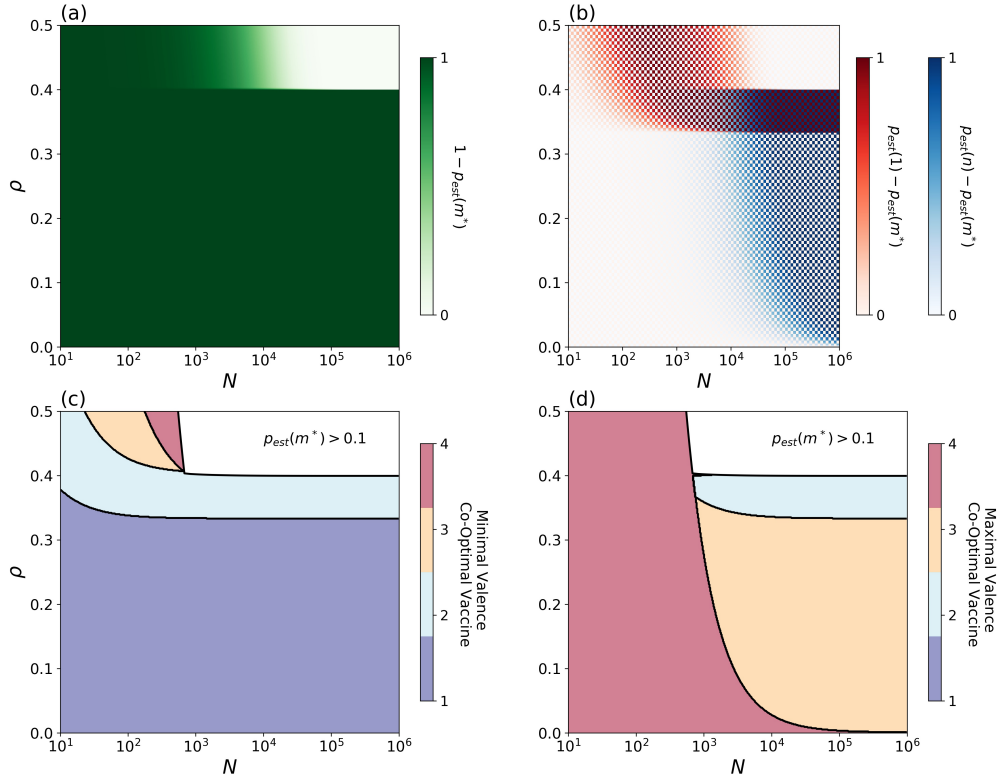


Figure 3.10: Optimal vaccination strategy for imperfect vaccines. (a) The probability of preventing the establishment of a resistant variant when the optimal vaccination strategy is applied. (b) The difference between the single-epitope and the optimal (red) and the full-epitope and the optimal (blue) vaccination strategies. The dark region in the overlapping region is an area of parameter space where the optimal strategy is better than both the single-epitope and the full-epitope vaccination strategies. The parameter values showing the (c) minimal and (d) maximal valence of near-optimal vaccination strategies. Coloured regions define areas where the different m -epitope strategies were (c) minimally or (d) maximally near-optimal. Within the white area the optimal vaccine strategy has a lower than 90% chance to prevent the establishment of a resistant strain ($p_{\text{est}}(m^*) > 0.1$). For this figure $\tau = 10$, $p = 0.05$ and $R_0 = 2$, ρ – degree of imperfection of the vaccine, N – number of infected individuals. When $\rho \geq 1 - (1 - 1/R_0)/V$, no variant, not even the wildtype, can be contained so we show ρ ranging from 0 to $1/R_0$ ($V = 1$).

Similarly to the previous section, we determine near-optimal minimal (the fewest number of epitopes) and maximal (the maximum number of epitopes) vaccination strategies. The solution space for these near-optimal strategies shows a straightforward pattern, with the single-epitope vaccine being the minimum near-optimal strategy for low values of ρ (Fig. 3.10c) while the full-epitope vaccine strategy being maximum near-optimal strategy for low values of N (Fig. 3.10d). The optimal vaccine model better reduces resistant variant establishment when there are multiple epitopes in the pathogen (Fig. 3.11a), when the disease duration in individuals is short (Fig. 3.11b), when the pathogen is not highly infectious (Fig. 3.11c) and when a large proportion of the population is vaccinated (Fig. 3.11d). We did not investigate

the interaction of these parameters.

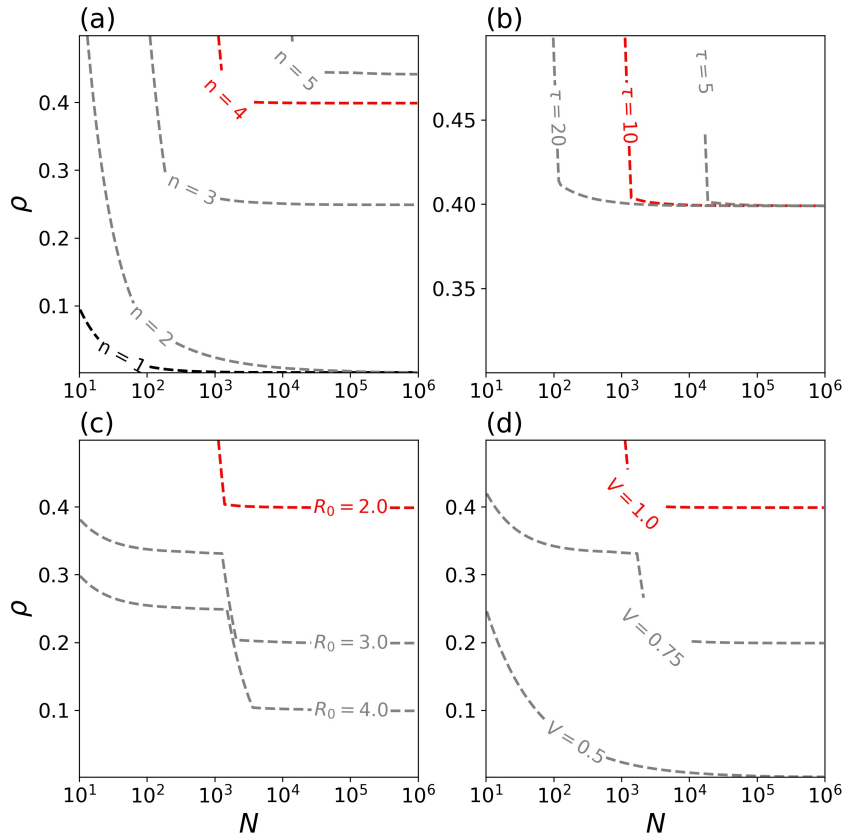


Figure 3.11: Efficacy of optimal vaccination strategy in preventing the establishment of resistant strains. Above the dashed contour lines the optimal vaccine strategy has a lower than 90% chance to prevent the establishment of a resistant strain. The figure shows the probability of establishment of a resistant strain as a function of N and ρ for n (a), τ (b), R_0 (c) and V (d). All red lines correspond to a base parameter choice in the model of $n = 4$, $V = 1$, $\tau = 10$, $R_0 = 2$. Across panels, each parameter (a) n , (b) τ , (c) R_0 and (d) V are varied keeping all the other three parameters constant at the base value.

3.4 Discussion

It has been previously suggested that the application of multiple epitopes, both distributed across the population [MWM21] and within the same individual [Suh97], can be a way of preventing evolution of a vaccine-resistant pathogen. The rationale for their use is that a pathogen will have a harder time adapting to such vaccines. We struggled to come up with a simple term describing the complex vaccination strategies we propose here. In the literature of pesticide and antibiotic resistance the equivalent of a vaccination strategy of inoculating individuals with different single-epitope vaccines is called a mosaic vaccination approach, with this term used by McLeod et al. [MWM21] for the same approach in a 2-epitope model. The full-epitope strategy, also in the antibiotic resistance literature, is referred to as pyramid or combination approach [REX16, REX13]. Meanwhile, mosaic vaccines or multi-epitope vaccines

are those that target more than one epitope, different strains or even different pathogens [Suh97]. The strategies we consider in our model do not fit into any of these definitions and for lack of a better word we just refer to them as complex strategies.

Among the goals of a vaccination campaign against a pathogen is to prevent its evolution towards vaccine resistance. Our results suggest that the optimal vaccination strategy utilizing a combination of different epitope vaccines, is better at reducing the emergence of vaccine resistance than the traditional approach of using one single-epitope vaccine and the full-epitope vaccine strategy when there is considerable selection against vaccine-induced antibodies. This may have broad practical implications for immunologists and policymakers tackling emergent and established pathogens but our results come with a set of caveats. The assumptions and simplifications in our model were motivated either by a reasonable approximation of biological reality or by mathematical simplicity. We prioritized the exploration of the model with a perfect vaccine and very few immunocompromised individuals over the model where all individuals in the population were the same but the vaccine was imperfect. This was done because we were initially motivated by the phenomenon of rapid pathogen evolution in immunocompromised individuals seen in SARS-CoV-2 patients but also because in our analytical model we have an infinite population size, making it mathematically more convenient to deal with a set small number of immunocompromised individuals. However, since all vaccines have less than 100% efficacy [LKRY⁺22, GMNR01], the model analyzing imperfect vaccines is arguably applicable to a wider range of pathogens.

Perhaps the greatest simplification in our model, driven by lack of collective knowledge of the relevant biological processes, is the process of selection of pathogens in an infected individual driven by the immune system. By parsimony we assume that selection against all presented epitopes is equal and that all mutations have an equal probability of emergence. Our model assumes that all different immune types, regardless of the number of immunized epitopes n , clear the pathogen within the same time. Put differently, we treat all vaccines as equally strong, as long as they provide immunity against the pathogen. We also assume that selection by the immune system can be mathematically described by a simple Bernoulli process. We explored three different ways of modeling the selection in the individual driven by the immune system: the Bernoulli model, the Infinite model and the Wright-Fisher model [HC06]. We obtained broadly similar results on a set of test parameters for these three models (SFig. 1) and we selected the Bernoulli model for its mathematical simplicity.

We also assumed that vaccine resistance is caused by a single mutation rather than by a series of mutations on a complex protein fitness landscape of the epitope, as observed in the Spike protein of SARS-CoV-2 [SGH⁺20]. However, there is no general biological understanding of a generic fitness landscape of an epitope that could be used as a model template. Roughly the accumulation of several mutations to achieve resistance should be mathematically equivalent to reducing the mutation rate of a single mutant. Therefore, exploring complex intra-protein fitness landscapes of epitopes should lead to qualitatively different results. However, the fitness landscape of the interacting epitopes may have an influence on our results. In our model, we assume a specific relationship between pathogen fitness and the number of non-mutated epitopes. For example, consider a 2-epitope pathogen case, with an infected individual vaccinated with a 2-epitope vaccine. We assume that in this individual a variant with a mutation in one epitope has a fitness advantage over the wildtype because the mutant pathogen will be recognized by fewer antibodies. More generally, we assume that the relationship between the fitness of the pathogen and the number of mutated epitopes is linear (SEq. 4 and SEq. 5). However, if this relationship is different, our results may be affected. Specifically, if the immune

system is just as effective against a pathogen with a single functional epitope as against the wildtype with multiple functional epitopes, it is likely that the full-epitope vaccination strategy may be optimal in a much larger range of parameter values.

Due to differences in immune detection and presentation [BZA⁺21, RSL⁺22], imprinting [SSK⁺22, YS21] and immunodominance [HDLC17, AAY18, Ado98, HYC⁺22] the actual memory immunity induced by a multi-epitope vaccine may be smaller than m , the number of presented epitopes. Immunodominance, the tendency of the immune system to prioritize producing antibodies to one epitope over others [Ado98, AI12], may have a particularly strong effect on the dynamics of the model. First, the relationship between fitness and the number of mutated epitopes will not be linear. Second, this relationship may be different in different infected individuals. In either case, immunodominance is expected to further reduce the relative efficacy of the full-epitope vaccination strategy: in the extreme case of strong immunodominance and when all individuals are vaccinated with a full-epitope vaccine, such a strategy can cause individual immune systems to choose the same single immune response to a dominant epitope [AAY18]. However, an intermediate m-epitope vaccination strategy under immunodominance will force some immune systems to produce an immune response to a non-dominant epitope, diversifying the immune system responses in the population and ultimately leading to a reduction of the probability of spread of a vaccine-resistant variant. The immune response triggered by a natural infection is expectedly similar to that of a high m-epitope vaccine. Therefore, previous mosaic vaccination might direct natural population level immune protection to become more diverse and resilient in the light of immunodominance.

In the absence of vaccines or for small vaccine rollouts, natural immunity is expectedly the major driver of antigenic evolution. The applicability of our model to scenarios of low vaccine rollout is therefore limited, as natural immunity is not explicitly included in the analysis. Knowledge about the distribution of immune response profiles induced by natural infections could however be implemented in the presented modeling procedure with relative ease. Here we refrained from making any detailed assumptions about such profiles.

In the limit of high vaccine rollout, in particular if V , fraction of vaccinated individuals in the population, is above the herd immunity threshold, classical epidemiology predicts pathogen extinction. We argue that our model stays relevant even within this “safe” regime for the following reasons: (1) An immunocompromised individual could remain infected, even if all contacts have recovered or otherwise immune, (2) vaccines often obey limited effectivity, rendering herd immunity to become unachievable and (3) vaccine hesitancy, slow vaccine rollout and local reservoirs remain as barriers to high levels of population immunity.

Finally, we assumed no recombination in the pathogen, which is not applicable in many pathogens [PLAG⁺15], and relaxing this assumption will influence the results of the model. On an intuitive level, recombination will lead to a faster rate of emergence of strains resistant to multiple epitopes within the host, potentially canceling any fitness reducing effects of resistance mutations [eCHGL07]. In our model, this is equivalent to a larger τ (see Fig. 3.7), thus, we anticipate that an intermediate m-epitope vaccination strategy would be optimal for a greater set of parameters for a pathogen with recombination. It would be interesting to consider a formal model that incorporates recombination, however, it is beyond the scope of our current work.

In sum, different levels of preexisting immunity due to infection, immunodominance, adaptivity, crossimmunity, mutation rates, recombination and pathogen clearance, will influence the results of the basic model, potentially violating assumptions and change the symmetrical outcomes

driven by equal use of different vaccines in the population. Consequently, under more complex real scenarios the best vaccination strategy may not treat all epitopes equally.

The strategy of vaccination of a population with different vaccines to control for risks of evolution of resistant strains has not been studied in detail, however, the concept of controlling risk by diversifying the solution strategy has been used in a wide variety of fields. Perhaps the most impactful example is that of the Modern Portfolio Theory [Mar52] that defined the practical diversification of stocks and securities in investment portfolios in a broadly similar manner. The benefits of genetically diverse crops over the genetically uniform monoculture has been appreciated for well over a century [Tor80, DeA08]. The use of genetically diverse crops leads to higher yields, less damage from parasites [WPY⁺21] and ensures overall food supply stability [RT19]. At the extreme, monocultures are susceptible to drastic out of control epidemics, with the Irish potato famine [Gib22] and the Panama disease, that struck banana production in the 1950's and with an evolved strain threatening banana production today [Plo15] being notable examples. In fact, similar results to ours were obtained in modeling studies tailored to agricultural systems [DDMF17, RPB⁺18, MMB15]. Diversification on a genetic level is also common in nature, with many species practicing a bet-hedging strategy [GSSK15, Sim11, CMR10] to minimize the risk associated with uncertainties in the future. These bet-hedging strategies allow the species to deal with uncertainties of progeny dispersal or environmental variability, but perhaps the most pertinent examples are of disease-host interactions. The benefits of genetic diversity of immune response of the population have been the subject of study for many species, which show that increased diversity of immune response increases the chances to control the spread of the disease in the population [Som05, CLN⁺18, AK15, vHEB⁺16, KL12, UKS⁺10] and within whole ecosystems [HHRM11, SO01], ultimately shaping the co-evolution of pathogens and hosts [SO01, LD00, GM02, vHEB⁺16].

We are not aware of any ongoing efforts using a mixed vaccine approach, whereby different individuals in the population would receive a vaccine tailored to different epitopes. For some pathogens there may be biological limitations in creating such mixed vaccine batches, therefore, here we will discuss the potential benefits of a mixed vaccination approach only in hypothetical terms. The specific application of such an approach to thwart a particular pathogen would require detailed expertise in that particular pathogen, which is outside our area of expertise. However, if a mixed vaccine strategy is technically feasible to apply against a specific pathogen, the experts working on this pathogen may consider the following conceptual advantages.

Many of the potential benefits of mixed vaccine strategy could be obtained by a single-epitope mixed vaccine approach, whereby the population is inoculated by different vaccines with each inducing an immune response to a single epitope. The R&D and manufacturing of several single-epitope vaccines may be simpler than researching a complex full-epitope vaccine. A mixed single-epitope strategy while having the benefit of reducing the establishment probability of vaccine-resistant strains can also be quickly adapted to evolving threats. Consider a 5-epitope pathogen and a population that is vaccinated by a mix of 5 single-epitope vaccines. If vaccine-resistance to one of the epitopes evolves, the failed vaccine can be discontinued until it can be updated to be effective against the evolved epitope. Meanwhile, an all out outbreak in the population is prevented by the four other still functional single-epitope vaccines and eventually the updated single-epitope vaccine is reintroduced. In case of vaccination of the population with a single 5-epitope vaccine, when it fails it can lead to a serious global infection event that cannot be controlled until a new 5-epitope vaccine is updated. Updating a 5-epitope vaccine may also take a long time, exacerbating the effects of the ongoing outbreak.

Our results show that the advantage of using complex strategies of vaccination is substantially stronger when a high fraction of the population is vaccinated, we believe that the most likely application of such strategies will be outside the human population due to vaccination hesitancy. Furthermore, many people may hesitate receiving a random vaccine, especially if there are any differences, however minor, in their efficacy. Perhaps people will have different reasons to choose different vaccines and the necessary vaccine diversity can be maintained. However, none of these issues apply to animals and it seems likely that initially the application of such complex mosaic vaccinations may be in livestock [FLT⁺19].

Records, Evolution and Innovation

4.1 Introduction

The previous two chapters focused on evolutionary epidemiology – a field that sits at the intersection of two disciplines frequently studied on their own within complex systems research: evolution and epidemiology.

The *probability of fixation* constitutes a fundamental concept in evolution (and population genetics in particular). Assume that an individual with relative selective advantage, $(1 + s)$, gives rise to a number of offspring, n , that follows a Poisson distribution,

$$p(n) = \sum_{n=0}^{\infty} \frac{e^{-(1+s)}}{n!} (1+s)^n \quad . \quad (4.1)$$

The probability of fixation, Q , can then be numerically evaluated with an implicit equation (see [Gav04]),

$$1 - Q = e^{-(1+s)Q} \quad . \quad (4.2)$$

In epidemiology, a similar central concept can be found in the *expected size of a pandemic*, which measures the fatality of a pandemic as the total fraction of recovered individuals at the end of the pandemic, r_{∞} . Not every individual becomes infected and eventually recovered, due to growing herd immunity in the population. Based on the baseline SIR-model (*susceptible-infected-recovered* [KR08]) one can derive r_{∞} for an initially fully susceptible population as

$$1 - r_{\infty} = e^{-R_0 r_{\infty}} \quad . \quad (4.3)$$

Both expressions, Eqs. 4.2 and Eq. 4.3, show a very similar structure and result from an underlying Poisson branching process, which can also be expressed as a percolation problem of disease spread on a random contact network [PSCVMV15]. Rather than viewing a pandemic purely as a sequential, step-by-step spread of contagion, one can also consider the global network topology through which the disease might propagate. In this view, the key question becomes whether the initially infected individual belongs to a giant component that spans

the population. Whether studying genetic drift or disease transmission, one can approach the problem therefore by considering the entire ensemble of possible random networks or branching trees, where the average degree, $\langle k \rangle$, is tuned to match the selection coefficient $(1 + s)$ or the reproductive rate R_0 .

There is a philosophical dimension to this line of thought, reminiscent of Poincaré’s Demon. If we had complete knowledge of the network topology through which a contagion eventually spreads, or the future phylogeny of an organism, we could, in principle, transform a stochastic process into a deterministic one.

This tension between stochastic and deterministic descriptions reappears in other domains governed by evolutionary principles. One particularly compelling example is innovation. At first glance, innovation seems inherently unpredictable, involving the emergence of ideas or artifacts that lie outside the current system. Yet statistical patterns in innovation suggest otherwise: innovations may appear novel to us, but they often emerge from within a larger unknown – yet “out there” – structure of theoretical possibilities that we, as a collective, are continuously exploring.

A key concept here is the so called *adjacent possible* – the subset of all conceivable innovations that are accessible from the current state of the system. Popularized by Stuart Kauffman [Kau00], this idea has inspired several formal models that attempt to characterize the dynamics of exploration and novelty. In this chapter, we will introduce and compare a selection of these models. We will argue that the adjacent possible is shaped not just by combinatorial potential but also by feedback loops and the obsolescence of older forms.

One domain where these ideas converge particularly well is evolutionary theory. As we will show, deep connections exist between models of neutral evolution, urn models of innovation, and the statistics of records.

Outline of this chapter

First, we will review some results from the mathematical field of *record statistics* (the statistics of a sequence of maxima in a growing set of random numbers). To this end we start off with a review of a key model in understanding the adjacent possible, the so called *sample space reducing process* [CMHT15] and show that its dynamics offer an alternative and complimentary view on the mathematics of records. We will discuss several well known key properties of a record series, such as the number of records until time T , the waiting time between records and some extensions from the case of i.i.d. random variables. Along this journey, we will comment on corresponding structures in the case of the sample space reducing process. Some of these connections have not been mentioned in the literature before.

Secondly, we continue with a review of a certain class of models for evolution and models for innovation. Starting with the case of neutral evolution, we introduce the *Dirichlet process*, which itself shows interesting connections to the statistics of records, which have been little appreciated in the literature. We continue with a model of evolution where selection is non-zero, but constant, and show that this model can be compared to a particular instantiation of a broader class of processes, called *Pitman Yor process*, which are extensions of the Dirichlet process. The Pitman Yor process, again, can be extended to cover a class of proposed urn models for innovation [TLSS14]. Record statistics and the sample space reducing process reside in this broader class of models and we argue that models of innovation could potentially be reformulated, assuming a latent record process. Finally, we note a general lack of models,

explicitly incorporating the obsolescence of old innovations, which is a key feature of systems that innovate.

Building on the previous two chapters, we propose a generalized scheme for innovation models, inspired by record statistics, the sample space reducing process, and the Dirichlet process. This scheme allows a classification of different models according to the laws by which the adjacent possible collapses as the system innovates. We propose a simple extension of record statistics, the *record process with stochastic reset*, which allows us to model obsolescence explicitly within this generalized scheme. By the end of this chapter, we hope to develop a model, which is capable to reproduce the statistics of systems that innovate.

We close with some potential fields of application, some of which we will explore in the next chapter.

4.2 Record Statistics

The concept of the adjacent possible describes the space of novel configurations – be they ideas, technologies, biological forms, or social constructs – that become attainable when a new discovery or innovation occurs. It represents the evolving frontier between what exists and what could plausibly emerge next. Crucially, the adjacent possible does not encompass all imaginable futures; rather, it defines the subset of next-step possibilities that are reachable from the current state. This space expands and retracts as new elements are introduced, constantly reshaping the landscape of potential innovation.

4.2.1 Sample Space Reducing Processes

The adjacent possible can be understood as a dynamic pool of options. The process by which innovation unfolds exhibits strong *path-dependence*, where each present innovation could constrain the emergence of future innovations. A useful conceptual framework to describe this is the *sample space reducing process* (SSRP) [CMHT15]. In this framework, the accessible outcome space shrinks stochastically over time as new states are visited and others become excluded.

A SSRP is typically defined with the metaphor of a stair case (see Fig. 4.1a): The stair case has N steps. Each step represents a state $x \in \{1, \dots, N\}$ that can be visited by a ball, which jumps from state to state. A ball at state x_t can only transition downwards, $x_{t+1} < x_t$, but aside from this constraint, picks any state with equal probability, $p(x_{t+1}|x_t) = 1/(x_t - 1)$ for $x_{t+1} < x_t$ and $x_t \neq 1$. Once the ground state, $x_t = 1$, is reached, the process is reset to a random state, $p(x_{t+1}|x_t = 1) = 1/N$. This defines a state-transition matrix, $p(x_{t+1}|x_t)$, for which the long term steady state can be computed in order to arrive with the marginalized state visiting probability $p(x)$. This computation yields that

$$p(x) = \frac{p(1)}{x} \quad , \quad (4.4)$$

where $p(1)$ is the visiting probability of the ground state $x = 1$. The state visiting probability follows a power law with exponent -1 , also often referred to as *Zipf's Law* in the literature [Zip16]. This baseline formulation of the SSRP is also called the *slowly driven SSRP*, as resets only occur once the ground state is reached (Fig. 4.1b).

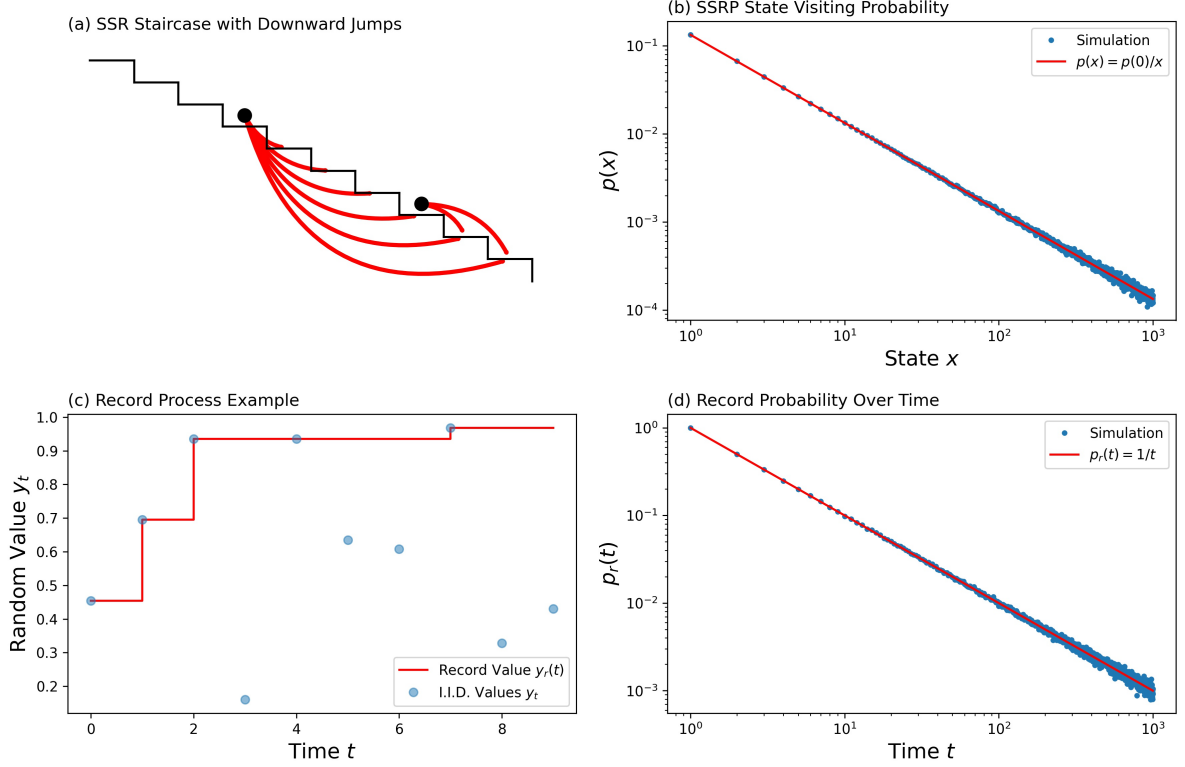


Figure 4.1: **Sample Space Reducing Processes (SSRP) and Record Processes (RP):** **(a)** The SSRP can be conceptualized as a ball (black) transitioning to lower states on a staircase. Any lower state is reached from the current state with equal probability, indicated by red lines. Once the ground state is reached, the ball is reinitialized at a random state, x . **(b)** The steady state probability for the ball in (a) to visit a state x is following a power law with exponent -1 . **(c)** shows an example of an i.i.d. record process. The blue dots are i.i.d. random values, y_t , one sampled for any of the ten time steps t shown. The red line is the record curve $y_r(t)$. **(d)** The probability that a record is obtained in an i.i.d. RP at time t decays as a power law with exponent -1 .

We further define a formulation of the SSRP in a continuous state space. As the stair case step sizes become infinitesimally small and as the number of steps becomes very large $N \rightarrow \infty$, the stair case takes the shape of a continuous slope, where states are defined on the rescaled continuous space $x \in [0, 1)$. The SSRP on such an infinitely large stair case would never reach the ground state, and is therefore never restarted. To prevent this, we can define a small value x_∞ , below which a restart is defined to occur and send $x_\infty \rightarrow 0$.

In the next subsection we will show that this framework is naturally connected to the statistics of record-breaking events.

4.2.2 The Statistics of Records in i.i.d. Random Variables

Records, which are events that exceed all previous observations, have long captured the human imagination, as seen in the enduring popularity of the Guinness World Records. They are also mathematically well-studied phenomena [Cha52, Nev00, MS24]. In the simplest setting, consider a sequence $\{y_0, y_1, \dots, y_t, \dots\}$ of independent and identically distributed (i.i.d.) random variables, y , sampled from a continuous distribution, $p(y)$, also referred to as the

parent distribution. We call $y_r(t) = \max_{i < t} y_i$ the current record value and refer to the process, which gives rise to such records, as *record process* (RP) (Fig. 4.1c). The probability that the t -th value, y_t is a record, $p_r(t) := p(y_t > \max_{i < t} y_i)$, is given by a surprisingly simple and universal formula:

$$p_r(t) = \frac{1}{t} \quad (4.5)$$

This result follows from symmetry: any of the t i.i.d. samples is equally likely to be the maximum. Therefore, in the absence of any other information, the last value has a $1/t$ chance of being a record (Fig. 4.1d). This universality holds for any continuous underlying distribution $p(y)$. The symmetry argument can further be extended to show that correlations between record events vanish, $p_r(t_0, t_1, \dots, t_m) = p_r(t_0)p_r(t_1) \cdots p_r(t_m)$ [Wer12]. All permutations of ranks among the first t_m observations are equally likely, and the probability that the specified indices are successive maxima factorises to $1/(t_0 \cdots t_m)$.

From the perspective of the adjacent possible, each record can be interpreted as a novel achievement or innovation – an expansion into unexplored territory. Yet, each new record also constrains future possibilities by setting a higher bar for what counts as novelty. The probability of discovering a new record, $p_r(t)$, decreases over time, effectively shrinking the adjacent possible. Consequently, the expected waiting time, $\langle \tau \rangle$, for a new record grows.

We define the cumulative variable $x_r(t)$ as the probability to find a value larger than the current record, $y_r(t)$,

$$x_r(t) = \int_{y_r(t)}^{\infty} p(y) dy \quad , \quad x_r \in [0, 1] \quad . \quad (4.6)$$

The observation of new records y_r translates into a continuous SSRP for the variable x_r as defined above, where each record y_r halves the probability to find a new record on average.

Moreover, the apparent structural correspondence between the record probability $p_r(t)$, Eq. 4.5, and the state visiting probability $p(x)$, Eq. 4.4, in the sample space reducing (SSR) process can be made mathematically precise [Yad17]. Given that a record occurs at time t , we can ask: What is the probability that the previous record occurred at time $\tau < t$? Using the result above, this probability is

$$p(y_\tau = \max_{i < t} y_i \mid y_t > y_\tau) = \frac{1}{\tau} \cdot \left(1 - \frac{1}{\tau + 1}\right) \cdot \left(1 - \frac{1}{\tau + 2}\right) \cdots \left(1 - \frac{1}{t - 1}\right) \quad (4.7)$$

where the product conveniently collapses to become

$$p(y_\tau = \max_{i < t} y_i \mid y_t > y_\tau) = \frac{1}{t - 1} \quad (4.8)$$

This expression, $p(\tau \mid t)$, exactly replicates the uniform transition probabilities, $p(x_{t+1} \mid x_t)$, as defined in the SSRP, revealing a mathematical equivalence between both processes. In fact, the record process for i.i.d. random variables follows the same statistics as the SSRP, but in "reverse time". Note, that it is considerably harder to derive Eq. 4.4 only considering the language of the SSRP [CMHT15], while it is easily seen using a symmetry argument if one adopts the language of the record process.

This duality allows results developed within the SSRP framework to be directly transferred to the domain of record statistics. In the following subsections, we will explore several general properties of this minimal model.

4.2.3 Properties of SSRP and RP

The number of records and the number of jumps

It is well known that the expected number of records in a sequence of T i.i.d. random variables grows logarithmically with T . This can be shown by summing the individual probabilities $p_r(t)$ for each time step $t = 1, 2, \dots, T$,

$$\langle R(T) \rangle = \sum_{t=1}^T \frac{1}{t} \approx \log T + \gamma, \quad (4.9)$$

where $\gamma \approx 0.577$ is the Euler – Mascheroni constant.

An equivalent logarithmic scaling holds for SSRPs. If the initial system has N accessible states, the expected number of steps required to reach the ground state is also approximately $\log N$.

To uncover the full distribution of the number of records (or steps in the SSRP), we employ the method of generating functions [Wer12]. The generating function of $R(T)$ is given by

$$\sum_{m=1}^{\infty} p(R(T) = m) z^m = \langle z^{R(T)} \rangle. \quad (4.10)$$

Next we define the independent indicator variables σ_t , which are 1 if a record was obtained at time t , and 0 otherwise, allowing us to rewrite the generating function as

$$\langle z^{R(T)} \rangle = \langle z^{\sum_{i=1}^T \sigma_i} \rangle = \prod_{i=1}^T \langle z^{\sigma_i} \rangle. \quad (4.11)$$

In the last step we used the independence of the σ_i . We know that σ_i is 1 with probability $1/i$ and can therefore conclude

$$\prod_{i=1}^T \langle z^{\sigma_i} \rangle = \prod_{i=1}^T \left(\frac{z}{i} + \left(1 - \frac{1}{i}\right) \right) = \frac{1}{T!} \prod_{i=1}^T (z - 1 + i). \quad (4.12)$$

The last product can be identified as the generating function of the unsigned Stirling numbers of the first kind, $s(T, m)$, an expression, which counts the number of permutations of T elements with exactly m cycles,

$$\prod_{m=1}^T (z - 1 + m) = \sum_{m=0}^T s(T, m) z^m. \quad (4.13)$$

A *cycle* in a permutation, π , is an ordered sequence $(a_1 a_2 \dots a_k)$ such that $\pi(a_i) = a_{i+1}$ for $1 \leq i < k$ and $\pi(a_k) = a_1$. The probability of observing $R(T)$ records therefore shows deep

connections to combinatorics and is given by the fraction of permutations with $R(T)$ cycles out of all possible permutations, $T!$,

$$p(R(T) = m) = \frac{s(T, m)}{T!} . \quad (4.14)$$

The same expression holds for the distribution of the number of jumps in the SSRP. In the limit of large T , this distribution converges to a *log-Poisson distribution* [SJ13]:

$$p(R | T) \approx \frac{1}{R!} [\log T]^R e^{-\log T} = \frac{1}{T \cdot R!} [\log T]^R , \quad (4.15)$$

which naturally reproduces the expected mean $\langle R(T) \rangle \sim \log T$ and further shows that $\text{Var}(R(T)) \sim \log T$.

The distribution of waiting times and jump sizes

Next we want to know the distribution of *waiting times*, ℓ , between consecutive records. We can make use of the fact that this corresponds to the number of cycles of length, ℓ , in a random permutation. There are $\binom{T}{\ell}$ ways to choose ℓ elements, and among those, there are $(\ell - 1)!$ ways to arrange them into a cycle. The number of permutations of the remaining $T - \ell$ elements is $(T - \ell)!$, so the number of permutations where a specific ℓ -cycle is present is

$$\binom{T}{\ell} (\ell - 1)! (T - \ell)! = \frac{T!}{\ell} . \quad (4.16)$$

Now, since there are $T!$ total permutations, the probability that a given subset of size ℓ forms a cycle is $1/\ell$, and consequently the probability of observing a waiting time,

$$p(\ell) = \frac{1}{\ell} . \quad (4.17)$$

This result mirrors the distribution of jump sizes in the SSRP [Yad16], which is an observation, that has not been explicitly stated in the literature to date.

Renormalization and Self-similarity

Records processes are self-similar. Indeed, one can define a renormalization scheme to demonstrate this explicitly. Suppose we coarse-grain the time axis into supercells of size m , such that each supercell contains m consecutive samples. If we define a "coarse-grained record" as occurring when the maximum of a supercell exceeds all previous values, then the coarse-grained process still follows the same power-law decay, with time rescaled as $t \rightarrow t/m$. This follows from the fact that the maximum of m i.i.d. random variables is again an i.i.d. random variable, though drawn from a different distribution. For example, if the original values x_i are sampled from the uniform distribution on $[0, 1]$, then the maximum of m such values follows a Beta distribution:

$$P\left(\max_{i < n} x_i\right) = \text{Beta}(x | m, 1) = mx^{m-1} , \quad (4.18)$$

which remains i.i.d., but is increasingly skewed toward 1 as m increases. Hence, the coarse grained record process preserves its essential statistical structure under rescaling, namely that the probability of a record decays as $1/t$, regardless of the coarse graining of the data.

Growing Populations and Step Sizes

If the number of samples drawn at each time step is not constant but instead increases over time, the statistics of records are modified accordingly. Let $m(t)$ denote the number of samples drawn at time t . The probability that at least one of these $m(t)$ samples exceeds all previous values – i.e., is a record – is given by

$$p_r(t) = \frac{m(t)}{\sum_{\tau=1}^t m(\tau)} . \quad (4.19)$$

Now, consider the case where $m(t)$ grows as a power law, $m(t) = t^\alpha$, with $\alpha > 0$. Then,

$$p_r(t) \sim \frac{1}{t} \quad (4.20)$$

for large t , and where we have approximated the sum in Eq. 4.19 with an integral. Only in the limiting case, if $m(t)$ grows exponentially, $m(t) \sim e^{\beta t}$, does the record probability approach a constant,

$$p_r(t) \sim \text{const.} \quad (4.21)$$

This result can be interpreted as a transition between sublinear (diminishing) and superlinear (persistent) record regimes.

A parallel result holds in the SSRP – although it is harder to show if one restricts the analysis to the language of the SSRP [CMHT16]. There, the analogue of increasing $m(t)$ is drawing from a broader or more concentrated prior distribution over state visits. It might be helpful to think that the length of steps on the stair case is not constant in this case. If the prior becomes increasingly biased toward high value states (analogous to increasingly drawing more samples per time), the SSRP takes a longer time to reach the ground state.

For the general case, let's call $S(t+1) = \sum_{\tau=0}^{t+1} m(\tau)$ and together with Eq. 4.19 we can write

$$S(t+1) = S(t) + m(t+1) = S(t) + p_r(t+1) S(t+1) . \quad (4.22)$$

After rearranging, we recover the recursive relation

$$S(t+1) = \frac{S(t)}{(1 - p_r(t+1))} , \quad (4.23)$$

which can be used to evaluate $m(t)$ for various desired $p_r(t+1)$. This recipe can be considered as a new complementary procedure (to [CMHT16]) for evaluating the statistics of a SSRP with arbitrary prior distribution.

4.2.4 Extensions of Record Statistics

The statistics of records were extended in multiple ways to incorporate both time-dependent [Wer12] and correlated random variables [MPS20]. Franke, Wergen and Krug [FWK10] showed that the exact choice of the parent distribution does matter if the parent distribution is not stationary, but is drifting or changing its variance [Wer13]. Notably, they found solutions for the typical tail behavior of all three limiting distributions of extreme value statistics [DHF06] if the distribution is moving with a constant speed $v \geq 0$,

$$p_t(y) = p_0(y - vt) \quad , \quad (4.24)$$

where p_t is the probability density of the random variable y at time t . For the case of the exponential family of *Gumbel* class, the record probability of a moving parent distribution takes a particularly simple form [FWK10],

$$p_r(t) = \frac{1 - e^{-v}}{1 - e^{-tv}} \quad . \quad (4.25)$$

For $v \rightarrow 0$, the standard result of $p_r(t) = 1/t$ is obtained. Note that a proper rescaling also leaves this distribution invariant under coarse-graining into supercells of size m . One can show that Eq. 4.25 stays invariant if time and speed are renormalized as $v \rightarrow mv$ and $t \rightarrow t/m$. On a broader scale the same statistics are observed but assuming a faster speed v . In an important application of this extended RP, Wergen has successfully shown that temperature records in our current climate can only be understood assuming an underlying shifting temperature distribution [WK10].

For the case of correlated random variables, the most studied system is that of the random walk. The random walk is strongly correlated since the state of the random walker depends its previous position. The record probability of the random walk can be evaluated to be related to the average growth of the variance of the random walk, approximately following a power-law with power $-1/2$ [MPS20],

$$p_r(t) \sim \frac{1}{\sqrt{t}} \quad . \quad (4.26)$$

For a compilation of many more interesting extensions we will refer the reader to the extraordinary doctoral thesis of Gregor Wergen [Wer12, WHK14].

In this section, we have reviewed the sample space reducing process, the statistics of records, their properties and considered some extensions of the baseline i.i.d. RP. In the next chapter, we will introduce some models, which are used to understand evolution and innovation. Interestingly, for certain parameter choices, these models reduce to the i.i.d. RP. This situates the introduced RP and SSRP within a broader class of innovation models.

4.3 Models of Evolution and Innovation

Record statistics have also been used to model evolutionary dynamics on random, fully-epistatic fitness landscapes [JK05]. The central idea is that evolution – often summarized as "survival of the fittest" – can be viewed as a process in which the current genotype is replaced by one of higher fitness, analogous to a record-breaking event in a sequence of random variables. When

the fitness landscape is assumed to be uncorrelated and random, the evolutionary dynamics closely resemble those of record statistics in the i.i.d. case. Before turning to evolution under selection, however, it is instructive to examine neutral evolution, where the fitness landscape is flat and all genotypes are equally fit. This selection-free regime is where many foundational results in population genetics originate.

4.3.1 Neutral Evolution and the Dirichlet Process

Coalescence and the Chinese Restaurant Process

One of the simplest models of neutral evolution is the *Wright-Fisher model* [Gil04b], which assumes random replacement of individuals in a population of fixed size N . In this model, eventually a single genotype will dominate the entire population – a process known as *fixation*. The probability that a given genotype eventually fixes is $1/N$, since all genotypes have equal fitness and thus equal initial chance of becoming fixed.

The ancestry of the fixed genotype can be traced backward in time, forming a genealogical tree until the origin of the simulation. If mutations are introduced into the model, such trees may coalesce to different ancestors, resulting in multiple distinct lineages in the population at a given time [Kin82]. In the limit of a large population, $N \rightarrow \infty$, the distribution of the number of distinct genotypes (trees originating at unique mutations) present at a given time can be described analytically and is known as *Ewen's Sampling Formula* [Ewe72].

In order to understand this better – and as a step towards deriving Ewen's sampling formula (we will not derive it fully in this work, see [EE04] for a complete derivation)–, we start with a population at time t and ask, (1) what is the rate at which two individuals in a population sample of size k originate from the same ancestor at time $t - 1$ (they *coalesce*), and (2) what is the rate at which an individual in the sample at time t originated from a de-novo mutation.

Given a typically high dimensional genotype space, it is often assumed that every mutation is creating a unique and never-before-seen genotype. The rate at which such (neutral) mutations arrive grows linearly with k , as every individual mutates with the same probability,

$$r_m(k) = \mu k \quad , \quad (4.27)$$

where μ is a parameter which tunes the mutation rate. The rate at which two individuals coalesce on the other hand is proportional to the number of ways in which two individuals can be picked within the sample of size k ,

$$r_c(k) = \binom{k}{2} \frac{1}{N} = \frac{k(k-1)}{2N} \quad . \quad (4.28)$$

If we rescale $\mu = \theta/2N$, we ensure that both rates scale with N in the same way and θ is a rescaled mutation rate. Consider now a process in which the sample of size k grows at every timestep to become a sample of size $k + 1$, either via a coalescence event or a mutation event. The probabilities for each of the two possibilities can be computed using the above rates.

$$p_m(k) = \frac{r_m(k+1)}{r_c(k+1) + r_m(k+1)} = \frac{\theta}{\theta + k} \quad (4.29)$$

$$p_c(k) = \frac{r_c(k+1)}{r_c(k+1) + r_m(k+1)} = \frac{k}{\theta + k} \quad (4.30)$$

This time-forward process where the sample continuously grows, is often modeled by the *Hoppe urn* [Hop84] or the *Chinese Restaurant Process* (CRP) [Pit95].

The Hoppe urn is a particular version of a broader class of thought experiments, called urn models. One of the easiest versions of an urn model is given by the *Pólya urn*. Consider a container, or an *urn*, which is initially filled with N balls of two different colors, where each color is initially present at abundances, a and b , such that $N = a + b$. At every time step in the model, a ball is picked at random and replaced along with an additional ball of the same color. The Hoppe urn represents an extension of the Pólya urn, where an additional colorless "mutation" ball exists, which has a mass θ , indicating the relative weight of picking it. If $\theta = 1$ the mutation ball is equally likely picked as any other ball. In general the probability of picking the mutation ball is given by Eq. 4.29 where k is the number of colored balls in the urn (excluding the colorless mutation ball). If the mutation ball is chosen, it is replaced along with an additional ball of an entirely new color, which was not present in the system before. For all other balls the replacement rule stays the same as in the Pólya urn.

The CRP is an equivalent formulation of this idea, in which customers enter an initially empty restaurant one after each other. The first customer picks any of the tables. The second customer either joins the first customer on their table, or starts a new table. In the general case, if k customers have entered the restaurant already, the probability for customer No. $k + 1$ to start a new table is given by Eq. 4.29. (The name of the CRP reinforces clichés about a nationality and I would have liked to avoid using this name, but unfortunately, this is the name used in the literature since ages and its use has not yet become obsolete.)

The CRP shares connections with RP and SSRP. In both frameworks, new "types" (genotypes, tables, or records) emerge with a probability that decreases over time, modulated by a tunable parameter θ which represents the mutation rate. If $\theta = 1$, the CRP reduces to the standard record statistic behavior observed in i.i.d. random variables. More generally, θ biases the rate at which new types appear: When $\theta \ll 1$, new types are rare, and existing types dominate. When $\theta \gg 1$, novelty is frequent.

From this one can calculate the distribution of the number of distinct types $R(k)$, $p(R(k))$, specifically genotypes, in a sample of size k taken from an infinitely large population. We can derive this using a similar method as already introduced for the number of records, using a generating function and the indicator function σ_i , which is 1 with probability $\theta/(\theta + i - 1)$,

$$\langle z^i \rangle = \prod_{i=1}^k \langle z^{\sigma_i} \rangle = \prod_{i=1}^k \left(\frac{z\theta}{\theta + i - 1} + \frac{i - 1}{\theta + i - 1} \right) = \frac{1}{\theta \cdot (\theta + 1) \cdots (\theta + k - 1)} \sum_{i=1}^k (s(k, i)\theta^i) z^i. \quad (4.31)$$

In the last step we have again used the definition of the Stirling numbers Eq. 4.13. We therefore find the density of type counts, R , to be

$$p(R(k) = m) = \frac{\Gamma(\theta)}{\Gamma(\theta + k)} s(k, m) \theta^m, \quad (4.32)$$

which becomes Eq. 4.14 for $\theta = 1$.

The full distribution of genotype counts is governed by Ewens' sampling formula, a fundamental result in population genetics (see [EE04] for a full derivation),

$$p(a_1, a_2, \dots, a_k \mid \theta, k) = \frac{k! \cdot \Gamma(\theta)}{\Gamma(\theta + k)} \prod_{i=1}^k \frac{\theta^{a_i}}{i^{a_i} a_i!}, \quad (4.33)$$

where a_i is the number of types present i times in a sample of size k , such that

$$\sum_{i=0}^k a_i i = k. \quad (4.34)$$

Ewens' formula also has a correspondence in record statistics, but to understand it, we first have to introduce the *Dirichlet Process*.

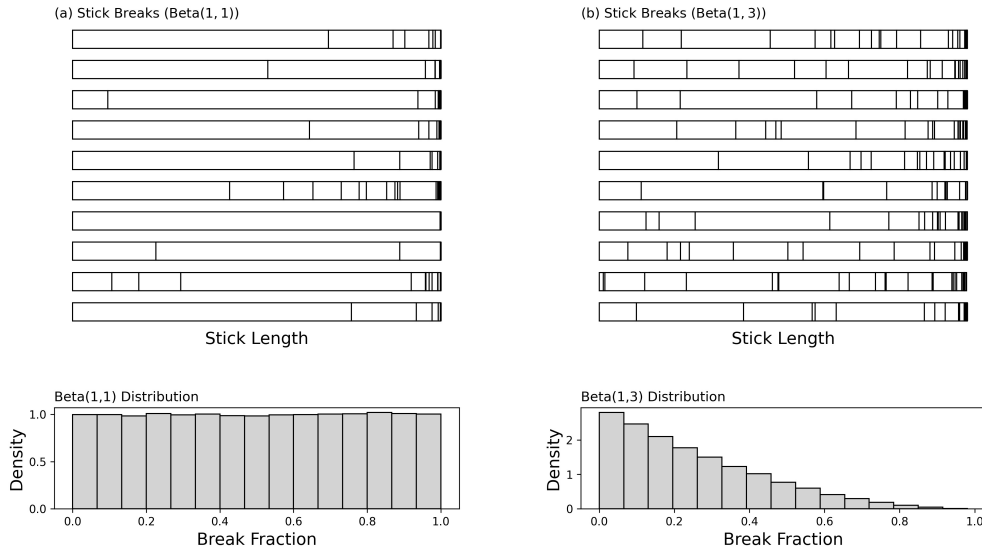


Figure 4.2: **The Dirichlet Process and the Stick Breaking Construction:** The stick breaking process is exemplified for $\theta = 1$ **(a)** and for $\theta = 3$ **(b)**. A stick is broken consecutively (indicated by vertical black lines on each of the ten+ten sticks), such that the break point on the remaining stick fractions is determined by a $\text{Beta}(1, \theta)$ distribution. For $\theta = 1$, samples from the beta distribution are uniformly distributed **(a)**. For $\theta > 1$ samples from the beta distribution are skewed to values near zero **(b)**, effectively decreasing the average size of a piece from the stick.

The Dirichlet Process

The CRP possesses a key property known as *exchangeability*. This means that the probability of a particular configuration does not depend on the order in which customers are assigned to tables. We can verify this by observing that the probability of any particular seating arrangement depends only on the final partition structure, not on the sequence of arrivals. For general random variables X_i exchangeability can be expressed mathematically as

$$p(X_1 = x_1, \dots, X_n = x_n) = p(X_{\sigma(1)} = x_1, \dots, X_{\sigma(n)} = x_n) \quad (4.35)$$

for any permutation σ of $\{1, 2, \dots, n\}$. Since Ewen's sampling formula Eq. 4.33, does not explicitly depend on the order in which customers arrive, but only on the configuration $\{a_1, a_2, \dots, a_k\}$ it can be proofed that this property holds for the CRP.

In such cases a fundamental result from Bayesian inference applies: *de Finetti's Theorem* [Ald06]. This theorem states that any infinitely exchangeable sequence of random variables, $\pi = \{x_1, x_2, \dots\}$, can be represented as a mixture model, that is, there exists an underlying distribution from which the observed data are conditionally independent and identically distributed (i.i.d.) samples,

$$p(\pi) = \int_{\phi} \prod_{i=1}^{\infty} p(x_i|\phi) p(\phi) d\phi \quad , \quad (4.36)$$

where ϕ is an auxiliary variable, which mixes possible underlying distributions $p(x_i|\phi)$. In the context of the CRP, this implies the existence of a random distribution of distributions from which observations are drawn, which is called *Dirichlet Process* (DP).

Although the DP is conceptually elegant, its full specification requires integration over an infinite collection of distributions, making exact computation intractable. Fortunately, there exists an alternative and constructive approximation known as the Stick-Breaking Process (SBP).

In the SBP, a unit-length stick is recursively broken into pieces, with each piece representing the proportion of the population (or customers) assigned to a particular genotype (or table). The lengths of the fragments are drawn from Beta distributions, $\text{Beta}(1, \theta)$ (see Fig. 4.2), producing an infinite sequence whose distribution follows the so-called *GEM distribution*, $\text{GEM}(\theta)$ (named after Griffiths, Engen, and McCloskey) [EE04].

More formally, let $V_m \sim \text{Beta}(1, \theta)$ independently for $m = 1, 2, \dots$. The GEM distribution of resulting stick lengths π_i is given by

$$\pi_1 = V_1, \quad \pi_k = V_k \prod_{j=1}^{k-1} (1 - V_j) \quad \text{for } k \geq 2 \quad . \quad (4.37)$$

This distribution over distributions (sequences of breakings) corresponds precisely to the one predicted by the de Finetti's Theorem in the case of the CRP, and gives an i.i.d. distribution from which customer-assignments are sampled.

The following scheme is computationally convenient: Sample uniformly distributed random values $x \in [0, 1]$ on the entire unit-length stick. Break the stick m times, with break points sampled from $\text{Beta}(1, \theta)$, until the cumulative length of broken stick parts $b_m = \sum_{k=1}^m \pi_k$ exceeds x . Then assign x to table m . If x is smaller than b_m , but $b_j < x \leq b_{j+1}$, then assign x to table j . This way samples from the GEM distribution are generated on demand, without a need to construct the complete GEM distribution (see Fig. 4.2 for a visualization).

The stick-breaking process might also immediately remind us of record statistics. Recall the random variable, x_r , Eq. 4.6, which models the retraction of the adjacent possible. Any parent distribution $p(y)$ can be transformed into a distribution $p(x)$, which is uniform on the interval $x \in [0, 1]$ and which will give rise to the same statistics of record times,

$$p(x) = \int_{-\infty}^x p(y) dy \quad , \quad x \in [0, 1] \quad , \quad (4.38)$$

	I.I.D Record Process	Chinese Restaurant Process
Probability of record / new table at t -th element	$p_r = 1/t$	$p_m = \theta/(t + \theta)$
Distribution of number m of records (tables) in k elements	$p(m k) = \frac{1}{k!} s(k, m)$	$p(m k) = \frac{\Gamma(\theta)}{\Gamma(\theta+k)} s(k, m) \theta^m$
Element of type j	Uniform random value x falling in between two records $x \in (x(t_{j-1}), x(t_j))$	Customer at table j / Uniform random value x falling in between two break points $x \in (b_{j-1}, b_j]$
Type count i in k elements	Number of inter-record values $\sum_{t=1}^k x_t$ with $x_t \in (x(t_{j-1}), x(t_j))$	Number of customers at table j
Probability for the histogram of type counts $\{a_i\}$	$p(a_1, a_2, \dots, a_k k) = \prod \frac{1}{i^{a_i} a_i!}$	Ewen's sampling formula: $p(\{a_i\} \theta, k) = \frac{k! \Gamma(\theta)}{\Gamma(\theta+k)} \prod_{i=1}^k \frac{\theta^{a_i}}{i^{a_i} a_i!}$

Table 4.1: Comparison of various measures on the Chinese restaurant process compared with the i.i.d. record process.

where $p(y)$ is the distribution of i.i.d. random variables from which records are sampled. For a sequence of record times, t_i , the corresponding values, $x(t_i) = x_r(t)$, will reproduce the distribution of stick-breaking positions, where $\theta = 1$. In this case the Beta distribution, $\text{Beta}(1, 1) = U(1)$, is a uniform distribution $U(1)$. This perspective also allows us to identify a correspondence between the number of random values x that fall in the interval $(x(t_i), x(t_{i+1}))$, so in between two records, and the histogram of type counts a_i is given by Ewen's sampling formula Eq. 4.33 (see Table 4.1 for an overview of corresponding quantities),

$$p(a_1, a_2, \dots, a_k | k) = \prod \frac{1}{i^{a_i} a_i!} \quad . \quad (4.39)$$

which we derive from Eq. 4.33 by setting $\theta = 1$ and where a_i denotes the histogram of in-between record counts. Importantly, this observation offers a new understanding of the CRP as being governed by a latent RP/SSRP. For $\theta \neq 1$ the record process can no longer be described by a stationary distribution, $p(x)$.

As a result on the side, we note a structural parallel: tracing the coalescence-tree of a population is a backward-in-time process, estimating when lineages merge into common ancestors, while the CRP is a forward-in-time process, generating new types. This duality echoes the mirror symmetry observed earlier between the SSRP and the RP, where one describes the shrinking of possibility forward in time, and the other describes it in reverse, suggesting that the SSRP might be a useful tool in studying coalescence.

In this subsection, we have introduced the Dirichlet Process, which can be intuitively understood with the CRP and the SBP. We also showed that the SBP with $\theta = 1$ can be identified as an i.i.d. RP and use this to identify Ewen's sampling formula as another property of the i.i.d. RP. In the next subsection, we will move our attention to evolution with selection and introduce a generalization of the Dirichlet Process.

4.3.2 Evolution with selection and the Pitman Yor Process

Evolution with constant selection

We have the interesting situation in which both neutral evolution – in particular the observation of new species in a sample of size k and directed evolution on the fully epistatic fitness landscape – are described by related mathematical structures, namely the CRP and the statistics of i.i.d. records.

However, evolution in random fitness landscapes is not very realistic, and fitness landscapes with large neutral plateaus and/or extended local maxima are more realistic [Gav04]. Therefore, we have to make our model of evolution based on records more flexible.

It is instructive to first assume a linear fitness landscape in which any beneficial mutation adds a constant selective advantage s . We assume that the population size stays constant and that such beneficial mutations arrive with rate μ . Whether a beneficial mutation is eventually fixed in the population is however determined by genetic drift and depends on the current population fitness f . Assuming a birth-death process [Gil04b], the probability of fixation of a beneficial allele with relative fitness, $r = (f + s)/f$, is given by

$$p_{fix} = \frac{1 - r^{-1}}{1 - r^{-N}} \quad , \quad (4.40)$$

which in the limit of a large population becomes (see also chapter 2)

$$p_{fix} = 1 - r^{-1} = \frac{s}{f + s} \quad . \quad (4.41)$$

If the current fitness, f , is assumed to be the result of n sequential beneficial mutations, $f = \sum_{i=1}^n s$, Eq. 4.41 reminds of the probability of a record,

$$p_{fix}(n) = \frac{1}{n + 1} \quad , \quad (4.42)$$

but instead of summing all mutations, be they fixed or not, it is as if time, here replaced by the number of fixed mutations n , is only counted if a record in fitness space is actually obtained. We can show that the probability of fixation decreases as $\sim (\mu t)^{-1/2}$. The probability of fixation in a population of individuals with n beneficial mutations, is given by Eq. 4.42. On average, the waiting time, $\langle \tau \rangle$, for the next beneficial mutation to fix, is therefore $\langle \tau \rangle = n + 1$. The cumulative waiting time t before n mutations are fixed, is therefore growing like $t \sim n^2$, and the number of mutations therefore grows as $n(t) \sim \sqrt{t}$. After differentiating $n(t)$ we see that the speed at which mutations are fixed and consequently the probability of fixation decays as $p_{fix}(t) \sim t^{-1/2}$ [KTP09].

The Pitman Yor Process

The described behavior of evolution on a linear fitness landscape is closely related to a common extension of the Dirichlet Process, dubbed the *Pitman-Yor process*, or two-parameter *Poisson-Dirichlet process* [PY97]. In the Pitman-Yor process (PYP), the stick-breaking scheme is adapted such that the stick is broken with probabilities

$$V_m \sim \text{Beta}(1 - \alpha, \theta + n\alpha) \quad , \quad (4.43)$$

where m is the current count of break points and $\alpha \in [0, 1]$. The GEM-like distribution of the resulting stick lengths π_i is given by

$$\pi_1 = V_1, \quad \pi_m = V_m \prod_{j=1}^{m-1} (1 - V_j) \quad \text{for } m \geq 2. \quad (4.44)$$

A stick-breaking process constructed in this way also offers a simple interpretation in terms of a CRP. It is a process in which there is an increasing probability to start a new table depending on how many tables are already present,

$$p_m(k) = \frac{\theta + \alpha n}{k + \theta}, \quad p_c(k) = \frac{k - \alpha n}{k + \theta}, \quad (4.45)$$

where n is the number of tables after k customers.

The PYP is intended to model innovations more explicitly, in particular the idea that an existing diversity of types (tables) is triggering further innovations. Secondly, it is well known that the PYP can reproduce power-law distributions more accurately, while remaining an exchangeable process, which is useful for many applications of nonparametric Bayesian methods that use the PYP to generate prior distributions.

For the purpose of this work, it is interesting to note that for fixed θ , asymptotically

$$p_m(k) \sim k^{\alpha-1}. \quad (4.46)$$

The probability of choosing a new table is decaying like a power law, with an exponent that depends on the parameter α .

We next construct an approximate recursive scheme for the probability, $p(n+1)$, that a new table is chosen given that n tables are currently occupied,

$$p(n+1) = \frac{\theta + n\alpha}{\theta + n\alpha + 1 - \alpha} p(n) \quad , \quad p(0) = 1 \quad , \quad (4.47)$$

where we used the fact that the shown fraction is the mean of the Beta($1 - \alpha, \theta + m\alpha$) distribution. Compare this to the law observed for the case of a constant fitness gradient, Eq. 4.42, for which we derive

$$p(n+1) = \frac{n}{n+1} p(n) \quad . \quad (4.48)$$

Verify that for $\alpha = 1/2$ and $\theta = 0$, the same average dynamics of the adjacent possible is observed.

We may also attempt to identify a record statistical process, which mimics the observed average dynamics Eqs. 4.47 & 4.46. In such a process, the distribution from which values are drawn, has to change upon the observation of a new record. Take the case of $\alpha = 1/2$ for example, and assume that random values are initially drawn uniformly from the interval $[0, s]$. If each observed record extends the sampling space by s , such that after m records each random variable is uniformly drawn from $[0, s + ms]$, the probability of a record will also decay as $p_m(t) \sim 1/\sqrt{t}$.

In this subsection we have introduced the PYP, an extension the the Dirichlet process, and argued how the PYP relates to the RP. The crucial observation is, that such a RP requires a form of feedback, where each record changes the sampling distribution. In the following, we will introduce yet a broader class of models, which are designed to capture the statistics of innovation.

4.3.3 Urn models of Innovation

We have seen that innovation can be studied using urn models and over the years several versions of the CRP were proposed. Of particular interest in this context is the model by Tria et al. [TLSS14], called the *urn model with triggering*, which was recently shown to feature an equivalence with the PYP for a certain parameter choice of the model [TLS18, TRLT24].

The urn model with triggering is an extension of the Hoppe urn. Instead of a colorless ball which triggers the addition of a new form, the observation of new colors itself triggers the addition of further new colors. Each newly observed color therefore contributes positively to the probability of finding further novelties.

This model has been used to study the statistics of novelty observed in various systems, such as literature corpora and online communities [TLSS14]. In correspondence with the PYP, the probability of observing a new type in these systems often decays as a power law,

$$p_{\text{novelty}}(t) \sim t^{-\beta}, \quad \beta \in (0, 1] \quad , \quad (4.49)$$

or equivalently (but more often reported), the number of distinct types grows as

$$D(t) \sim t^{1-\beta} \quad , \quad (4.50)$$

a relationship usually referred to as *Heaps' Law* [Hea78].

These models of innovation – the Chinese Restaurant Process (CRP), the Pitman-Yor Process (PYP), and the urn model with triggering – exhibit rich-get-richer dynamics, or *preferential attachment*, which is the reason for the emergence of power laws in the abundance of types. These distributions are of interest, as they are typically observed in many complex evolving systems. It is interesting to note that the outlined connection between the CRP, RPs, and SSRPs offers a way to understand these power laws from a new and complementary perspective.

4.3.4 Obsolescence and the Bak Sneppen model

While there exist strong similarities between models of evolution and models of innovations, there is also a crucial difference: Most urn models assume that the urn gradually fills, but is never emptied. Therefore, balls added very early in the process can still affect the dynamics later on in the process. This is crucially different in the context of evolution, where extinctions can reset the system and create space for new species. In the context of economics, this dynamic is referred to as *creative destruction*, a term coined by Joseph Schumpeter [Sch13]. It describes a situation in which an innovation triggers the obsolescence of existing technologies.

Due to its similarities to record statistics, we want to mention the *Bak-Sneppen model* here [BS93], another model aimed to understand evolution. In the Bak-Sneppen model, a population of N individuals is assigned random fitness values f_i for each individual i . At every time step,

the fitness of the least fit individual is replaced with a new fitness drawn from a distribution. In this way, the whole population becomes successively more optimized, effectively introducing a process close to a RP. The average of the whole population is adapting with a rate predicted by record statistics.

In this formulation, however, the mean fitness is – except for brief fluctuations in the fitness of individual members – increasing deterministically. In the original Bak-Sneppen model, individuals are arranged on a ring. Along with the least fit individual, also its two neighbors are replaced. This can trigger waves of fitness breakdowns on the circular topology, and importantly, it can effectively trigger a reset of the population fitness. This is a form of obsolescence.

In this section, we have progressively introduced the Dirichlet process, the PYP and the urn model with triggering, as models of evolution and innovation, which can be understood as extension of the baseline i.i.d. RP. We argued that a RP, which is ultimately able to reproduce the statistics of systems that innovate, has to incorporate a form of feedback: The distribution from which record-values are drawn is changing upon the observation of a record. Secondly, we argue that obsolescence plays a key role in such systems. In the next section we attempt to construct a model of innovation, inspired by the RP, which incorporates both such features.

4.4 Generalized Models of Innovation

In the previous sections we have outlined, that there are rich connections between models of innovation, evolution and the statistics of records. Further, we have argued that there are two key ingredients, that a generalized theory of innovation should have: (1) innovations restrict and expand the adjacent possible (2) innovations can trigger obsolescence.

	Probability of Novelty $p_r(t)$	Average Collapse of the Adjacent Possible
Record Process	$\frac{1}{t}$	$p_r(n+1) \approx \left(1 - \frac{1}{2}\right) p_r(n)$
Coupon Collector	$\sim \left(1 - \frac{1}{N}\right)^t$	$p_r(n+1) = \left(1 - \frac{1}{N-n+1}\right) p_r(n)$
Chinese Restaurant Process	$\frac{\theta}{\theta+t}$	$p_r(n+1) \approx \left(1 - \frac{1}{1+\theta}\right) p_r(n)$
Pitman Yor Process	$\sim \frac{1}{t^{1-\alpha}}$	$p_r(n+1) \approx \left(1 - \frac{1-\alpha}{1-\alpha+n\alpha+\theta}\right) p_r(n)$
Evolution on linear fitness landscape	$\sim \frac{1}{\sqrt{t}}$	$p_r(n+1) = \left(1 - \frac{1}{n}\right) p_r(n)$
Prime Numbers	$\sim \frac{1}{\log(t)}$	$p_r(n+1) \approx \left(1 - \frac{1}{t}\right) p_r(n)$

Table 4.2: This table compiles various example systems, in which the adjacent possible is collapsing both as a function of the number of novelties, n , and as a function of time, t .

4.4.1 Collapse of the Adjacent Possible

In its simplest form, the adjacent possible is characterized by the probability, p_r , that a novelty is obtained, be it a record, a mutation or a colored ball. In models of innovation and evolution,

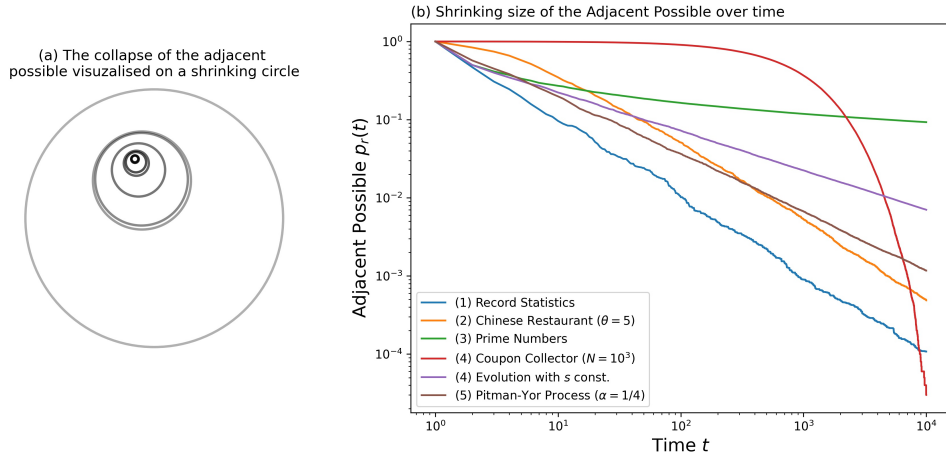


Figure 4.3: **Collapse of the adjacent possible for different processes:** **(a)** A sketch of how the size of the adjacent possible is changing over time: Random points within the unit circle are sampled at each timestep. Starting with an adjacent possible, which spans the whole unit circle, each observed point, that falls within the adjacent possible, is causing the adjacent possible to shrink. This creates a nested sequence of circles with decreasing area. **(b)** The size of the adjacent possible, $p_r(t)$, for different processes. In an i.i.d. record process, the adjacent possible shrinks exponentially with each novelty (record), yielding an effective marginal decrease in time of $p_r(t) = t^{-1}$ (blue line). If the retraction is slower than exponential in the number of innovations, other process classes are recovered, such as the Chinese restaurant process (brown) or the probabilistic dynamics of prime numbers (green). For further details see Table 4.2.

the restriction of the adjacent possible often depends on the total number of observed novelties n . See Table 4.2 and Fig. 4.3 for a summary of examples.

We can also conceive of more general forms for this reduction of the adjacent possible. A well-known example from mathematics is the coupon collector's problem, where the probability, $p_r(m)$, of obtaining a new coupon diminishes gradually with the number of found coupons, n . Early coupons are easy to collect, but the last few significantly shrink the adjacent possible,

$$p_{coupon}(n) = \frac{N - m}{N} = \left(1 - \frac{1}{N - n + 1}\right) \cdot p_{coupon}(n - 1) \quad , \quad (4.51)$$

where N is the total number of distinct coupons. Another illustrative example treats the emergence of prime numbers as an innovation process. As one counts from 2 to ∞ , through all integers t , the density of primes is successively diluted. Every obtained prime number restricts the density of primes at larger integers. This process can be studied from the perspective of probabilistic number theory and the adjacent possible is effectively decaying as

$$p_{prime}(t) = \prod_{p \text{ is prime}}^{t-1} \left(1 - \frac{1}{p}\right) \quad , \quad (4.52)$$

which can also be approximated using the recursive law,

$$p_{\text{prime}}(n) = \left(1 - \frac{1}{t}\right) p_{\text{prime}}(n-1) \quad , \quad (4.53)$$

where n is the approximate number of primes found at integer t . Using Euler's product formula one can show that the "probability" of encountering a new prime at integer t decays like,

$$p_{\text{prime}}(t) \sim \frac{1}{\log t} \quad , \quad (4.54)$$

which, for large t , yields the Prime Number Theorem [Apo13],

$$\pi(t) \sim \int_2^t \frac{1}{\log \tau} d\tau \quad , \quad (4.55)$$

where $\pi(t)$ is the so called prime counting function. The distribution of prime numbers features strong structural similarities to the Dirichlet process. A corresponding formulation of Ewen's distribution for prime numbers is given by *Billingsley's theorem* Patrick Billingsley. On the distribution of large prime divisors. *Periodica Mathematica Hungarica*, 2(1-4):283–289, 1972. In general, we may express the restriction of the adjacent possible on average through a recursive relation of the form,

$$p_r(n+1) = f(n, t) \cdot p_r(n) \quad , \quad (4.56)$$

allowing us to classify evolving systems accordingly. As seen in Table 4.2 this setup can explain the average dynamics of diverse systems and produces a broad range of rates for arriving novelties. The PYP, for instance, induces a reduction of the adjacent possible by Eq. 4.47, interpolating between models of evolution on a random fitness landscape and those on a constant gradient. To understand this behavior, we must recognize that each innovation not only restricts but also expands the space of possibilities. Nevertheless, the overall size of the adjacent possible still tends to decline in these models.

4.4.2 A simple record process with obsolescence

Corominas-Murtra et. al. developed a generalization of the SSRP, which they called *sample space reducing process with driving* Bernat Corominas-Murtra, Rudolf Hanel, and Stefan Thurner. Understanding scaling through history-dependent processes with collapsing sample space. *Proceedings of the National Academy of Sciences*, 112(17):5348–5353, 2015. Unlike the above described SSRP (also called the slowly driven regime), the SSRP with driving introduces a constant probability that the system is reset from the current state to a random state.

We take this concept of occasional sample space extension and apply it to the stochastic reduction of the adjacent possible in classical record processes over i.i.d. random variables. When a record is found, rather than reducing the adjacent possible, we allow it to reset to it's maximal size. In this context, the record memory – i.e., the maximum observed value – is erased and reinitialized. The previous records becomes obsolete. We call this the *record process with stochastic reset*. Concretely, in the record process with stochastic reset random

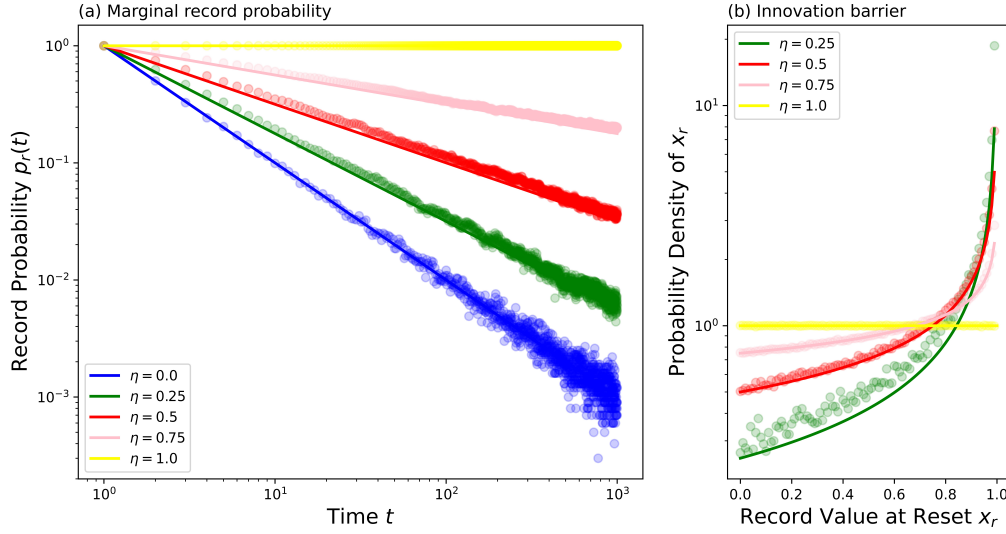


Figure 4.4: **Record process with stochastic reset:** In the record process with reset, each record can trigger a reset of the current record value with probability η . In panel **(a)** we show numerically, that for any value of $\eta \in [0, 1]$ the resulting marginal record probability decays as $t^{\eta-1}$ (colored points: simulation results, colored lines: $t^{\eta-1}$). While resets occur at any record with equal probability, the value of the record, $x_r \in [0, 1]$, at which a reset occurs, is skewed towards higher values of x_r . In **(b)** we show numerically that the probability that a reset happens at value x_r is following roughly a $\text{Beta}(\eta^{-1}, 1)$ distribution (colored points: simulation results, colored lines: $\text{Beta}(\eta^{-1}, 1)$).

values x are drawn at random from a distribution $p(x)$ and without loss of generality we can pose that $p(x)$ is uniform on the unit interval, $[0, 1]$. At each time t one random variable, x_t , is drawn and compared to the maximum value observed until then, x_r , in order to evaluate whether a record was obtained. If a new maximum is observed, $x_t > x_r$, we draw a Bernoulli random variable, ν , with success probability $\eta \in [0, 1]$. If $\nu = 1$, we reset the current maximum to zero, $x_r = 0$, effectively restarting the process, otherwise we update $x_r = x_t$. Dynamically equivalent to this is a formulation in which each reset serves as a full shift of the parent distribution $p(x)$, such that new values are then drawn from $p^*(x) = p(x - v)$, where $v > 1$. This too triggers the obsolescence of previous record values.

This simple extension leads to intriguing results. Most notably, the probability of observing a record at time t is increased by the occasional rejuvenation of the record process. The decay rate of t^{-1} becomes $t^{\eta-1}$, where η is the rate of stochastic resetting upon the occurrence of a record, which we demonstrate numerically (Fig. 4.4a). See the appendix for an analytic derivation of the observed behavior in the limit of large t .

We assume that resets are triggered by record events, motivated by the idea that some records could trigger creative destruction. In this model, innovations originate from within the system rather than being imposed externally. In contrast to external driving, where resets occur after an exponential waiting time regardless of record history, this internal mechanism allows the system to evolve autonomously.

Alternatively we can also perceive this model as an approximate description of evolution in correlated fitness landscapes (see Fig. 4.5). Assume a population with offspring distribution,

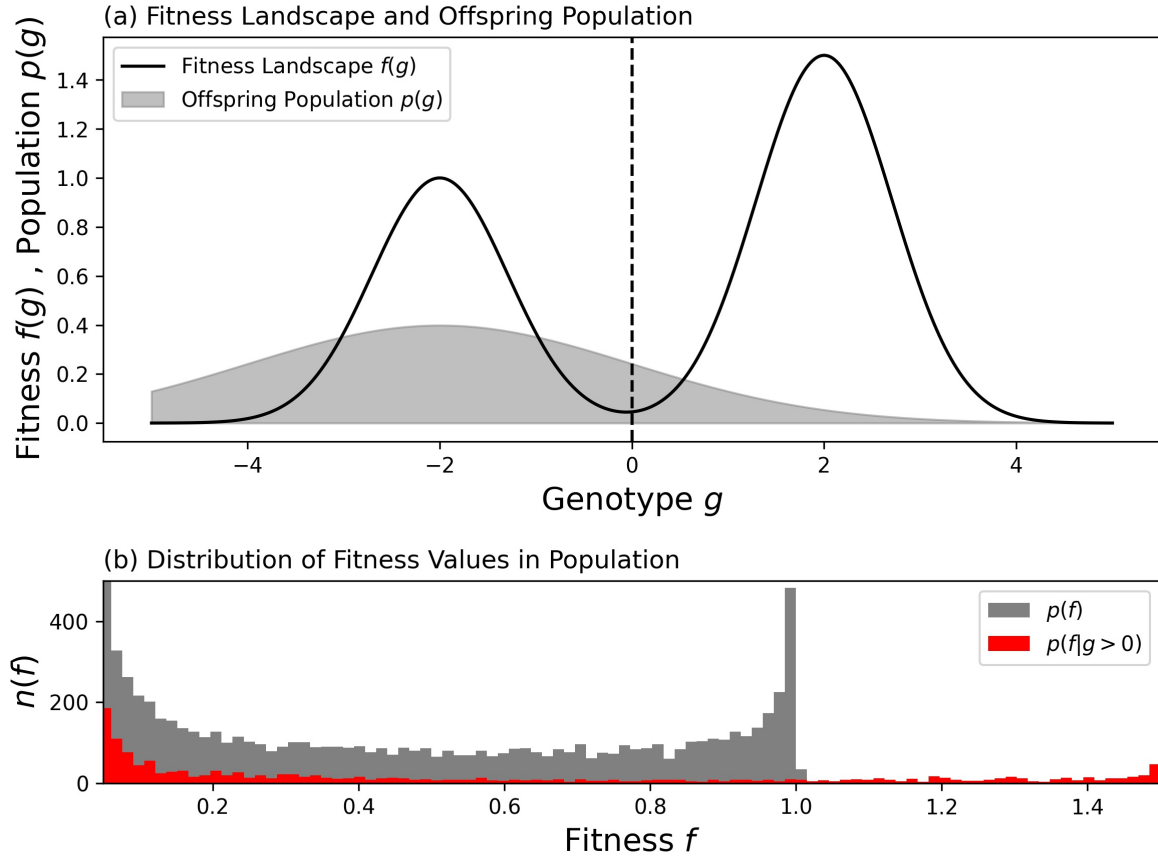


Figure 4.5: **Record process with reset as a model of evolution on rugged fitness landscapes:** **(a)** Shown is a fitness landscape, $f(g)$, (thick line) in a 1D genotype space. The grey filled curve, indicates the offspring distribution, $p(g)$ of genotypes, g , in a population residing on the left peak, $g < 0$. If $g^* = g > 0$ we say that, g^* , belongs to the more optimal fitness peak on the right. **(b)** We sample fitness values from the population and show their histogram in grey. We additionally color all fitness values contributed by a genotype with $g^* = g > 0$ in red. Above the innovation barrier of $f = 1$, there exists a small probability for an offspring fitness to be sampled exclusively from the more optimal fitness peak. If this happens, this offspring individual, g^* , will be selected and its offspring distribution, $p^*(g)$, shifts to the peak at $g = 2$.

$p(g)$, over genotypes, g . Each genotype is assigned a fitness value $f(g)$, determined by the fitness landscape. We assume that selection is strong and that the parent population at time t is described by a much more narrow distribution, around the fittest individual of the offspring at time $t - 1$, with genotype g_r . The genotype g_r will have approximately the highest fitness observed until time t , $f(g_r)$, and can therefore be called a genotype with record-fitness. Assume that $p(g)$ is supported in the vicinity of a local maximum in the fitness landscape, such that offspring has a fitness drawn from the local fitness peak (peak at $g = -2$ in Fig. 4.5). In this setup the population fitness, $g_r(t)$, will exclusively evolve towards this local maximum and will approximately follow a record process.

However, if a neighboring – more optimal – peak can be reached by a small fraction of the offspring population, the process is better described by a record process with reset. If an individual with genotype, $g_r^*(t)$, belonging to this fraction (the more optimal peak at $g = 2$ in Fig. 4.5) attains a fitness higher than the rest of the population, this will not only constitute a fitness record, $f(g_r^*)$, but also trigger a shift in the offspring distribution $p^*(g)$ as the new population average after selection, $g_r^*(t)$, has digressed far in genotype space: $|g_r^*(t) - g_r(t - 1)|$ is relatively large. In this notation the star, $*$, denotes that a genotype belongs to the more optimal fitness peak or that an offspring population originates from such a genotype. After a few generations, the distribution $p^*(g)$ will be mostly supported on the new fitness peak. The reset in the record process is identified with the shift in the offspring distribution as a “special” record genotype, g_r^* , is obtained.

This view implies the presence of an apriori unknown *innovation barrier* b , defined as the height of the lower fitness peak. Records exceeding this barrier can be considered part of a more optimal fitness peak (Fig. 4.5b). We can measure the distribution, $p(b)$, of such an innovation barrier for the RP with stochastic reset (Fig. 4.4b), by recording the value, $b \approx x_r^*$, at which the record process experiences a reset.

Previous work [CMHZZ18] has also examined state dependent driving rates in SSRP. This idea can be naturally extended to describe the collapse and expansion of the adjacent possible. The possible record-value dependent reset probability, $\eta(x_r)$, takes the role of the state dependent driving rate, $\rho(x)$ in the SSRP.

4.5 Applications and Outlook

The methods outlined above may prove useful for studying systems of innovation in which a notion of fitness can be defined. One clear application is found in sports, where performance metrics provide a natural fitness measure. We have also discussed evolution in considerable depth. In fact, both macroevolutionary data [SJ13] and results from experimental evolution [ECL96] suggest that power laws govern phenomena such as the distribution of mass extinction events and the growth of *E. coli* fitness over time. In fact the *distribution of fitness effects* can be considered an analogous concept to the adjacent possible defined here.

Beyond these examples, several additional domains present themselves. In the context of learning, *experience curve effects* are often found to follow the so called *power law of practice* [Sno26]. Similar trends are observed in economics, where learning curves describe efficiency improvements. Wright’s law states that the price of a good tends to decline as a power law with cumulative production [NFBT13]. In numerical optimization, many search algorithms exhibit power-law-distributed waiting times between improvements, suggesting that these methods could also benefit from a record-statistics perspective [BP02].

As we have seen, even systems without an explicit notion of fitness – such as the emergence of words in a text or the distribution of prime numbers – can be fruitfully analyzed using tools from record statistics. In the next chapter, we will apply these concepts to systems of sports and athletics.

Appendix

Analytical argument for the observed numerical behavior in Fig. 4.4

Recall from the main text that in the record process with stochastic reset random values x are drawn at random from a distribution $p(x)$ and without loss of generality we can pose that $p(x)$ is uniform on the unit interval, $[0, 1]$. At each time t one random variable, x_t , is drawn and compared to the maximum value observed until then, x_r , in order to evaluate whether a record was obtained. If a new maximum is observed, $x_t > x_r$, we draw a Bernoulli random variable, ν , with success probability $\eta \in [0, 1]$. If $\nu = 1$, we reset the current maximum to zero, $x_r = 0$, effectively restarting the process, otherwise we update $x_r = x_t$.

We found numerically, that the probability of observing a record at time t is increased by the occasional rejuvenation of the record process. The decay rate of t^{-1} becomes $t^{\eta-1}$, where η is the rate of stochastic resetting (Fig. 4.4a). Here we show the behavior of this record probability in the RP with reset in the limit of large t . We show that

$$p_r(t) \sim t^{\eta-1} \quad \text{for } t \rightarrow \infty. \quad (4.57)$$

First, we consider the underlying SSRP with constant driving $\lambda(x) = \eta$, for which the state visiting probability, $p(x)$, is known [CMHT15],

$$p(x) \propto x^{\eta-1}. \quad (4.58)$$

The size of the sample space $x \in [1, N]$. In the limit of $N \rightarrow \infty$, the process becomes a continuous sample space reducing process and can be cast on the interval $(0, 1]$ by a transformation $x \rightarrow x/N$. This rescaled variable corresponds to the size of the adjacent possible, $1 - x_r$, in the record process. Given x_r , the probability to find a record is given by the same expression, $p_r = 1 - x_r$, and the average waiting for the next record is $\tau(x_r) = 1/p_r(x_r)$, which grows as x_r approaches 1. In the long term limit the distribution of $1 - x_r$ **at record times** is given by Eq. 4.58. The long term distribution $\psi(\tau)$ of waiting times is therefore a mixture of exponential waiting times,

$$\psi(\tau) \approx \int_{1/N}^1 p(x) x e^{-x\tau} dx \quad N \rightarrow \infty. \quad (4.59)$$

where $p(x)$ is a rescaled and normalized version of Eq. 4.58, $p(x) = Ax^{\eta-1}$. Let $r(t)$ denote the long term probability density to observe a record at time t . We can set up a renewal equation, which reads

$$r(t) = \psi(t) + \int_0^t \psi(t - \tau) r(\tau) d\tau = \psi(t) + (\psi * r)(t). \quad (4.60)$$

Taking Laplace transforms ($\tilde{f}(s) = \int_0^\infty e^{-st} f(t) dt$) converts the time-domain convolution into multiplication,

$$\tilde{r}(s) = \frac{\tilde{\psi}(s)}{1 - \tilde{\psi}(s)}. \quad (4.61)$$

Equivalently, expanding r as the sum over n -fold convolutions of ψ gives

$$r(t) = \sum_{n=1}^{\infty} \psi^{*n}(t) \iff \tilde{r}(s) = \sum_{n=1}^{\infty} (\tilde{\psi}(s))^n = \frac{\tilde{\psi}(s)}{1 - \tilde{\psi}(s)}, \quad (4.62)$$

i.e. the “ n -fold convolution” statement in Laplace space.

We compute the Laplace transform of $\psi(\tau)$ and by a change of order of integrations we obtain

$$\tilde{\psi}(\tau) \approx \int_{1/N}^1 p(x) \frac{x}{s+x} dx \quad N \rightarrow \infty \quad . \quad (4.63)$$

We rewrite this as

$$1 - \tilde{\psi}(\tau) \approx \int_{1/N}^1 p(x) dx - \int_{1/N}^1 p(x) \frac{x}{s+x} dx = \int_{1/N}^1 p(x) \frac{s}{s+x} dx \quad N \rightarrow \infty \quad . \quad (4.64)$$

The small- s (Laplace) limit corresponds to the large- t (time) limit by the Final Value Theorem and the Tauberian Theorem, hence it is enough to evaluate the behavior of $\tilde{\psi}$ for small s to study the asymptotics of $r(t)$. We perform a change of variable $x = su$, $dx = sdu$ and

$$\int_{1/N}^1 A x^{\eta-1} \frac{s}{s+x} dx = A s^{\eta} \int_{1/(Ns)}^{1/s} \frac{u^{\eta-1}}{1+u} du \quad . \quad (4.65)$$

In the large N limit and for any fixed small s , $Ns \rightarrow \infty$, the bounds of the integral become $(0, \infty)$ and independent of s ,

$$1 - \tilde{\psi}(\tau) \approx A s^{\eta} \int_0^{\infty} \frac{u^{\eta-1}}{1+u} du = \frac{A\pi}{\sin(\pi\eta)} s^{\eta} = B s^{\eta} \quad (4.66)$$

Plugging the small- s expansion into the renewal solution,

$$\tilde{r}(s) = \frac{1 - B s^{\eta}}{B s^{\eta}} \sim \frac{1}{B} s^{-\eta} \quad \text{for } s \rightarrow 0 \quad . \quad (4.67)$$

Using the Laplace pair $\mathcal{L}\{t^{\beta-1}/\Gamma(\beta)\}(s) = s^{-\beta}$ for $\beta > 0$, we invert to get the long-time behavior

$$r(t) \sim \frac{1}{B \Gamma(\eta)} t^{\eta-1} \sim t^{\eta-1} \quad \text{as } t \rightarrow \infty \quad . \quad (4.68)$$

Sport as a Model of Cultural Evolution

Abstract Athletic performance offers a uniquely quantifiable setting for studying cultural evolution, linking record statistics and evolutionary theory. Analyzing Olympic records and endurance sports, we find broadly universal patterns: linear progress in record values and counts across much of the 20th century. Population growth explains part of this trend but cannot alone account for sport evolution. Instead, progress reflects both expanding participation and evolutionary dynamics of innovation, inheritance, and selection. Applying concepts from population genetics, we find that variance predicts response, consistent with a heritability-like signal in performance. At the level of extremes, world records occur in clustered bursts and we propose a record process with restart reproducing this burstiness.

5.1 Introduction

Natural evolution offers one of the most powerful explanatory frameworks in science. Through variation, inheritance, and selection, populations of organisms adapt to their environments [Dar59]. The success of evolutionary theory has motivated researchers to ask whether the same principles might extend more broadly, beyond biology, to the cultural and technological domains [BR85, CKF17, Mes16]. Here, too, novelty arises, traits are transmitted by copying, and competition determines which variants persist.

The challenge in extending evolutionary reasoning to culture lies in defining what it means for a trait to be “fit”. In biology, fitness is anchored in reproductive success; in culture, different fields have adopted different proxies. In linguistics, fitness may be tied to frequency of use [CMFS11]; in economics, to profitability [Alc50]; in engineering, to efficiency or durability [NFBT13]; in art, perhaps to influence or longevity [Sim91]. Not always, but sometimes successfully, models of cultural evolution can even offer a notion of selection for fitness that parallels natural systems.

A central prediction of evolution under directional selection is that the mean fitness of a population tends to increase. Yet the formalism of evolutionary theory links not only to mean values but also to higher statistical moments such as variance, which governs the response to selection [FM83]. Moreover, in many empirical settings, the most striking signals of change are not averages but records – the rare, extreme achievements that push the boundaries of what is possible. Unlike in natural systems, where records often remain unnoticed at first, cultural domains frequently maintain detailed archives of record progressions: fastest machines, tallest

buildings, cheapest energy storage, farthest thrown javelin, and many more. The connection between those well-known classic results in population genetics on mean/variance on one hand and the statistics of extremes has been previously explored in the literature [Orr02, Gil84]. In this work we are particularly interested in connections between evolution and the statistics of records [KJ05].

In the previous chapter we introduced the statistics of records along with some extensions and related models for innovation. In this chapter we apply these methods using sports as our primary example. Athletic competitions, and the Olympic Games in particular, provide clear measures of fitness in the form of performance outcomes. Record progressions in sports allow us to study cultural evolution with a degree of quantitative precision that is rarely possible elsewhere.

In this chapter, we start by observing a general pattern in the cumulative statistics of athletic performance over various sport events, namely periods of unbound progression. We offer two possible explanations for this and provide evidence for the plausibility of both explanations. Related observations to the phenomenon of unbound progression can be made on the level of record times alone. We offer an evolutionary record process model, capable of reproducing distinctive properties of real-world record data.

5.2 Results

Athletic performance across most Olympic sports has increased markedly over the last century and does so in a broadly universal manner across events. Strikingly, the trajectory is approximately linear up to the 1980s, after which performance gains slow. This pattern holds not only for the record value, $y_r(t)$, but also the cumulative number of records, $n(t)$ (Fig. 5.1). Consider a naive model in which every year a “world-best” performance, y_t , is drawn from the same continuous stationary distribution, $p(y)$. While transient linear growth in the record value, $y_r(t)$, can be achieved for certain choices of $p(y)$, $n(t) \sim \log(t)$, a logarithmic growth in the observed number of records would be expected for any such i.i.d. record process, independent of the choice $p(y)$ [Gli78]. We therefore consider two (non-exclusive) explanations that are both compatible with the near-linearity, breaking the assumption of temporal independence of $p(y)$:

(1) Expanding participation. If access to sport – through equipment, coaching, and recruitment – grows exponentially (or logistically within its exponential phase), then record-breaking occurs at an approximately constant rate (see methods). Moreover, finalists can be viewed as the top m order statistics drawn from an underlying performance distribution. When the sample size $N(t)$ grows exponentially and the base distribution features an exponential tail for larger performance values, the expected trends of the top m extremes become approximately linear (Fig. 5.2; see methods). Distributions with exponential tails, could be a good model for how performance is distributed, since eventual results in sport events can be thought of as the combined outcome of several small additive factors.

(2) Cultural-evolutionary dynamics. Sports may be driven by an evolutionary process in which effective methods are discovered, copied, and refined. Here, “survival of the fittest” is replaced by *Citius, Altius, Fortius* (Faster, Higher, Stronger – the motto of the Olympic games). Practice, innovation, and learning accumulate. At the individual level, performance results from multiple latent, “genotype-like” contributors, such as technical proficiency, training regimes, nutrition, and athlete body morphology, which are differentially selected. Population

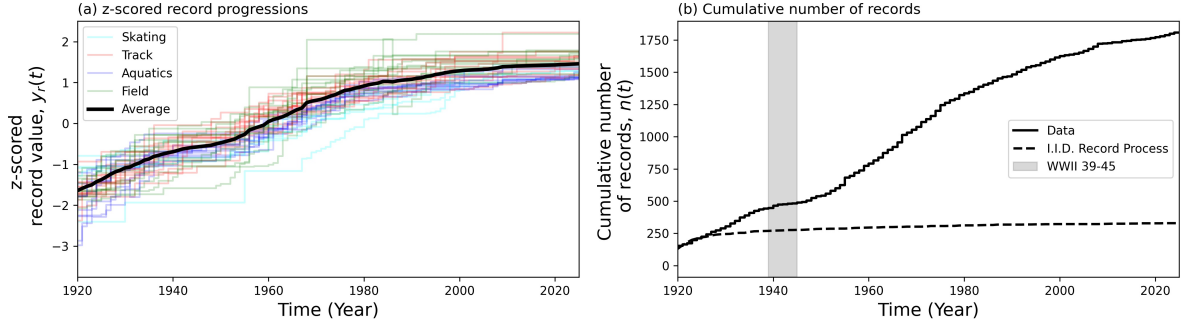


Figure 5.1: Average progression of performance across multiple sports, measured as the z-scored record value **(a)** and the cumulative number of records **(b)** across all sport events in the dataset (see Methods). For **(a)**, z-scores were computed as $z(t) = (y_r(t) - \bar{y}_r) / \sigma(y_r)$ with $\bar{y}_r = \sum_i y_r(t_i)$ and $\sigma(y_r)^2 = \sum_i (y_r(t_i) - \bar{y}_r)^2$; for events where lower values correspond to better performance (time controlled events), we used $-y_r$ in the calculation. With this standard z-scoring procedure the data collapses naturally. Progression curves are color-coded by sport class, and the average record curve across all sports is also shown (black thick). Only record trajectories beginning before 1920 are included. Overall, between 1920 and 1980 record values increased approximately linearly, with a marked interruption due to World War II (grey vertical area). Since 1980, progress has stagnated in most sports. **(b)** The cumulative number of records features a similar linear growth regime throughout much of the 20th century (black line). Assuming our data was produced by an i.i.d. record process where each year there are $m = 100$ possible occasions for record breaking, the observed number of records would grow *much* more slowly (black dashed line), which is in stark contrast to the observation.

genetics models can be devised to explain distributions of individual fitness with roughly constant variance and a directional drift toward higher fitness (performance) ([Ewe04, DF07]; Fig. 5.3).

Empirical evidence for both mechanisms. The transition from linear to sublinear growth in both the total number of records and the value of the record value – as seen in Fig. 5.1 – is suggestive of a logistic growth model, Fig. 5.2, where the latent sample size growth is related to the size of the world population [Yan75, EK09]. The world population has transitioned to a sub-exponential growth by 1980 [Uni22]. Yet, not every human can be assumed to have equal access to recruiting on the elite sport level and the accessibility of athletic careers might not have grown proportional to the overall world population. Unfortunately, performance on the elite level as in the Olympic games does not contain information on the underlying sampling distribution. To isolate participation effects, we analyze a particular data set for which complete results over multiple decades are available: the Comrades Cup Super-Marathon in South Africa. Both record values and the exact number of participants are available annually. For this extreme-distance event, growth in $N(t)$ explains most of the performance of the observed elite population (top m individuals) over the years (Fig. 5.4).

Evidence also supports evolutionary dynamics. Breeder’s equation in quantitative genetics predicts a positive correlation between within-population variance and the rate of evolutionary response (see methods) and serves as a measure of the heritability of traits. Heritability h^2 is defined as the proportion of variation in a population trait that can be attributed to inherited genetic factors and enters the Breeder’s equation as the linear dependence between response R_t , e.g. how much a trait improved on average after selection, and selection potential

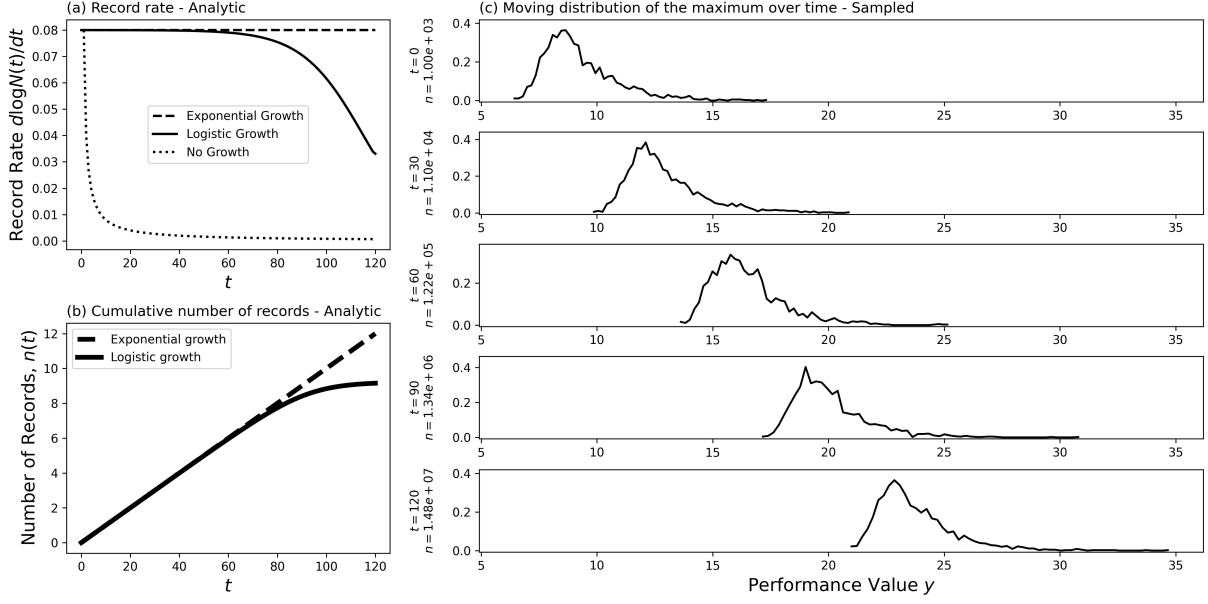


Figure 5.2: Extremal and record statistics in growing populations or sample sizes. Panel (a) shows three record rates, which are good estimates for the probability of a record at time t (see methods): The dashed line is the constant record rate, for exponentially growing sample sizes, $N(t) = N_0 e^{rt}$, with $r = 0.08$. The full line is for logistically growing sample sizes, which is constant at first but decays. Finally, dotted is the rapidly decaying record rate for sample sizes which do not grow. In panel (b) the expected cumulative number of records at time t is shown, which is obtained by integrating the record rates. Exponentially or logistically growing populations feature a linear growth in the number of records as observed in Fig. 5.1. In (a,b) the time axis (0-120) and the growth parameters were chosen to invoke a comparison with observed true record traces (Fig. 5.1) but are artificial. Panel (c) shows the distribution of the maximum value obtained from an i.i.d. exponential distribution, where the sample size grows exponentially. The distribution does not change its shape but is shifting towards larger performance values, y , at constant speed r .

S_t , e.g. the difference in trait between the average population and the selected population, $R_t = h^2 S_t$. We borrow this concept to study quantify the heritability of cultural traits, where we identify S_t with the yearly variance in performance, σ_t^2 , and R_t as the average increase in performance. Our minimal model satisfies a relationship between variance and evolutionary response (Fig. 5.3b), and, empirically, the correlation between within-Olympiad variance among finalists and the inter-Olympiad (four-year) improvement is predominantly positive across events, indicating an inheritance-like signal (Fig. 5.5). Furthermore, many sports exhibit well-documented phenotypic shifts [Eps14, NO01]. Across events, the stronger the contemporaneous phenotypic difference between medalists and the finalist field, the larger the long-run change in that phenotype over the Olympic era. This result parallels Breeder's equation at the level of morphology (Fig. 5.6), where we identify S_t with the phenotypic difference between medalists and other finalists and R_t with the overall phenotypic change.

Implications. A comprehensive theory of progress in sport must combine (i) changing population size with (ii) evolutionary improvement. Record curves frequently display piecewise behavior, echoing the evolutionary concept of punctuated equilibria: periods of relative stasis alternate with bursts of record-breaking, often triggered by innovations in technique [BTT⁺08]. High jump is the classic example: the introduction of a new jumping techniques, such as the

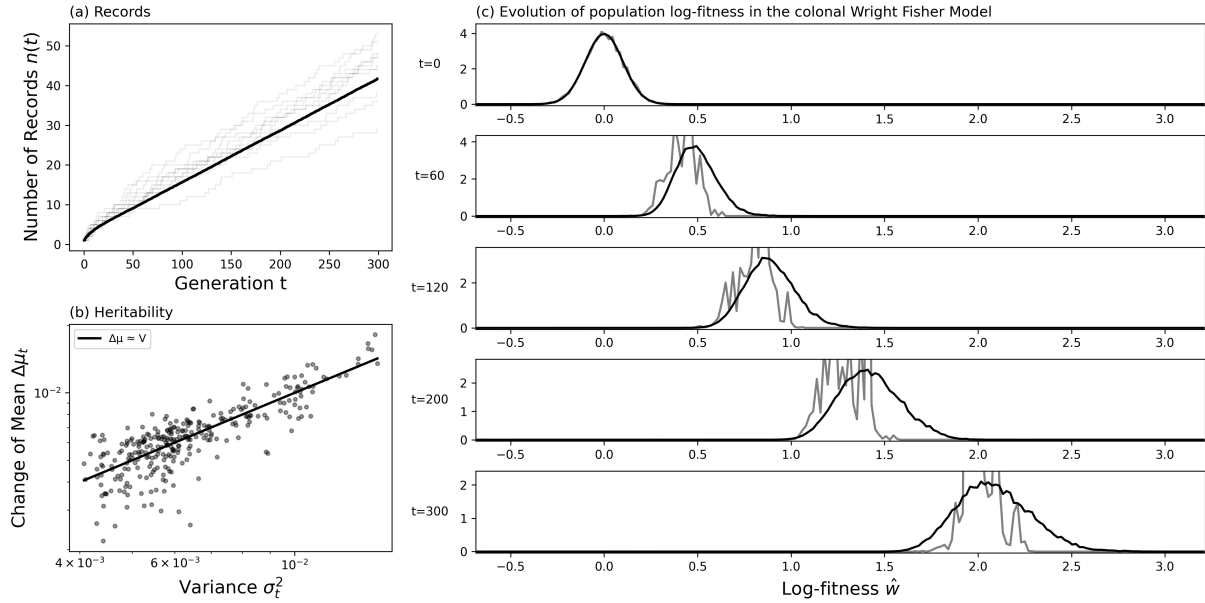


Figure 5.3: Behavior of a Wright-Fisher model with selection and mutation in log-fitness space (see methods for computational details). **(c)** The evolutionary dynamics shift an initially normally distributed population in log-fitness space at time $t = 0$ to higher fitness values. The population size was $N = 5000$. For a single evolving population the form of the distribution changes, while retaining a similar variance on average (grey), while the average of 1000 such populations displays a widening of variance (black) (this is similar to a biased random walk [MSW12]). In panel **(a)** this linear trend in fitness space results in a linearly growing number of records. Here the average behavior is shown (thick black line) together with some example trajectories for the cumulative record count (light black lines). In panel **(b)** the correlation between variance and change for mean from one generation to the next is shown. This relation is referred to as *Breeder's equation* and is an indicator for evolutionary dynamics.

Fosbury flop, precipitated a cascade of records, Fig. 5.5a. Such innovations transiently inflate performance variance beyond the predictions of simple evolutionary models. Consequently, predictive models of progress in sport must explicitly account for innovation events alongside participation growth and selection.

Record Data. In the following we will focus on *record statistics*, a domain with a robust mathematical foundation and striking universality properties. For independent and identically distributed (i.i.d.) draws from any continuous base distribution, the *occurrence times* of (upper) records are distribution-free: the probability that the t -th observation is a new record equals $1/t$, and the record indicators are independent ([Gli78], methods and Fig. 5.1). This universality makes record-process models natural building blocks on the path toward a general theory of the cultural evolution of sport. A second reason to focus on records is their ready availability. While historical sport data on the full field of participants in a sport competition is available in some cases (as we have seen for the example of the Olympics and the Comrades Cup), comparative studies – possibly to domains outside of sports – can only be reliably performed for record values.

As stated above, we find that the marginal probability to see a record is roughly constant in time if averaged over all records (Fig. 5.1) in accordance with the hypothesis of exponential sample size growth. However, records are not binomially distributed in time, rather several sports

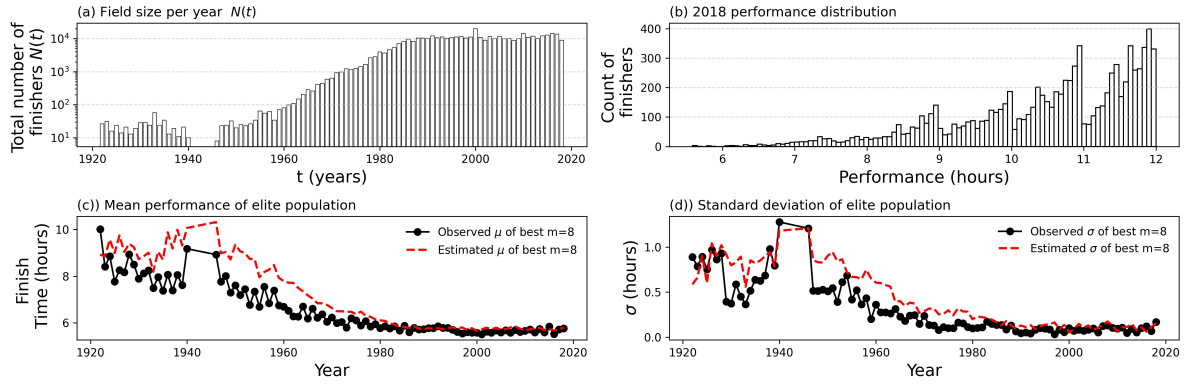


Figure 5.4: Data from the *Comrades Cup*, one of the oldest and largest ultra-marathons held yearly since the early 19th century. It connects the cities Durban and Pietermaritzburg in South Africa, but the direction of the race alternates. In **(a)** we show the total number of finishers in the dataset. The number of participants has grown over the years, with a very clear exponential regime (vertical axis is logarithmic) between 1950 and 1990. After that the maximal number of participants was limited. In 1946 the recorded number of runners was lowest $N = 8$ participants. In **(b)** we show the performance distribution of all finishers for the most recent year in the dataset, 2018. The performance distribution decays strongly towards faster times and features peculiar oscillations solely triggered by a human obsession with whole numbers. In **(c,d)** we show the mean and standard deviation of the elites in the dataset, namely the top m athletes of every year (black line with dots). It features yearly oscillations due to the alternation of race direction, one direction being slightly more challenging. We compare this with an estimate of a completely naive non-parametric model: For every year t we sample performances from **(b)** (which is assumed to be the “true” distribution of performance) $N(t)$ times and take the m best results. Generally $m \leq \min_t N(t)$, and we choose $m = 8$ to get the largest sample that can be compared to the data. The mean and standard deviation of this synthetic elite population is shown in red. Given the simplicity of the model, the data’s general shape of both mean and standard deviation is well reproduced. The model is generally overestimating the finish times of historic runners, which might be due to an increasing fraction of amateurs in the population of athletes over the years.

feature record curve signatures which show correlations. An important statistical signature here is the distribution of the waiting time between records, which follows a power law with power close to -2 and cutoff at around 10 years (see Fig. 5.7). This observation can not be explained by a simple drift model (as in Fig. 5.1&5.3) where the waiting time distribution decays exponentially. Further, the distribution of records in time is clustered with a positive burstiness value, of about 0.2 (Fig. 5.7). Burstiness is defined as $B = (\sigma - \mu)/(\mu + \sigma)$ where μ and σ are the mean and the standard deviation of the waiting time distribution [GB08]. For bursty, clumped data traces $1 > B > 0$, Poissonian uncorrelated traces $B = 0$ and for regular periodic traces $-1 < B < 0$.

Record Processes. The observed clustering in the number of records implies punctuated phases of progression and is a complimentary result to the previously discovered correlation between performance variance and performance growth. In contrast to the simple Wright Fisher model, which we introduced in Fig. 5.3, higher performance states are not equally likely to be reached at all times and depend on the availability of innovations. Borrowing further concepts from evolutionary biology, we may pose that evolution in some sport disciplines is situated on a

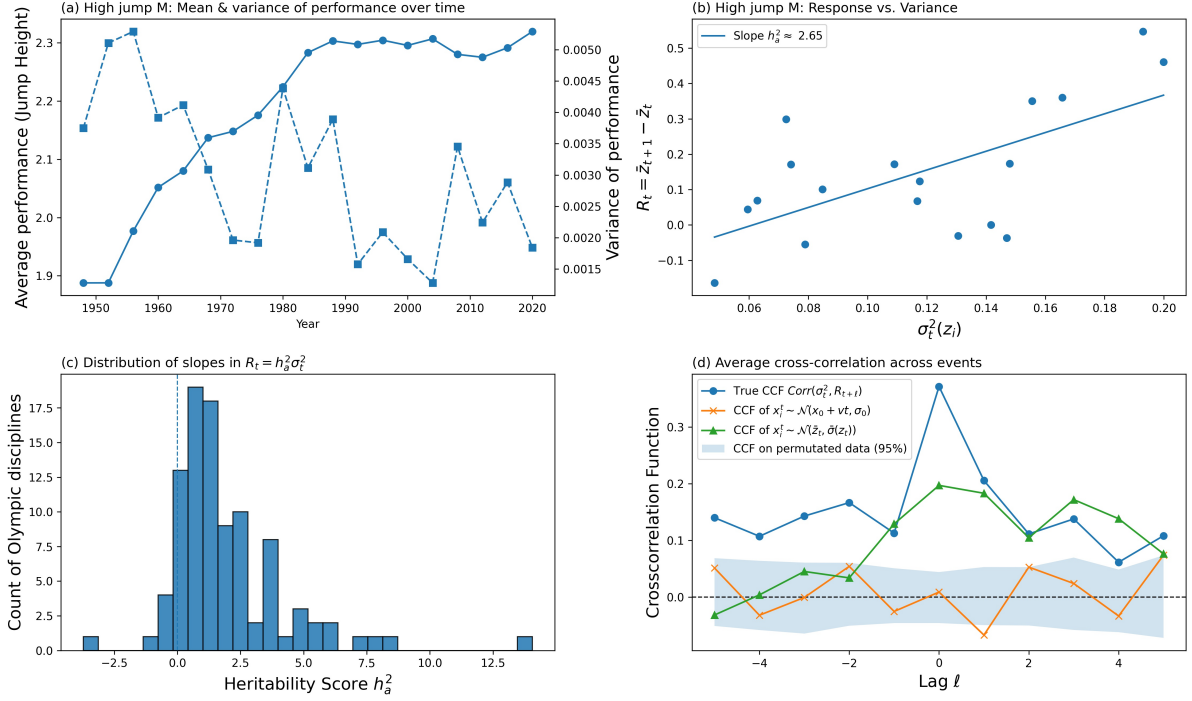


Figure 5.5: Data from the Olympics on the evolution of performance since 1960. We calculate mean and variance on a yearly basis for the finalists of different Olympic disciplines with quantifiable performance (these include Track & Field, Aquatics, Cycling, Canoe and Rowing). The number of finalists varies across sports, from around 6-8 in 100 m Sprint to more than 100 participants in Marathon, where the only conducted event is the final. **(a)** For several sports, and exemplarily shown for the prominent case of Men's High Jump, the average performance has increased in a piece-wise fashion. In this example, another observation can be made: The variance among athletes competing for High Jump in a particular year is high in phases of rapid overall progression. This mirrors an often observed dynamic in evolutionary biology, where the population variance is proportional to the speed of evolution. In **(b)** we show an analogous result, namely that the correlation between variance at time t , $\sigma_t^2(z_i)$, and response, $R_t = \bar{z}_{t+1} - \bar{z}_t$ is positive. z_i is the z-scored performance results $z_i = (y_i - \sum_t \bar{y}_t) / \sigma(\bar{y}_t)$. We calculate the slope of this relation and refer to it as the *heritability score*, h_a , building on a concept we borrowed from evolutionary biology, where heritability quantifies how much of trait variance can be inherited. In panel **(c)** we show the distribution of the measured heritability score for all disciplines in the dataset which were present for at least ten Olympiads. It is generally positive and most heritability scores are in between 0 and 2.5. . Finally, in panel **(d)** we present the cross correlation function at lag, ℓ , $\text{Corr}(R_{t+\ell}, \sigma_t^2)$, for three different processes: (1) a synthetic drift model, where athlete performances are sampled every four years from a linearly shifting distribution, $x_i^t \sim \mathcal{N}(x_0 + vt, \sigma_0)$, (orange line), (2) a more complex synthetic model, where performances are sampled from a normal distribution of z-scores guided by the average mean $\bar{z}$ and average variance $\bar{\sigma}(z)$ of the z-score in the population per year (green line), (3) finally the cross-correlation of the data itself (blue line). As a base line comparison we also show the 95% confidence interval into which the CCF of permuted data would fall, where σ_t^2 and R_t are scrambled up (blue shaded area). The linear model (1) does not create any correlations. Model (2) creates correlations which may be due to physical constraints, as both the progression and the variance of performances diminishes close to the theoretical optimum. The real data shows a pronounced peak at $\ell = 0$, which we interpret as an evolutionary effect.

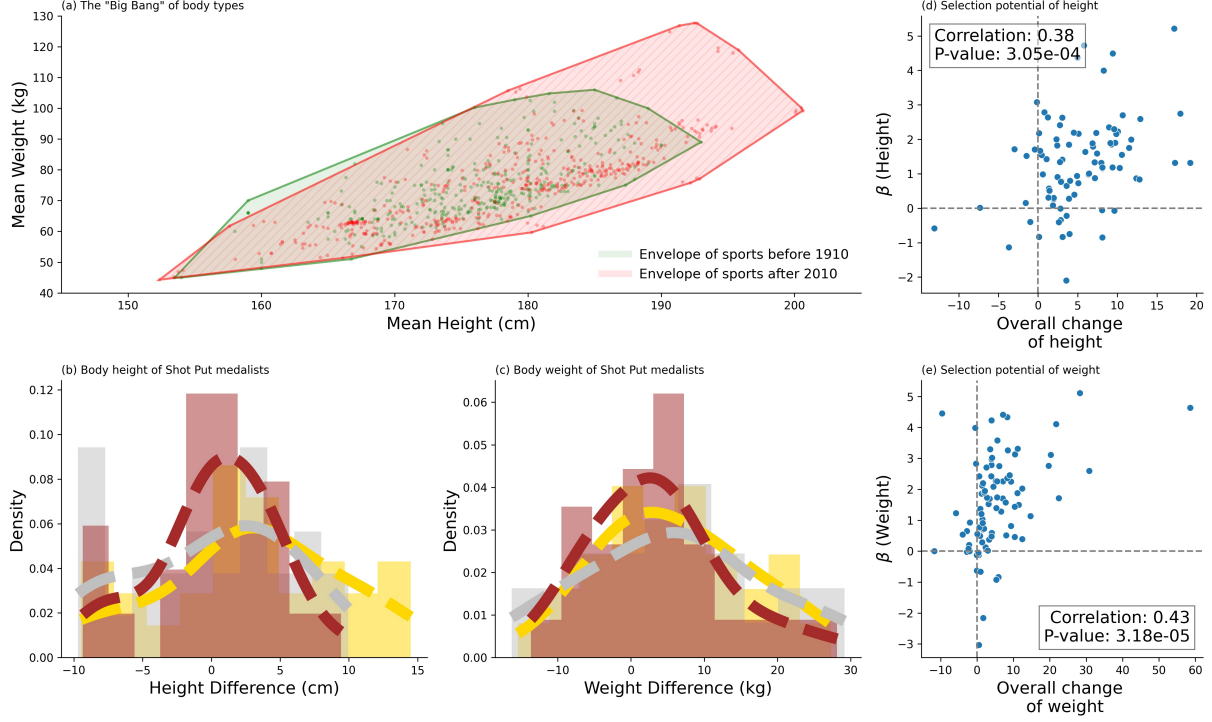


Figure 5.6: Data from the Olympics on the phenotypic variables height, h , and mass, m , of athletes. Panel **(a)**, the phenotype distribution has changed for Olympic events. Over the last century the overall distribution in the mass-height phasespace has expanded (from green dots (average mass of athletes in a sport before 1910) to red dots (after 2010)). This observation is also dubbed the “Big Bang of body types” [NO01], as athlete populations “explore” the space of possible body types to find the most adapted for their respective sport. We compute the convex hull in phasespace of all sports in the dataset before 1910 (green) and after 2010 (red) to demonstrate the expansion. In **(b,c)** we focus on the Shot Put, which has experienced a particularly strong increase in average height and mass of athletes. We see that the height difference, $h_m - \sum_i^n h_i/n$, (horizontal axis) between the height of the medal winners, h_m , (the densities, bronze, silver and gold are colored accordingly) and the average Shot Put athlete at the Olympics, h_i is positively skewed. Shot Put medalists are, averaged over all Olympic events, taller and heavier than the typical Shot Put athlete: Medalists are on average 5 kg heavier and 1 cm taller. For gold medalists the average difference is largest with 5.9 kg and 2.9 cm. If we look at the sample of all athletic events (not just Shot Put) **(d,e)**, we broadly observe positive correlations between the average phenotypic difference or “selection strength”, $\beta = \sum_t \left(\sum_1^3 h_m/3 - \sum_1^n h_i/n \right)$, and the overall change in phenotype over the Olympic period, $\sum_i^n h_i^{t>2010}/n - \sum_i^n h_i^{t<1910}/n$.

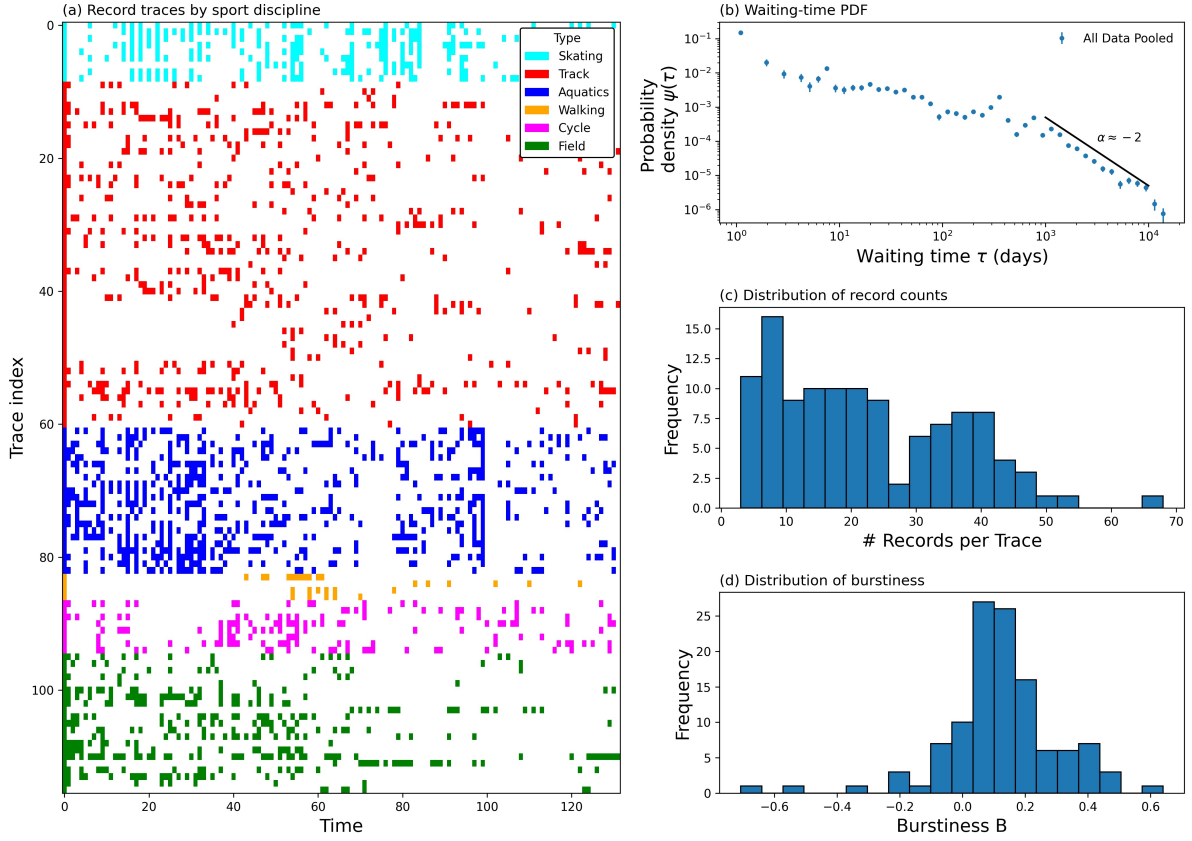


Figure 5.7: Introduction of the dataset of records in sport. In panel **(a)** we show record times since 1960 as binary data points. For this we first bin the time axis and record each bin as a record time if one or more records fall into it. The binary traces are further organized by sport type. While records have become rarer in the disciplines of Track and Field, frequent record breaking is still observed for swimming and technology driven sports, such as ice skating and race cycling. **(b)** We also calculate the waiting time distribution in all data traces together and observe three broad regimes. On the scale of up to a year, record breaking occurs with more or less constant or slightly decreasing rate. On the scale of one to ten years, the waiting time for the next record decays as a powerlaw with exponent $\alpha \approx -2$. Finally, beyond a waiting time of ten years, the probability density features a cutoff. In panel **(c)** we show the distribution of record counts, which falls into a broad range from a few to more than 50 since 1960. Finally, in **(d)** we feature the distribution of the burstiness, $B = (\sigma - \mu)/(\mu + \sigma)$ [GB08], of record traces, where μ and σ are the mean and the standard deviation of the waiting time distribution.

rugged fitness landscape [Wri32, Kau93, KL87]. In the end of chapter 3 we introduced a the record process with reset and argued that it could serve as a model of evolution on non-linear fitness landscapes (see methods). Curiously, the waiting time distribution in the record process with reset features a power law exponent in between -1 and -2. The exponent of -2, as seen in Fig. 5.7 emerges if every record resets the record process and therefore corresponds to a regime of fast resetting.

We propose a record process model, which combines the linear trend in cumulative records and the observed clustering of records (see methods). The adjacent possible as defined as the current probability to obtain a record after n previous records, would generally decrease in an exponential fashion for an i.i.d. record process. For a model of constant drift, the adjacent possible is expected to stay more or less constant. We can reconcile both perspectives by posing a minimal size of the adjacent possible. Each record diminishes the probability to find further records, but the probability can not fall below a certain fixed level given by the constant drift in the system.

More formally, we consider classic i.i.d. record process in which a record broken with probability $p_r = 1/(t - t_0)$ if the process started at time t_0 . The process is modified by (1) the addition of a constant base-line probability, p , for record breaking such that $p_r = (1 - p)/(t - t_0) + p$, modelling drift, and (2) a reset-probability η , by which we set $t_0 = t$, each time a record is obtained.

We use a likelihood-based approach to test the model against competing hypotheses, Fig. 5.8. In particular we are comparing an i.i.d. record model, a model with constant drift, a model on a logistically growing population and a record model with reset and drift. We find that the i.i.d record model fails to explain the record data. The record model with reset and constant drift is often a better candidate than the model with logistic growth and the binary model - in particular for those sports for which strong clustering is observed, such as Pole Vault. For sports with little clustering the record traces can be well explained by drift or logistic growth alone, such as Marathon. Further, the burstiness is typically underestimated by models which take into account drift alone. The reset model predicts the observed burstiness. Conversely, the reset model underpredicts the total number of observed records, which is better explained by models of drift alone.

5.3 Discussion

Our results reveal a broadly consistent picture of sport performance over the last century: long phases of near-linear record progression followed by a marked slowdown after 1980. Such patterns cannot be reconciled with simple i.i.d. record models, and while expanding participation explains part of the trend, earlier work has already emphasized that population growth alone cannot account for the richness of sport evolution [Yan75]. Our findings underscore that cultural-evolutionary processes – innovation, copying, and refinement – are central to understanding how records advance.

A central challenge in applying evolutionary reasoning outside biology is defining fitness. In most cultural domains, fitness is ambiguous: profit in economics, influence in art, efficiency in engineering. Sports are exceptional in that performance outcomes provide an unambiguous, quantitative measure of fitness. This makes them a rare case where methods developed for population genetics can be applied almost directly. Indeed, we observed empirical signatures of

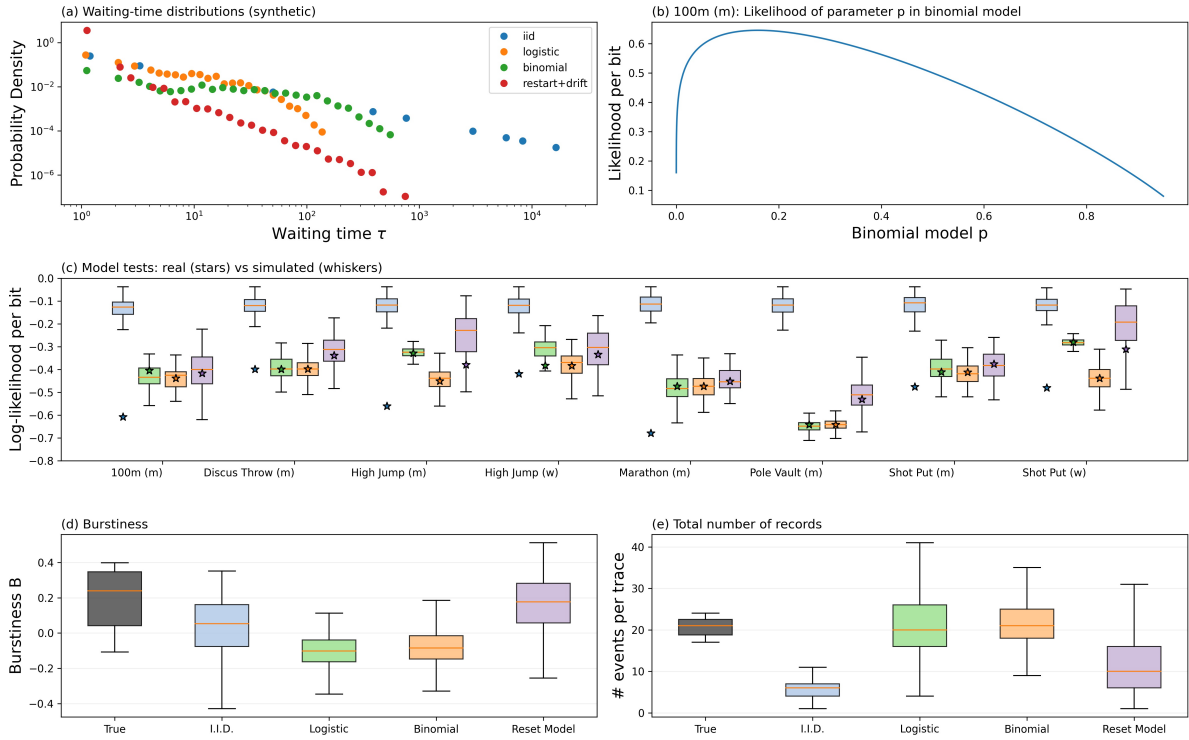


Figure 5.8: Comparison of record models for a set of sport records. We bin record traces into bins of half year size, as this is the time scale on which major athletic tournaments return. We are comparing four different models and their capability to explain the data. These models feature different waiting time distributions, which we show in panel (a). The waiting times in the i.i.d. model follow a power law with power -1 without cutoff. The binary and the logistic growth model feature exponential decay. The waiting times in the record model with reset decay as a powerlaw with a tunable exponent. We then compute the likelihood of each model given the data and estimate the parameter with a simple grid search. The likelihood of parameter p in the binary model is shown given the data of the 100 meter running event. In panel (c) we then show the compiled results for 8 different sports, which we considered representative of both drift dominated and innovation dominated dynamics. The loglikelihood of the data is indicated as a star. Furthermore we have used the optimal parameters given the data to simulate synthetic records with each model class. The distribution of loglikelihood of these synthetic record traces are shown as whisker box plots. Broadly, if the star falls into the range of the synthetic data, we can say that the model explains the data. In panel (d) we compare the burstiness of all real records in the database with the burstiness of simulated records under optimally parametrized models. The reset model produces burstiness scores similar to the data. In panel (e) we compare the total number of real records with the the number of simulated records under optimally parametrized models. The logistic and the binary model explain the data well, while the reset model generally predicts too few records.

Breeder's equation: variance predicts response, heritability can be measured, and phenotypic shifts accumulate across decades.

In this context, it is important to clarify what we mean by inheritance in sports. Not all determinants of performance are transmissible: weather conditions or chance events leave no legacy. In evolutionary biology these factors are also called environmental. But training regimes, nutritional practices, body morphologies, and technical innovations can be copied, imitated, and refined. It is through these channels that heritability in sports can be meaningfully defined, and why population-genetic concepts such as Breeder's equation are useful analogies.

At the level of records, the temporal distribution of achievements is not memoryless. Record traces cluster in time, producing heavy-tailed waiting-time distributions. Our record model with reset captures this burstiness but systematically underestimates the total number of records. This tension likely stems from the fact that resets are themselves triggered by records: while this generates clustering, it also penalizes strong drift, which would otherwise be needed to explain the absolute number of records observed. Thus, burstiness and overall frequency of records appear to be governed by partially conflicting mechanisms.

Correlations among records may also arise even without evolutionary dynamics, depending on the shape of the underlying performance distribution [WFK11, FWK12]. In the Gumbel class of exponentially tailed distributions – appropriate when outcomes are the sum of many small effects – record events remain uncorrelated. This makes Gumbel a plausible model for athletic performance. Yet as athletes approach physiological ceilings, distributions may shift towards the Weibull domain, where correlations naturally emerge. We have not tested this hypothesis directly, but it offers a possible alternative explanation for observed dependencies visible in the data. For details see [WFK11, FWK12].

These observations connect naturally to the metaphor of rugged fitness landscapes [Wri32] and punctuated equilibria [EG⁺72]. Record progress is often piecewise, with bursts of rapid improvement followed by long plateaus. Innovations such as the Fosbury flop exemplify epistatic change: multiple elements must be altered simultaneously to unlock a new performance regime. Such events reflect creative destruction, where established optima collapse in favor of radically different solutions [LKW24]. The observed boom–reconsolidate dynamics of variance and mean performance suggest that athletes are navigating complex, multi-peaked landscapes rather than smooth gradients.

This raises the question of open-endedness. For much of the 20th century, records advanced steadily, but the slowdown after 1980 suggests that the adjacent possible – while fluctuating in size – is shrinking overall and that the space of innovations is becoming depleted. A possible extension of the suggested models, which is to consider in future work, are time-dependent decreasing probabilities for reset. Our original inspiration lies in biology, where experiments such as Lenski's long-term evolution project demonstrate open-ended adaptation under constant conditions [EL03]. Whether sports can exhibit a comparable trajectory depends on whether new techniques, technologies, or body types continue to emerge.

More broadly, the lessons of sport extend well beyond athletics. Sports provide a uniquely well-measured testbed for cultural evolution, but the same dynamics apply to other domains where records are carefully tracked: video game speedrunning, technological performance indicators such as solar cell prices, or learning curves in cognition and optimization algorithms. In all of these cases, extremes rather than averages capture the essence of progress. Sports thus serve as a model system for studying how innovation unfolds.

5.4 Methods

5.4.1 Data Acquisition

In this work we have used four different dataset.

1. Data on the Comrades Cup, used to create Fig. 5.4, is publicly available on Finsh-Time.co.za at `results.finishtime.co.za/Comrades.aspx` [fin].
2. Data on physiological traits of Olympian athletes was obtained via scraping from `www.sports-reference.com/` by Kaggle user heesoo37 in 2018 [hee18]. As of today sport-reference.com no longer provides data on the Olympics. This dataset contains data on athletes weight and height, as well as the won medal if any. We used it to create results shown in Fig. 5.5. It does not contain information on individual performance.
3. Full performance results were obtained from `ww.olympiandatabase.com` an independently compiled database by Olympic enthusiasts [olynd]. From this database we built results shown in Fig. 5.5.
4. The data on records were taken from `www.worldathletics.com`, the official database for results in Track & Field events [wornd]. World record data for aquatic, cycle and skating events were taken from Wikipedia.

Record statistics in exponentially growing populations For continuous i.i.d. samples, the expected number of records, $n(t)$, grows as a harmonic sum $n(t) = H_n \approx \log t + \gamma$, since the the probability, $p_r(t)$, of a record a time t is universally given by $p_r(t) = 1/t$ [Gli78]. Assume the number of samples at each time step is growing like $N(t)$ and a record is counted only once for each timestep. Adding samples from $N(t)$ to $N(t+1)$ samples yields an expected increase in the cumulative number of records of

$$n(t+1) - n(t) \approx \frac{dn(N(t))}{dt} = \frac{dn}{dN} \frac{dN}{dt} = \frac{1}{N} \frac{dN}{dt} = \frac{d \log(N)}{dt} \quad (5.1)$$

If we consider a record process with growing sampling size, $N(t) = N_0 e^{rt}$, the cumulative number of records grows linearly,

$$n(t+1) - n(t) \approx r \quad (5.2)$$

More realistically for our setting, under logistic growth $\dot{N}(t) = rN(t)(1 - N(t)/K)$ and $N(t) = K/(1 + Ae^{-rt})$ we have,

$$\frac{dn(N(t))}{dt} = r \left(1 - \frac{N(t)}{K} \right) = \frac{rAe^{-rt}}{1 + Ae^{-rt}} \quad , \quad (5.3)$$

which is independent of K and decreasing after a transient phase of linear growth. Conversely, the record probability is constant for a while before it decays with a power of -1 for large t .

The mean of the maximum for exponential distributions. Assume random variables drawn from an exponential distribution $X_i \sim \text{Exp}(\lambda)$ and call, $M_n = \max_i X_i$, the maximum value of n such variables, then

$$\mathbb{E}[M_n] = \int_0^\infty (1 - F(x)^n) dx, \quad (5.4)$$

where $F(x) = 1 - e^{-x/\lambda}$, is the CDF of the exponential distribution and we have used that the X_i are i.i.d. Substituting $y = e^{-x/\lambda}$ gives

$$\mathbb{E}[M_n] = \lambda \int_0^1 \frac{1 - (1 - y)^n}{y} dy = \lambda H_n. \quad (5.5)$$

For distributions with exponential tail, the expected maximum value grows proportional to the number of observed records. For exponentially growing samples this implies a linear growth of the expected maximum value.

Wright-Fisher simulation with selection and mutation We simulate a population of constant size N with clonal reproduction and additive mutation in log-fitness. Denote by $\hat{w}_i^{(t)}$ the log-fitness of individual i at generation t . First, we draw the initial population $\hat{w}_i^{(0)} \sim \mathcal{N}(\mu_0, \sigma_0^2)$ independently for $i = 1, \dots, N$. Then we compute reproductive weights

$$w_i^{(t)} = \exp(\hat{w}_i^{(t)} - \max_j \hat{w}_j^{(t)}), \quad p_i^{(t)} = \frac{w_i^{(t)}}{\sum_{j=1}^N w_j^{(t)}}, \quad (5.6)$$

and sample N parents with replacement according to $\text{Multinomial}(N; p_1^{(t)}, \dots, p_N^{(t)})$, the multinomial distribution. We set $\hat{w}_{\text{par}(k)}^{(t)}$ to be the log-fitness of the parent of offspring k . Each offspring individual mutates independently with probability μ . If a mutation occurs, we draw an effect $\Delta_k \sim \mathcal{N}(\hat{w}_{\text{par}(k)}^{(t)}, \sigma_m^2)$ (independent of the parent) and set

$$\hat{w}_k^{(t+1)} = \hat{w}_{\text{par}(k)}^{(t)} + \Delta_k \quad (5.7)$$

otherwise we set $\hat{w}_k^{(t+1)} = \hat{w}_{\text{par}(k)}^{(t)}$. This produces the next generation.

Heritability and Selection In population genetics, *heritability* refers to the proportion of variation in a trait within a population that can be attributed to genetic differences among individuals, rather than to environmental factors. It is usually expressed as a value between 0 and 1: a value close to 1 means that most of the observed variation in the trait is due to genetic differences, while a value near 0 means that environmental influences dominate. Importantly, heritability does not measure how much of a trait in a single individual is “caused” by genes, but rather describes variation at the population level. A classic way to define heritability comes from the Breeder’s equation [FM83],

$$R = h^2 S \quad (5.8)$$

where R is the response to selection (the change in the mean of a trait from one generation to the next), S is the selection differential (the difference between the mean trait value of the selected parents and the population as a whole), and h^2 is the *narrow-sense heritability*. In this

framework, heritability quantifies how effectively selection on a trait translates into evolutionary change. In this work we extended this framework to the evolution of performance results in the Olympics. In analogy to the biological setting, we assume that the best performing athletes or techniques/ body types are “selected” and partly inherited by culture evolutionary processes of copying. We define R as the difference in mean performance, $R_t = \bar{z}_{t+1} - \bar{z}_t$, from one year to the next. Similarly we assume that the selection differential is proportional to the variance in performance, $S \sim \sigma_t(z)$. We generalize equation 5.8 to the setting in sports

$$R_t = h_a^2 \sigma_t^2 \quad (5.9)$$

and measure R_t via regression (Fig. 5.5). h_a^2 , the generalized heritability score, is not bound to the interval $[0, 1]$ since $\sigma_t(z)$ is only an estimate for the true unknown S_t . Using alternative definitions of the selection differential we may pose that the response is proportional to the difference of the average trait of the best performing individuals (which will eventually be selected) and the average performance overall $S \sim \bar{z}_t^{best} - \bar{z}_t$. We called this quantity “selection strength” β in Fig. 5.6. The therein observed correlation between trait evolution and β can therefore be interpreted as yet another instantiation of Breeder’s equation.

Record process with reset and drift We consider the record process with reset (which was already introduced in chapter 3). This process starts as a normal record process in which records are broken with probability $p_r = 1/t$. Every time we observe a record, we draw a Bernoulli random variable, ν , with success probability η to decide whether an “innovation” was obtained. If $\nu = 1$ is obtained at a record at time t_* , the probability of a record is updated to $p_r = 1/(t - t_*)$ for all times $t_* > t$, until the next innovation is found. If $\nu = 0$ no p_r stays unaffected. This process generates a marginal record probability which decays as a powerlaw $p_r \sim t^{\eta-1}$. Furthermore, the waiting time distribution in between records decays as $\psi(\tau) \sim t^{-1-\eta}$. We extend this framework in two ways. (1) We introduce drift, by setting the probability of a record at any time to

$$p_r(t > t_*) = \frac{1 - p}{t - t_*} + p \quad , \quad (5.10)$$

where t_* is the time of the last innovation event. Secondly, we add a parameter q , which models a possible overbinning effect. In equation t is the number of attempts we had at obtaining a record. If at every timestep one record is observed, t is equivalent to true time. However, if timesteps are chosen to be small, there might be bins which are completely empty. The probability of obtaining a record is then, given by

$$p_r = \frac{1 - q^t}{t} \quad , \quad (5.11)$$

where q is the probability that there is no event in a timebin. Further, we set $p = 0$ and $t_* = 0$. We proof this equation.

A record at time t requires (i) that a draw occurs at t (probability $1 - q$) and (ii) that this draw is a record given the N_{t-1} previous draws. Thus

$$p_r = (1 - q) \mathbb{E} \left[\frac{1}{N_{t-1} + 1} \right], \quad N_{t-1} \sim \text{Binomial}(t - 1, 1 - q). \quad (5.12)$$

We use the following identity, valid for $N \sim \text{Binomial}(m, 1 - q)$,

$$\begin{aligned}
 \mathbb{E}\left[\frac{1}{N+1}\right] &= \sum_{n=0}^m \frac{1}{n+1} \binom{m}{n} (1-q)^n q^{m-n} \\
 &= \frac{1}{m+1} \sum_{n=0}^m \binom{m+1}{n+1} (1-q)^n q^{m-n} \\
 &= \frac{1}{(m+1)p} \sum_{n=0}^m \binom{m+1}{n+1} (1-q)^{n+1} q^{m-n} \\
 &= \frac{((1-q) + q)^{m+1} - q^{m+1}}{(m+1)(1-q)} .
 \end{aligned} \tag{5.13}$$

With $m = t - 1$ this yields the exact formula 5.11 which decays as $1/t$ after a transient roughly constant initial period. The full model equation or the probability of a record then becomes

$$p_r(t > t_*) = \frac{1-p}{t-t_*} (1 - q^{t-t_*}) + p \quad . \tag{5.14}$$

where p is a rescaled parameter by q .

Burstiness parameter We adopt the burstiness score as introduced by Goh & Barabasi [GB08],

$$B = \frac{\sigma_\tau - \mu_\tau}{\mu_\tau + \sigma_\tau} \tag{5.15}$$

where μ_τ and σ_τ are the mean and standard deviation of the waiting times τ in binary data, in our case the waiting times between record events. $B \in [-1, 1]$ and positive values indicate clumped data.

Likelihood estimation of record process models In order to compare different generative models of record-breaking, we adopt a *maximum-likelihood approach*. The central idea is that each candidate model defines a probability distribution over record traces, i.e. sequences of binary outcomes $(x_1, \dots, x_T)$ with $x_t = 1$ if a record occurs at time t , and $x_t = 0$ otherwise. Given observed data, the likelihood of a model is the probability that the model generates exactly this trace. Taking the logarithm yields the *log-likelihood*, which is additive in time and numerically more stable. Model parameters are then estimated by maximizing this log-likelihood.

Formally, if $p_t(\theta)$ denotes the model's predicted probability of a record at time t under parameters θ , the log-likelihood of an observed trace $\{x_t\}$ is

$$\ell(\theta) = \sum_{t=1}^T \left[x_t \ln p_t(\theta) + (1 - x_t) \ln(1 - p_t(\theta)) \right]. \tag{5.16}$$

The log-likelihood per bit is simply $\bar{\ell} = \ell/T$. In the i.i.d. record process we have $p_t = 1/t$. Thus

$$\ell_{\text{iid}} = \sum_{t=1}^T \left[x_t \ln\left(\frac{1}{t}\right) + (1 - x_t) \ln\left(1 - \frac{1}{t}\right) \right]. \tag{5.17}$$

If the underlying sampling size grows logistically, the implied record probability is a function $p_t(\vartheta)$ of growth parameters $\vartheta = (k, x_0)$, $p_t(k, x_0) = N(t) / \sum_k^t N(k)$, with $N(t) = K / (1 + \exp(-k(t - x_0)))$.

$$\ell_{\text{logistic}}(\vartheta) = \sum_{t=1}^T \left[x_t \ln p_t(\vartheta) + (1 - x_t) \ln(1 - p_t(\vartheta)) \right]. \quad (5.18)$$

The log-likelihood is independent of the capacity parameter K .

In the binomial model the record probability is constant $p_t \equiv p \in (0, 1)$, so

$$\ell_{\text{binomial}}(p) = \sum_{t=1}^T \left[x_t \ln p + (1 - x_t) \ln(1 - p) \right] = n(T) \ln p + (T - n(T)) \ln(1 - p), \quad (5.19)$$

where $n(t) = \sum_{t=1}^T x_t$ is the total number of records. This is an accurate model of records under constant drift.

Recursive likelihood estimation of record process model with reset Let the age $a \geq 1$ be the number of steps since the last reset (or since the beginning). Conditional on age a at the start of step t , define the record/event probability by

$$h(a; q, p) = \frac{(1 - q^a)(1 - p)}{a} + p, \quad (5.20)$$

where $p \in [0, 1)$ is an age-independent (“drift”) event channel and $q \in [0, 1)$ controls how the age-dependent component accumulates with age a . Let $\pi_t(a)$ be our current belief for the age distribution at the start of step t , given observations up to $t - 1$. The (marginal) record probability at time t is

$$p_t(\eta, p, q) = \sum_{a \geq 1} \pi_t(a) h(a; q, p) = \sum_{a \geq 1} \pi_t(a) \left[\frac{(1 - q^a)(1 - p)}{a} + p \right]. \quad (5.21)$$

Initially we set $\pi_1(1) = 1$ and $\pi_1(a) = 0$ for $a > 1$. Given an observation $x_t \in \{0, 1\}$ at time t , we update π_{t+1} by

$$\pi_{t+1}(a) = \begin{cases} \eta, & \text{if } x_t = 1 \text{ and } a' = 1, \\ (1 - \eta) \cdot \frac{\pi_t(a - 1) h(a - 1; q, p)}{p_t}, & \text{if } x_t = 1 \text{ and } a' \geq 2, \\ \frac{\pi_t(a - 1) (1 - h(a - 1; q, p))}{1 - p_t}, & \text{if } x_t = 0 \text{ and } a' \geq 2, \\ 0, & \text{if } x_t = 0 \text{ and } a' = 1, \end{cases} \quad (5.22)$$

The lines in this conditional equation denote the following:

1. We observe a record ($x_t = 1$) and an innovation (reset). With probability η the any previous age falls back to $a = 1$.

2. We observe a record ($x_t = 1$) but no innovation (no reset). We use *Bayes' theorem* to construct the posterior probability that the age was a , given that a record is observed. $\pi_t(a | x_t = 1) = p(x_t = 1 | a)\pi(a)/p(x_t = 1) = h(t | a)\pi(a)/p_t$. In this case the age is advanced by 1, $\pi_{t+1}(a + 1 | x_t = 1) = \pi_t(a | x_t = 1)$. This happens with overall no-innovation probability $(1 - \eta)$.
3. We observe no record ($x_t = 0$). Similarly to the case above, we use Bayes theorem to find our beliefs about the age a at time t , given that no record occurs at time t , $\pi_t(a | x_t = 0) = p(x_t = 0 | a)\pi(a)/p(x_t = 0) = \pi(a)(1 - h(a))/(1 - p_t)$.
4. If no record is observed, a is never reset.

Finally, after constructing p_t iteratively (together with the latent distribution π_t , we calculate the log-likelihood as

$$\ell_{\text{reset/drift}}(\eta, p, q) = \sum_{t=1}^T \left[x_t \ln p_t(\eta, p, q) + (1 - x_t) \ln(1 - p_t(\eta, p, q)) \right] . \quad (5.23)$$

Likelihood-optimization We perform numeric likelihood-optimization via a naive grid search of the model parameters for the logistic and the reset model with drift. The i.i.d. record model can not be optimized and is directly evaluated using Eq. 5.17. The binomial model can be analytically optimized from Eq. 5.19, yielding the optimal parameter $\hat{p}_{\text{binomial}} = n/T$, where n is the total number of records in the sequence.

Bibliography

- [A⁺08] F. E. Andre et al. Vaccination greatly reduces disease, disability, death and inequity worldwide. *Bulletin of the World Health Organization*, 86:140–146, 2008.
- [A⁺20a] A. Aleta et al. Modeling the impact of social distancing, testing, contact tracing and household quarantine on second-wave scenarios of the covid-19 epidemic. *medRxiv*, 2020.
- [A⁺20b] E. Andreano et al. Sars-cov-2 escape in vitro from a highly neutralizing covid-19 convalescent plasma. *bioRxiv*, 2020.
- [A⁺20c] V. A. Avanzato et al. Case study: Prolonged infectious sars-cov-2 shedding from an asymptomatic immunocompromised individual with cancer. *Cell*, 183:1901–1912.e9, 2020.
- [AAY18] Meghan O. Altman, Davide Angeletti, and Jonathan W. Yewdell. Antibody immunodominance: The key to understanding influenza virus antigenic drift. *Viral Immunology*, 31(2):142–149, 2018.
- [ABB⁺23] Chiara Agrati, Barbara Bartolini, Veronica Bordoni, Franco Locatelli, Maria Rosaria Capobianchi, Antonino Di Caro, Concetta Castilletti, and Giuseppe Ippolito. Emerging viral infections in immunocompromised patients: a great challenge to better define the role of immune response. *Frontiers in immunology*, 14:1147871, 2023.
- [Ado98] Luciano Adorini. Immunodominance. In *Encyclopedia of Immunology*, pages 1290–1292. Elsevier, 1998.
- [AHD14] B. M. Althouse and L. HÃ¶lbert-Dufresne. Epidemic cycles driven by host behaviour. *Journal of the Royal Society Interface*, 11:20140575, 2014.
- [Al12] Ali Akram and Robert D. Inman. Immunodominance: A pivotal principle in host response to viral infections. *Clinical Immunology*, 143(2):99–115, 2012.
- [AK15] B. Ashby and K. C. King. Diversity and the maintenance of sex by parasites. *Journal of Evolutionary Biology*, 28(3):511–520, 2015.
- [Alc50] Armen A Alchian. Uncertainty, evolution, and economic theory. *Journal of political economy*, 58(3):211–221, 1950.
- [Ald06] David J Aldous. Exchangeability and related topics. In *École d’Été de Probabilités de Saint-Flour XIIIâ€”1983*, pages 1–198. Springer, 2006.

- [AMB21] D. Adak, A. Majumder, and N. Bairagi. Mathematical perspective of covid-19 pandemic: Disease extinction criteria in deterministic and stochastic models. *Chaos, Solitons & Fractals*, 142:110381, 2021.
- [Apo13] Tom M Apostol. *Introduction to analytic number theory*. Springer Science & Business Media, 2013.
- [ATS20] Y. Alimohamadi, M. Taghdir, and M. Sepandi. Estimate of the basic reproduction number for covid-19: A systematic review and meta-analysis. *Journal of Preventive Medicine and Public Health*, 53:151–157, 2020.
- [AVTC20] R. M. Anderson, C. Vegvari, J. Truscott, and B. S. Collyer. Challenges in creating herd immunity to sars-cov-2 infection by mass vaccination. *Lancet*, 396:1614–1616, 2020.
- [B⁺20] C. O. Barnes et al. Sars-cov-2 neutralizing antibody structures inform therapeutic strategies. *Nature*, 588:682–687, 2020.
- [B⁺21] L. R. Baden et al. Efficacy and safety of the mrna-1273 sars-cov-2 vaccine. *New England Journal of Medicine*, 384:403–416, 2021.
- [BA80] J. G. Breman and I. Arita. The confirmation and maintenance of smallpox eradication. *The New England Journal of Medicine*, 303(22):1263–1273, 1980.
- [BGAC04] M. F. Boni, J. R. Gog, V. Andreasen, and F. B. Christiansen. Influenza drift and epidemic size: The race between generating and escaping immunity. *Theoretical Population Biology*, 65:179–191, 2004.
- [Bil72] Patrick Billingsley. On the distribution of large prime divisors. *Periodica Mathematica Hungarica*, 2(1-4):283–289, 1972.
- [BMSN97] S. Bonhoeffer, R. M. May, G. M. Shaw, and M. A. Nowak. Virus dynamics and drug therapy. *Proceedings of the National Academy of Sciences of the United States of America*, 94(13):6971–6976, 1997.
- [Bon08] M. F. Boni. Vaccination and antigenic drift in influenza. *Vaccine*, 26 Suppl 3:C8–14, 2008.
- [BP02] Stefan Boettcher and Allon G Percus. Optimization with extremal dynamics. *complexity*, 8(2):57–62, 2002.
- [BR85] Robert Boyd and Peter J. Richerson. *Culture and the Evolutionary Process*. University of Chicago Press, Chicago, 1985.
- [BS93] Per Bak and Kim Sneppen. Punctuated equilibrium and criticality in a simple model of evolution. *Physical review letters*, 71(24):4083, 1993.
- [BSD⁺21] G. A. Bazykin, O. Stanevich, D. Danilenko, A. Fadeev, K. Komissarova, A. Ivanova, M. Sergeeva, K. Safina, E. Nabieva, G. Klink, S. Garushyants, J. Zabutova, A. Kholodnaia, I. Skorokhod, V. V. Ryabchikova, A. Komissarov, and D. Lioznov. Emergence of y453f and delta69-70hv mutations in a lymphoma patient with long-term covid-19, 2021.

- [BSKA20] O. N. Bj  rnstad, K. Shea, M. Krzywinski, and N. Altman. Modeling infectious epidemics. *Nature Methods*, 17:455–456, 2020.
- [BTT⁺08] G. Berthelot, V. Thibault, M. Tafflet, S. Escolano, N. El Helou, X. Jouven, et al. The citius end: World records progression announces the completion of a brief ultra-physiological quest. *PLOS ONE*, 3(2):e1552, 2008.
- [BTW⁺18] Dan H. Barouch, Frank L. Tomaka, Frank Wegmann, Daniel J. Stieh, Galit Alter, Merlin L. Robb, Nelson L. Michael, et al. Evaluation of a mosaic hiv-1 vaccine in a multicentre, randomised, double-blind, placebo-controlled, phase 1/2a clinical trial (approach) and in rhesus monkeys (nhp 13-19). *The Lancet*, 392(10143):232–243, 2018.
- [BV16] O. N. Bj  rnstad and C. Viboud. Timing and periodicity of influenza epidemics. *Proceedings of the National Academy of Sciences USA*, 113:12899–12901, 2016.
- [BZA⁺21] Arman A. Bashirova, Wanjing Zheng, Marjan Akdag, Danillo G. Augusto, Nicolas Vince, Krista L. Dong, Colm O’hUigin, and Mary Carrington. Population-specific diversity of the immunoglobulin constant heavy g chain (ighg) genes. *Genes and Immunity*, 22(7-8):327–334, 2021.
- [C⁺18] H. Chabas et al. Evolutionary emergence of infectious diseases in heterogeneous host populations. *PLoS Biology*, 16:e2006738, 2018.
- [C⁺20a] B. Choi et al. Persistence and evolution of sars-cov-2 in an immunocompromised host. *New England Journal of Medicine*, 383:2291–2293, 2020.
- [C⁺20b] R. Chowdhury et al. Dynamic interventions to control covid-19 pandemic: A multivariate prediction modelling study comparing 16 worldwide countries. *European Journal of Epidemiology*, 35:389–399, 2020.
- [C⁺21] S. Cele et al. Escape of sars-cov-2 501y.v2 from neutralization by convalescent plasma. *Nature*, 593:142–146, 2021.
- [CC10] B. Charlesworth and D. Charlesworth. *Elements of Evolutionary Genetics*. Roberts & Company, 2010.
- [CCH⁺21] T. Carleton, J. Cornetet, P. Huybers, K. C. Meng, and J. Proctor. Global evidence for ultraviolet radiation decreasing covid-19 growth rates. *Proceedings of the National Academy of Sciences USA*, 118:e2012370118, 2021.
- [CCR⁺20] Bina Choi, Manish C. Choudhary, James Regan, Jeffrey A. Sparks, Robert F. Padera, Xueting Qiu, Isaac H. Solomon, et al. Persistence and evolution of sars-cov-2 in an immunocompromised host. *The New England Journal of Medicine*, 383(23):2291–2293, 2020.
- [CCS20] G. Cacciapaglia, C. Cot, and F. Sannino. Second wave covid-19 pandemics in europe: A temporal playbook. *Scientific Reports*, 10:15514, 2020.
- [CGP05] Y. Cao, D. T. Gillespie, and L. R. Petzold. Avoiding negative populations in explicit poisson tau-leaping. *Journal of Chemical Physics*, 123:054104, 2005.

[Cha52] KN53463 Chandler. The distribution and frequency of record values. *Journal of the Royal Statistical Society: Series B (Methodological)*, 14(2):220–228, 1952.

[CKF17] Nicole Creanza, Oren Kolodny, and Marcus W Feldman. Cultural evolutionary theory: How culture evolves and why it matters. *Proceedings of the National Academy of Sciences*, 114(30):7782–7789, 2017.

[CLN⁺18] Hélène Chabas, Sébastien Lion, Antoine Nicot, Sean Meaden, Stineke van Houte, Sylvain Moineau, Lindi M. Wahl, Edze R. Westra, and Sylvain Gandon. Evolutionary emergence of infectious diseases in heterogeneous host populations. *PLoS Biology*, 16(9):e2006738, 2018.

[CM10] Lawrence Corey and M. Juliana McElrath. Hiv vaccines: Mosaic approach to virus diversity. *Nature Medicine*, 16(3):268–270, 2010.

[CMFS11] Bernat Corominas-Murtra, Jordi Fortuny, and Ricard V Solé. Emergence of zipf’s law in the evolution of communication. *Physical Review E—Statistical, Nonlinear, and Soft Matter Physics*, 83(3):036115, 2011.

[CMHT15] Bernat Corominas-Murtra, Rudolf Hanel, and Stefan Thurner. Understanding scaling through history-dependent processes with collapsing sample space. *Proceedings of the National Academy of Sciences*, 112(17):5348–5353, 2015.

[CMHT16] Bernat Corominas-Murtra, Rudolf Hanel, and Stefan Thurner. Extreme robustness of scaling in sample space reducing processes explains zipf’s law in diffusion on directed networks. *New Journal of Physics*, 18(9):093010, 2016.

[CMHZZ18] Bernat Corominas-Murtra, Rudolf Hanel, Leonardo Zavoian, and Stefan Thurner. How driving rates determine the statistics of driven non-equilibrium systems with stationary distributions. *Scientific Reports*, 8(1):10837, 2018.

[CMR10] Dylan Z. Childs, C. J. E. Metcalf, and Mark Rees. Evolutionary bet-hedging in the real world: Empirical evidence and challenges revealed by plants. *Proceedings. Biological Sciences / the Royal Society*, 277(1697):3055–3064, 2010.

[D⁺20a] B. L. Dickens et al. Modelling lockdown and exit strategies for covid-19 in singapore. *Lancet Regional Health – Western Pacific*, 1:100004, 2020.

[D⁺20b] Y. Dong et al. A systematic review of sars-cov-2 vaccine candidates. *Signal Transduction and Targeted Therapy*, 5:237, 2020.

[D⁺21] N. G. Davies et al. Estimated transmissibility and impact of sars-cov-2 lineage b.1.1.7 in england. *Science*, 372, 2021.

[Dar59] Charles Darwin. *On the Origin of Species by Means of Natural Selection, or the Preservation of Favoured Races in the Struggle for Life*. John Murray, London, 1859.

[DDG20] E. Dong, H. Du, and L. Gardner. An interactive web-based dashboard to track covid-19 in real time. *Lancet Infectious Diseases*, 20:533–534, 2020.

[DDMF17] Ramses Djidjou-Demasse, Benoît Moury, and Frédéric Fabre. Mosaics often outperform pyramids: Insights from a model comparing strategies for the deployment of plant resistance genes against viruses in agricultural landscapes. *The New Phytologist*, 216(1):239–253, 2017.

- [DeA08] M. DeArce. Correspondence of charles darwin on james torbitt's project to breed blight-resistance potatoes. *Archives of Natural History*, 35(2):208–222, 2008.
- [DF07] Michael M Desai and Daniel S Fisher. Beneficial mutation–selection balance and the effect of linkage on positive selection. *Genetics*, 176(3):1759–1798, 2007.
- [DHF06] Laurens De Haan and Ana Ferreira. *Extreme value theory: an introduction*. Springer, 2006.
- [eCHGL07] Mian er Cong, Walid Heneine, and J. Gerardo García-Lerma. The fitness cost of mutations associated with human immunodeficiency virus type 1 drug resistance is modulated by mutational interactions. *Journal of Virology*, 81(6):3037–3041, 2007.
- [ECL96] Santiago F Elena, Vaughn S Cooper, and Richard E Lenski. Punctuated evolution caused by selection of rare beneficial mutations. *Science*, 272(5269):1802–1804, 1996.
- [EE04] Warren John Ewens and WJ Ewens. *Mathematical population genetics: theoretical introduction*, volume 27. Springer, 2004.
- [EG⁺72] Niles Eldredge, Stephen Jay Gould, et al. Punctuated equilibria: an alternative to phyletic gradualism. *Models in paleobiology*, 82:115, 1972.
- [EK09] Iddo Eliazar and Joseph Klafter. Record events in growing populations: Universality, correlation, and aging. *Physical Review E—Statistical, Nonlinear, and Soft Matter Physics*, 80(6):061117, 2009.
- [EL03] Santiago F Elena and Richard E Lenski. Evolution experiments with microorganisms: the dynamics and genetic bases of adaptation. *Nature Reviews Genetics*, 4(6):457–469, 2003.
- [Eps14] David Epstein. *The sports gene: Inside the science of extraordinary athletic performance*. Penguin, 2014.
- [ERT20] M. S. Eichenbaum, S. Rebelo, and M. Trabandt. The macroeconomics of epidemics. Technical report, National Bureau of Economic Research, 2020.
- [ES05] S. F. Elena and R. Sanjuán. Adaptive value of high mutation rates of rna viruses: Separating causes from consequences. *Journal of Virology*, 79:11555–11558, 2005.
- [Esh84] I. Eshel. On the survival probability of a slightly advantageous mutant gene in a multitype population: A multidimensional branching process model. *Journal of Mathematical Biology*, 19:201–209, 1984.
- [Ewe72] Warren J Ewens. The sampling theory of selectively neutral alleles. *Theoretical population biology*, 3(1):87–112, 1972.
- [Ewe04] Warren J. Ewens. *Mathematical Population Genetics I: Theoretical Introduction*, volume 27 of *Interdisciplinary Applied Mathematics*. Springer, New York, 2 edition, 2004.

- [F⁺20] S. Flaxman et al. Estimating the effects of non-pharmaceutical interventions on covid-19 in europe. *Nature*, 584:257–261, 2020.
- [F⁺21] N. R. Faria et al. Genomics and epidemiology of the p.1 sars-cov-2 lineage in manaus, brazil. *Science*, 372:815–821, 2021.
- [FH20] M. Fudolig and R. Howard. The local stability of a modified multi-strain sir model for emerging viral strains. *PLoS ONE*, 15:e0243408, 2020.
- [fin] Comrades marathon results. Online results portal. Accessed: 23 August 2025.
- [FLT⁺19] Giovanni Franzo, Matteo Legnardi, Claudia Maria Tucciarone, Michele Drigo, Marco Martini, and Mattia Cecchinato. Evolution of infectious bronchitis virus in the field after homologous vaccination introduction. *Veterinary Research*, 50(1):92, 2019.
- [FM83] Douglas Scott Falconer and Trudy FC Mackay. *Quantitative genetics*. Longman London, 1983.
- [FWK10] Jasper Franke, Gregor Wergen, and Joachim Krug. Records and sequences of records from random variables with a linear trend. *Journal of Statistical Mechanics: Theory and Experiment*, 2010(10):P10013, 2010.
- [FWK12] J Franke, G Wergen, and J Krug. Correlations of record events as a test for heavy-tailed distributions. *Physical review letters*, 108(6):064101, 2012.
- [G⁺04] B. T. Grenfell et al. Unifying the epidemiological and evolutionary dynamics of pathogens. *Science*, 303:327–332, 2004.
- [Gav04] Sergey Gavrilets. *Fitness landscapes and the origin of species*, volume 41. Princeton university press, 2004.
- [GB08] K-I Goh and A-L Barabási. Burstiness and memory in complex systems. *Europhysics Letters*, 81(4):48002, 2008.
- [GBLD⁺21] Wilfredo F. Garcia-Beltran, Evan C. Lam, Kerri St. Denis, Adam D. Nitido, Zeidy H. Garcia, Blake M. Hauser, Jared Feldman, et al. Multiple sars-cov-2 variants escape neutralization by vaccine-induced humoral immunity. *Cell*, 184(9):2372–2383.e9, 2021.
- [GD08] Sylvain Gandon and Troy Day. Evidences of parasite evolution after vaccination. *Vaccine*, 26(Suppl 3):C4–C7, 2008.
- [Gib22] Amanda Kyle Gibson. Genetic diversity and disease: The past, present, and future of an old idea. *Evolution*, 76(S1):20–36, 2022.
- [Gil84] John H Gillespie. Molecular evolution over the mutational landscape. *Evolution*, pages 1116–1129, 1984.
- [Gil01] D. T. Gillespie. Approximate accelerated stochastic simulation of chemically reacting systems. *Journal of Chemical Physics*, 115:1716–1733, 2001.
- [Gil04a] J. H. Gillespie. *Population Genetics: A Concise Guide*. Johns Hopkins University Press, 2004.

- [Gil04b] John H Gillespie. *Population genetics: a concise guide*. JHU press, 2004.
- [Gli78] Ned Glick. Breaking records and breaking boards. *The American Mathematical Monthly*, 85(1):2–26, 1978.
- [GM02] S. Gandon and Y. Michalakis. Local adaptation, evolutionary potential and host-parasite coevolution: Interactions between migration, mutation, population size and generation time. *Journal of Evolutionary Biology*, 15(3):451–462, 2002.
- [GMNR01] S. Gandon, M. J. Mackinnon, S. Nee, and A. F. Read. Imperfect vaccines and the evolution of pathogen virulence. *Nature*, 414(6865):751–756, 2001.
- [GMNR03] Sylvain Gandon, Margaret Mackinnon, Sean Nee, and Andrew Read. Imperfect vaccination: Some epidemiological and evolutionary consequences. *Proceedings. Biological Sciences / the Royal Society*, 270(1520):1129–1136, 2003.
- [Gra20] B. S. Graham. Rapid covid-19 vaccine development. *Science*, 368:945–946, 2020.
- [GSSK15] Ard Jan Grimbergen, Jeroen Siebring, Ana Solopova, and Oscar P. Kuipers. Microbial bet-hedging: The power of being different. *Current Opinion in Microbiology*, 25:67–72, 2015.
- [H⁺20] K. B. Habersaat et al. Ten considerations for effectively managing the covid-19 transition. *Nature Human Behaviour*, 4:677–687, 2020.
- [Hal27] J. B. S. Haldane. A mathematical theory of natural and artificial selection, part v: Selection and mutation. *Mathematical Proceedings of the Cambridge Philosophical Society*, 23:838–844, 1927.
- [HC06] Daniel L. Hartl and Andrew G. Clark. *Principles of Population Genetics*. Sinauer Associates (Oxford University Press imprint), Sunderland, MA, 4 edition, 2006.
- [HDL17] Colin Havenar-Daughton, Jeong Hyun Lee, and Shane Crotty. Tfh cells and hiv bnabs, an immunodominance model of the hiv neutralizing antibody generation problem. *Immunological Reviews*, 275(1):49–61, 2017.
- [Hea78] Harold Stanley Heaps. *Information retrieval: Computational and theoretical aspects*. Academic Press, Inc., 1978.
- [hee18] heeso037. 120 years of olympic history: Athletes and results. Kaggle dataset, June 2018. Version published June 15, 2018.
- [HHRM11] Sarah E. Haas, Mevin B. Hooten, David M. Rizzo, and Ross K. Meentemeyer. Forest species diversity reduces disease risk in a generalist plant pathogen invasion. *Ecology Letters*, 14(11):1108–1116, 2011.
- [HK13] A. Huppert and G. Katriel. Mathematical modelling and prediction in infectious disease epidemiology. *Clinical Microbiology and Infection*, 19:999–1005, 2013.
- [HK20] J. Hilton and M. J. Keeling. Estimation of country-level basic reproductive ratios for novel coronavirus (sars-cov-2/covid-19) using synthetic contact matrices. *PLoS Computational Biology*, 16:e1008031, 2020.

- [Hop84] Fred M Hoppe. Pólya-like urns and the ewens' sampling formula. *Journal of mathematical biology*, 20(1):91–94, 1984.
- [HSF⁺19] Jue Hou, Shubham Shrivastava, Christopher C. Fraser, Hooi Linn Loo, Lan Hiong Wong, Victor Ho, Katja Fink, Eng Eong Ooi, and Jianzhu Chen. Dengue mosaic vaccines enhance cellular immunity and expand the breadth of neutralizing antibody against all four serotypes of dengue viruses in mice. *Frontiers in Immunology*, 10:1429, 2019.
- [HYC⁺22] Wan-Ting He, Meng Yuan, Sean Callaghan, Rami Musharrafieh, Ge Song, Murillo Silva, Nathan Beutler, et al. Broadly neutralizing antibodies to sars-related viruses can be readily induced in rhesus macaques. *Science Translational Medicine*, 14(657):eabl9605, 2022.
- [JK05] Kavita Jain and Joachim Krug. Evolutionary trajectories in rugged fitness landscapes. *Journal of Statistical Mechanics: Theory and Experiment*, 2005(04):P04008, 2005.
- [K⁺18] P. K. Kiyuka et al. Human coronavirus nl63 molecular epidemiology and evolutionary patterns in rural coastal kenya. *Journal of Infectious Diseases*, 217:1728–1739, 2018.
- [Kau93] Stuart A. Kauffman. *The Origins of Order: Self-Organization and Selection in Evolution*. Oxford University Press, New York, 1993.
- [Kau00] Stuart A Kauffman. *Investigations*. Oxford University Press, 2000.
- [KCD⁺21] Steven A. Kemp, Dami A. Collier, Rawlings P. Datir, Isabella A. T. M. Ferreira, Salma Gayed, Aminu Jahun, Myra Hosmillo, et al. Sars-cov-2 evolution during treatment of chronic infection. *Nature*, 592(7853):277–282, 2021.
- [KCGP06] Katia Koelle, Sarah Cobey, Bryan Grenfell, and Mercedes Pascual. Epochal evolution shapes the phylodynamics of interpandemic influenza a (h3n2). *Science*, 314(5807):1898–1903, 2006.
- [KFM⁺21] Stephen M. Kissler, Joseph R. Fauver, Christina Mack, Caroline G. Tai, Mallery I. Breban, Anne E. Watkins, Radhika M. Samant, et al. Viral dynamics of sars-cov-2 variants in vaccinated and unvaccinated persons. *The New England Journal of Medicine*, 385(26):2489–2491, 2021.
- [Kha20] B. S. Khatri. Stochastic extinction of epidemics: how long would it take for sars-cov-2 to die out without herd immunity? *medRxiv*, 2020.
- [Kin82] John Frank Charles Kingman. The coalescent. *Stochastic processes and their applications*, 13(3):235–248, 1982.
- [KJ05] Joachim Krug and Kavita Jain. Breaking records in the evolutionary race. *Physica A: Statistical Mechanics and its Applications*, 358(1):1–9, 2005.
- [KL87] Stuart A. Kauffman and Simon Levin. Towards a general theory of adaptive walks on rugged landscapes. *Journal of Theoretical Biology*, 128(1):11–45, 1987.

- [KL12] K. C. King and C. M. Lively. Does genetic diversity limit disease spread in natural host populations? *Heredity*, 109(4):199–203, 2012.
- [KNB⁺20] Tamalika Kar, Utkarsh Narsaria, Srijita Basak, Debashrito Deb, Filippo Castiglione, David M. Mueller, and Anurag P. Srivastava. A candidate multi-epitope vaccine against sars-cov-2. *Scientific Reports*, 10(1):10895, 2020.
- [KR08] Matt J Keeling and Pejman Rohani. *Modeling infectious diseases in humans and animals*. Princeton university press, 2008.
- [KR17] David A. Kennedy and Andrew F. Read. Why does drug resistance readily evolve but vaccine resistance does not? *Proceedings. Biological Sciences / the Royal Society*, 284(1851), 2017.
- [KTG⁺20] S. M. Kissler, C. Tedijanto, E. Goldstein, Y. H. Grad, and M. Lipsitch. Projecting the transmission dynamics of sars-cov-2 through the postpandemic period. *Science*, 368:860–868, 2020.
- [KTP09] Sergey Kryazhimskiy, Gašper Tkačik, and Joshua B Plotkin. The dynamics of adaptation on correlated fitness landscapes. *Proceedings of the National Academy of Sciences*, 106(44):18638–18643, 2009.
- [L⁺20] S. A. Lauer et al. The incubation period of coronavirus disease 2019 (covid-19) from publicly reported confirmed cases: Estimation and application. *Annals of Internal Medicine*, 172:577–582, 2020.
- [L⁺21] D. Y. Logunov et al. Safety and efficacy of an rad26 and rad5 vector-based heterologous prime-boost covid-19 vaccine: An interim analysis of a randomised controlled phase 3 trial in russia. *Lancet*, 2021.
- [LD00] C. M. Lively and M. F. Dybdahl. Parasite adaptation to locally common host genotypes. *Nature*, 405(6787):679–681, 2000.
- [LHN05] Erez Lieberman, Christoph Hauert, and Martin A. Nowak. Evolutionary dynamics on graphs. *Nature*, 433(7023):312–316, 2005.
- [LKRY⁺22] Marc Lipsitch, Florian Krammer, Gili Regev-Yochay, Yaniv Lustig, and Ran D. Balicer. Sars-cov-2 breakthrough infections in vaccinated individuals: Measurement, causes and impact. *Nature Reviews Immunology*, 22(1):57–65, 2022.
- [LKW24] Edward D Lee, Christopher P Kempes, and Geoffrey B West. Idea engines: Unifying innovation & obsolescence from markets & genetic evolution to science. *Proceedings of the National Academy of Sciences*, 121(6):e2312468120, 2024.
- [Luo19] T. H. Luong. *Mathematical Modeling of Vaccinations: Modified SIR Model, Vaccination Effects, and Herd Immunity*. PhD thesis, Portland State University, 2019.
- [LWC⁺22] Ainsley Ryan Yan Bin Lee, Shi Yin Wong, Louis Yi Ann Chai, Soo Chin Lee, Matilda Xinwei Lee, Mark Dhinesh Muthiah, Sen Hee Tay, Chong Boon Teo, Benjamin Kye Jyn Tan, Yiong Huak Chan, et al. Efficacy of covid-19 vaccines in immunocompromised patients: systematic review and meta-analysis. *Bmj*, 376, 2022.

- [M⁺20a] M. Makhoul et al. Epidemiological impact of sars-cov-2 vaccination: Mathematical modeling analyses. *Vaccines (Basel)*, 8:668, 2020.
- [M⁺20b] B. Malone et al. Artificial intelligence predicts the immunogenic landscape of sars-cov-2 leading to universal blueprints for vaccine designs. *Scientific Reports*, 10:22375, 2020.
- [Mar52] Harry Markowitz. Portfolio selection. *The Journal of Finance*, 7(1):77, 1952.
- [Mar18] M. E. Martinez. The calendar of epidemics: Seasonal cycles of infectious diseases. *PLoS Pathogens*, 14:e1007327, 2018.
- [MBM⁺21] Matthew McCallum, Jessica Bassi, Anna De Marco, Alex Chen, Alexandra C. Walls, Julia Di Iulio, M. Alejandra Tortorici, et al. Sars-cov-2 immune evasion by the b.1.427/b.1.429 variant of concern. *Science*, 373(6555):648–654, 2021.
- [MBVN20] P. Mecnas, R. T. R. M. Bastos, A. C. R. Vallinoto, and D. Normando. Effects of temperature and humidity on the spread of covid-19: A systematic review. *PLoS ONE*, 15:e0238339, 2020.
- [Mes16] Alex Mesoudi. Cultural evolution: a review of theory, findings and controversies. *Evolutionary biology*, 43(4):481–497, 2016.
- [MKPP21] Marco Marani, Gabriel G. Katul, William K. Pan, and Anthony J. Parolari. Intensity and frequency of extreme novel epidemics. *Proceedings of the National Academy of Sciences*, 118(35):e2105482118, 2021.
- [MMB15] A. Mikaberidze, B. A. McDonald, and S. Bonhoeffer. Developing smarter host mixtures to control plant disease. *Plant Pathology*, 64(4):996–1004, 2015.
- [MPS20] Satya N Majumdar, Arnab Pal, and Grégory Schehr. Extreme value statistics of correlated random variables: a pedagogical review. *Physics Reports*, 840:1–32, 2020.
- [MS24] Satya N Majumdar and Gregory Schehr. *Statistics of Extremes and Records in Random Sequences*. Oxford University Press, 2024.
- [MSW12] Satya N Majumdar, Grégory Schehr, and Gregor Wergen. Record statistics and persistence for a random walk with a drift. *Journal of Physics A: Mathematical and Theoretical*, 45(35):355002, 2012.
- [MWM21] David V. McLeod, Lindi M. Wahl, and Nicole Mideo. Mosaic vaccination: How distributing different vaccines across a population could improve epidemic control. *Evolution Letters*, 5(5):458–471, 2021.
- [Nev00] Valéry Nevzorov. Records: mathematical theory. *Translations of mathematical monographs*, 2000.
- [NFBT13] Béla Nagy, J Doyne Farmer, Quan M Bui, and Jessika E Trancik. Statistical basis for predicting technological progress. *PloS one*, 8(2):e52669, 2013.
- [NM00] M. Nowak and R. M. May. Chapter 3–4. In *Virus Dynamics: Mathematical Principles of Immunology and Virology*, page 16. 2000.

- [NO01] Kevin Norton and Tim Olds. Morphological evolution of athletes over the 20th century: causes and consequences. *Sports medicine*, 31(11):763–783, 2001.
- [Nor10] Dennis Normile. Rinderpest, deadly for cattle, joins smallpox as a vanquished disease. *Science*, 330(6003):435–435, 2010.
- [NRS14] Richard A. Neher, Christy A. Russell, and Boris I. Shraiman. Predicting evolution from the shape of genealogical trees. *eLife*, 3:e03568, 2014.
- [olynd] Olympian database. Online database entry, n.d. Accessed: 23 August 2025.
- [OR18] Sophie Ochmann and Max Roser. Smallpox - our world in data. Our World in Data, June 2018.
- [Orr02] H Allen Orr. The population genetics of adaptation: the adaptation of dna sequences. *Evolution*, 56(7):1317–1330, 2002.
- [OU08] H. A. Orr and R. L. Unckless. Population extinction and the genetics of adaptation. *American Naturalist*, 172:160–169, 2008.
- [Our21] Our World in Data. Coronavirus (covid-19) vaccinations. <https://ourworldindata.org/covid-vaccinations>, 2021.
- [OW97] S. P. Otto and M. C. Whitlock. The probability of fixation in populations of changing size. *Genetics*, 146:723–733, 1997.
- [P⁺20] F. P. Polack et al. Safety and efficacy of the bnt162b2 mrna covid-19 vaccine. *New England Journal of Medicine*, 383:2603–2615, 2020.
- [Pit95] Jim Pitman. Exchangeable and partially exchangeable random partitions. *Probability theory and related fields*, 102(2):145–158, 1995.
- [PLAG⁺15] Marcos Pérez-Losada, Miguel Arenas, Juan Carlos Galán, Ferran Palero, and Fernando González-Candelas. Recombination in viruses: Mechanisms, methods of study, and evolutionary consequences. *Infection, Genetics and Evolution*, 30:296–307, 2015.
- [Plo15] Randy C. Ploetz. Fusarium wilt of banana. *Phytopathology*, 105(12):1512–1521, 2015.
- [PR18] V. N. Petrova and C. A. Russell. The evolution of seasonal influenza viruses. *Nature Reviews Microbiology*, 16:47–60, 2018.
- [PSCVMV15] Romualdo Pastor-Satorras, Claudio Castellano, Piet Van Mieghem, and Alessandro Vespignani. Epidemic processes in complex networks. *Reviews of modern physics*, 87(3):925–979, 2015.
- [PSM⁺22] Delphine Planas, Nell Saunders, Piet Maes, Florence Guivel-Benhassine, Cyril Planchais, Julian Buchrieser, William-Henry Bolland, et al. Considerable escape of sars-cov-2 omicron to antibody neutralization. *Nature*, 602(7898):671–675, 2022.
- [PW08] Z. Patwa and L. M. Wahl. The fixation probability of beneficial mutations. *Journal of the Royal Society Interface*, 5(28):1279–1289, 2008.

- [PY97] Jim Pitman and Marc Yor. The two-parameter poisson-dirichlet distribution derived from a stable subordinator. *The Annals of Probability*, pages 855–900, 1997.
- [R⁺21] E. Rosenstrom et al. High-quality masks reduce covid-19 infections and deaths in the us. *medRxiv*, 2021.
- [REX13] REX Consortium. Heterogeneity of selection and the evolution of resistance. *Trends in Ecology & Evolution*, 28(2):110–118, 2013.
- [REX16] REX Consortium. Combining selective pressures to enhance the durability of disease resistance genes. *Frontiers in Plant Science*, 7:1916, 2016.
- [RKDK21] Simon A Rella, Yuliya A Kulikova, Emmanouil T Dermitzakis, and Fyodor A Kondrashov. Rates of sars-cov-2 transmission and vaccination impact the fate of vaccine-resistant strains. *Scientific Reports*, 11(1):15729, 2021.
- [RKMK24] Simon A Rella, Yuliya A Kulikova, Aygul R Minnegalieva, and Fyodor A Kondrashov. Complex vaccination strategies prevent the emergence of vaccine resistance. *Evolution*, 78(10):1722–1738, 2024.
- [RPB⁺18] Loup Rimbaud, Julien Papaix, Luke G. Barrett, Jeremy J. Burdon, and Peter H. Thrall. Mosaics, mixtures, rotations or pyramiding: What is the optimal strategy to deploy major gene resistance? *Evolutionary Applications*, 11(10):1791–1810, 2018.
- [RSA16] D. Rauff, C. Strydom, and C. Abolnik. Evolutionary consequences of a decade of vaccination against subtype h6n2 influenza. *Virology*, 498:226–239, 2016.
- [RSL⁺22] Magdalena L. Russell, Aisha Souquette, David M. Levine, Stefan A. Schattgen, E. Kaitlynn Allen, Guillermina Kuan, Noah Simon, et al. Combining genotypes and t cell receptor distributions to infer genetic loci determining v(d)j recombination probabilities. *eLife*, 11, March 2022.
- [RT19] Delphine Renard and David Tilman. National food production stabilized by crop diversity. *Nature*, 571(7764):257–260, 2019.
- [RWK21] N. Rochman, Y. Wolf, and E. Koonin. Substantial impact of post-vaccination contacts on cumulative infections during viral epidemics. *F1000Research*, 10:315, 2021.
- [Sch13] Joseph A Schumpeter. *Capitalism, socialism and democracy*. routledge, 2013.
- [SGH⁺20] Tyler N. Starr, Allison J. Greaney, Sarah K. Hilton, Daniel Ellis, Katharine H. D. Crawford, Adam S. Dingens, Mary Jane Navarro, et al. Deep mutational scanning of sars-cov-2 receptor binding domain reveals constraints on folding and ace2 binding. *Cell*, 182(5):1295–1310.e20, 2020.
- [Sim91] Karl Sims. Artificial evolution for computer graphics. In *Proceedings of the 18th annual conference on Computer graphics and interactive techniques*, pages 319–328, 1991.

- [Sim11] Andrew M. Simons. Modes of response to environmental change and the elusive empirical evidence for bet hedging. *Proceedings. Biological Sciences / the Royal Society*, 278(1712):1601–1609, 2011.
- [SJ13] Paolo Sibani and Henrik Jeldtoft Jensen. *Stochastic dynamics of complex systems: From glasses to evolution*. World Scientific, 2013.
- [SKK⁺21] Sandrine M. Soh, Yeongjun Kim, Chanwoo Kim, Ui Soon Jang, and Hye-Ra Lee. The rapid adaptation of sars-cov-2-rise of the variants: Transmission and resistance. *Journal of Microbiology*, 59(9):807–818, 2021.
- [Sno26] George S Snoddy. Learning and stability: a psychophysiological analysis of a case of motor learning with clinical applications. *Journal of Applied Psychology*, 10(1):1, 1926.
- [SO01] Kenneth A. Schmidt and Richard S. Ostfeld. Biodiversity and the dilution effect in disease ecology. *Ecology*, 2001. March.
- [Som05] Simone Sommer. The importance of immune gene variability (mhc) in evolutionary ecology and conservation. *Frontiers in Zoology*, 2:16, October 2005.
- [SPS⁺22] Sissy Therese Sonnleitner, Martina Prelog, Stefanie Sonnleitner, Eva Hinterbichler, Hannah Halbfurter, Dominik B. C. Kopecky, Giovanni Almanzar, et al. Cumulative sars-cov-2 mutations and corresponding changes in immunity in an immunocompromised patient indicate viral evolution within the host. *Nature Communications*, 13(1):2560, 2022.
- [SSK⁺22] Yana Safonova, Sung Bong Shin, Luke Kramer, James Reecy, Corey T. Watson, Timothy P. L. Smith, and Pavel A. Pevzner. Variations in antibody repertoires correlate with vaccine responses. *Genome Research*, 32(4):791–804, 2022.
- [Suh97] A. Suhrbier. Multi-epitope dna vaccines. *Immunology and Cell Biology*, 75(4):402–408, 1997.
- [T⁺21] H. Tegally et al. Sixteen novel lineages of sars-cov-2 in south africa. *Nature Medicine*, 2021.
- [The19] The Lancet. Measles eradication: A goal within reach, slipping away. *The Lancet*, 393(10182):1669, 2019.
- [TLS18] Francesca Tria, Vittorio Loreto, and Vito DP Servedio. Zipf’s, heaps’s and taylor’s laws are determined by the expansion into the adjacent possible. *Entropy*, 20(10):752, 2018.
- [TLSS14] Francesca Tria, Vittorio Loreto, Vito Domenico Pietro Servedio, and Steven H Strogatz. The dynamics of correlated novelties. *Scientific reports*, 4(1):5890, 2014.
- [Tor80] James Torbitt. Letter from j. torbitt ‘letter no. 12728’. Darwin Correspondence Project, September 1880.
- [TRLT24] Giulio Tani Raffaelli, Margherita Lalli, and Francesca Tria. Inference through innovation processes tested in the authorship attribution task. *Communications Physics*, 7(1):300, 2024.

- [UH11] H. Uecker and J. Hermisson. On the fixation process of a beneficial mutation in a variable environment. *Genetics*, 188:915–930, 2011.
- [UKS⁺10] Line V. Ugelvig, Daniel J. C. Kronauer, Alexandra Schrempf, Jürgen Heinze, and Sylvia Cremer. Rapid anti-pathogen response in ant societies relies on high genetic diversity. *Proceedings. Biological Sciences / the Royal Society*, 277(1695):2821–2828, 2010.
- [Uni22] United Nations, Department of Economic and Social Affairs, Population Division. *World Population Prospects 2022: Summary of Results*. United Nations, New York, 2022.
- [VB⁺20] J. J. Van Bavel et al. Using social and behavioural science to support covid-19 pandemic response. *Nature Human Behaviour*, 4:460–471, 2020.
- [VD18] E. M. Volz and X. Didelot. Modeling the growth and decline of pathogen effective population size provides insight into epidemic dynamics and drivers of antimicrobial resistance. *Systematic Biology*, 67:719–728, 2018.
- [vHEB⁺16] Stineke van Houte, Alice K. E. Ekroth, Jenny M. Broniewski, Hélène Chabas, Ben Ashby, Joseph Bondy-Denomy, Sylvain Gandon, et al. The diversity-generating benefits of a prokaryotic adaptive immune system. *Nature*, 532(7599):385–388, 2016.
- [WB21] T. C. Williams and W. A. Burgers. Sars-cov-2 evolution and vaccines: Cause for concern? *Lancet Respiratory Medicine*, 2021.
- [Wer12] Gregor Wergen. *Records statistics beyond the standard model-Theory and applications*. PhD thesis, Universität zu Köln, 2012.
- [Wer13] Gregor Wergen. Records in stochastic processes—Theory and applications. *Journal of Physics A: Mathematical and Theoretical*, 46(22):223001, 2013.
- [WFK11] Gregor Wergen, Jasper Franke, and Joachim Krug. Correlations between record events in sequences of random variables with a linear trend. *Journal of Statistical Physics*, 144(6):1206, 2011.
- [WGM⁺22] Brian J. Willett, Joe Grove, Oscar A. MacLean, Craig Wilkie, Giuditta De Lorenzo, Wilhelm Furnon, Diego Cantoni, et al. Sars-cov-2 omicron is an immune escape variant with an altered cell entry pathway. *Nature Microbiology*, 7(8):1161–1179, 2022.
- [WHK14] Gregor Wergen, Andreas Hense, and Joachim Krug. Record occurrence and record values in daily and monthly temperatures. *Climate dynamics*, 42(5):1275–1289, 2014.
- [WK10] Gregor Wergen and Joachim Krug. Record-breaking temperatures reveal a warming climate. *Europhysics Letters*, 92(3):30008, 2010.
- [WLZ⁺22] Yifan Wang, Caixuan Liu, Chao Zhang, Yanxing Wang, Qin Hong, Shiqi Xu, Zuyang Li, Yong Yang, Zhong Huang, and Yao Cong. Structural basis for sars-cov-2 delta variant recognition of ace2 receptor and broadly neutralizing antibodies. *Nature Communications*, 13(1):871, 2022.

- [WMC17] F. T. Wen, A. Malani, and S. Cobey. The beneficial effects of vaccination on the evolution of seasonal influenza. *bioRxiv*, 2017.
- [WMRA20] R. West, S. Michie, G. J. Rubin, and R. Aml  tt. Applying principles of behaviour change to reduce sars-cov-2 transmission. *Nature Human Behaviour*, 4:451–459, 2020.
- [wornd] World athletics. Official website, n.d. Accessed: 23 August 2025.
- [WPY⁺21] Yan-Ping Wang, Zhe-Chao Pan, Li-Na Yang, Jeremy J. Burdon, Hanna Friberg, Qi-Jun Sui, and Jiasui Zhan. Optimizing plant disease management in agricultural ecosystems through rational in-crop diversification. *Frontiers in Plant Science*, 12:767209, December 2021.
- [Wri32] Sewall Wright. The roles of mutation, inbreeding, crossbreeding, and selection in evolution. In *Proceedings of the Sixth International Congress of Genetics*, volume 1, pages 356–366, 1932.
- [XM13] Y. Xiao and S. M. Moghadas. Impact of viral drift on vaccination dynamics and patterns of seasonal influenza. *BMC Infectious Diseases*, 13:589, 2013.
- [Yad16] Avinash Chand Yadav. Survival-time statistics for sample space reducing stochastic processes. *Physical Review E*, 93(4):042131, 2016.
- [Yad17] Avinash Chand Yadav. Correspondence between a noisy sample-space-reducing process and records in correlated random events. *Physical Review E*, 96(3):032134, 2017.
- [Yan75] Mark CK Yang. On the distribution of the inter-record times in an increasing population. *Journal of Applied Probability*, 12(1):148–154, 1975.
- [YS21] Jonathan W. Yewdell and Jefferson J. S. Santos. Original antigenic sin: How original? how sinful? *Cold Spring Harbor Perspectives in Medicine*, 11(5), 2021.
- [ZHL⁺22] Jing Zhang, Zi Bo Han, Yu Liang, Xue Feng Zhang, Yu Qin Jin, Li Fang Du, Shuai Shao, et al. A mosaic-type trimeric rbd-based covid-19 vaccine candidate induces potent neutralization against omicron and other sars-cov-2 variants. *eLife*, 11, August 2022.
- [Zip16] George Kingsley Zipf. *Human behavior and the principle of least effort: An introduction to human ecology*. Ravenio books, 2016.
- [ZRZ⁺22] Xiyun Zhang, Zhongyuan Ruan, Muhua Zheng, Jie Zhou, Stefano Boccaletti, and Baruch Barzel. Epidemic spreading under mutually independent intra- and inter-host pathogen evolution. *Nature Communications*, 13(1):6218, 2022.

