THE UNIVERSITY OF CHICAGO

PREDICTIVE STRATEGIES IN TIME-VARYING ENVIRONMENTS

A DISSERTATION SUBMITTED TO
THE FACULTY OF THE DIVISION OF THE PHYSICAL SCIENCES
AND
THE FACULTY OF THE DIVISION OF BIOLOGICAL SCIENCES AND THE
PRITZKER SCHOOL OF MEDICINE
IN CANDIDACY FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

GRADUATE PROGRAM IN BIOPHYSICAL SCIENCES

BY
VEDANT SACHDEVA

CHICAGO, ILLINOIS
AUGUST 2022

To thirteen, eighteen, and twenty-two year old me. Hope I made you proud.

"Finally, from so little sleeping and so much reading, his brain dried up and he went completely out of his mind." - Miguel de Cervantes Saavedra, Don Quixote

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

First, and foremost, I'd like to thank Professors Stephanie Palmer and Arvind Murugan. From even before I stepped on to the University of Chicago's campus, both of you saw promise in me and I hope I was able to live up to a fraction of that. From each of you, I've benefited from an endless supply of problems, ranging from statistical puzzles to difficult science problems. I've also gained a deep appreciation for simplicity, both in terms of the actual day-to-day work of science and in communication of scientific results. If I'm lucky, my state after graduate school will begin to resemble some superposition of yours.

To Professors Ranganathan and Littlewood, I thank you for the exciting ideas and thoughts you've shared with me throughout the various discussions you've had. Professor Ranganathan, the discussions we've had about how to develop theory in biology will be a driving force in my career going forward. Professor Littlewood, the connection between statistical physics and neuroscience you draw are inspiring, and have forced me to think hard about how I can connect what I've learned in the classroom to what I do in my research.

Over the course of the past 5 years, I've had the fortune of collaborating with many talented scientists: Professors Walczak, Mora, and Wang, Dr. Husain, and Dr. Ravasio, each of whom I've tried to emulate in my pursuit of science. Through their ingenuity, I've had the luxury of working on exciting problems and it's made my PhD successful. Dr. Husain, in particular, your ability to hear an idea, think about it carefully, and within a few hours, have a full theory behind it worked out is inspiring.

I've also had the benefit of having the support of my family, my friends, and my hobbies. From them, I learned much about intensity, humility, and discipline. While I need to continue to develop in each of these qualities, with their help, I've come a long way from where I was, and it's made me a better person.

ABSTRACT

Adapting populations have often been characterized in the context of static environments. However, many natural environments vary on timescales similar to or faster than the rate of evolutionary adaptation. Examples of time-varying environments include the environment faced by the adaptive immune system when generating antibodies against highly mutagenic viruses or the environment faced by an organism attempting to escape from an incoming predator.

We begin by exploring adaptation against HIV, a virus that mutates on the same timescale as it takes for B-cells to evolve antibodies. In our study, we propose a conceptual framework in which there exists generalist and specialist phenotypes. Specialist phenotypes are strategies that confer high fitness in a given environment, and may not confer high fitness in any other environment. Generalist phenotypes, on the other hand, are 'jack-of-all-trades' strategies and work well across a family of environments, though they may not work as well as a specialist phenotype for a given environment. We are able to demonstrate that the preferred phenotype depends on the variation of the environmental landscape. Further, we identify that generalist phenotypes confer fitness by exploiting the correlation structure of the time-varying environment.

We then consider the sensory encoding scheme used by the retina against changing visual scenes. Unlike in the previous study, the natural environment changes much more rapidly than the evolutionary adaptation time of the sensory encoding scheme. Consequently, we explore how a sensory encoding scheme can be predictive for a fixed level of compression. Using the information bottleneck method, we explore the optimal sensory encoding schemes for a range of time-varying environments relevant to the visual system. In addition, we also explore the transferability of a sensory encoding scheme, identifying the best schemes when the autocorrelation structure of the environment itself varies.

CHAPTER 1

INTRODUCTION

There are many environmental threats that can challenge the survival of a given species. These threats can range from infection by foreign antigens to predation by other species. In order to survive against such threats, biological organisms undergo evolution. In evolution, organisms reproduce and mutate over the course of many generations. The rate of reproduction, often called fitness, depends on both the external environment and the expressed phenotype of each organism. During reproduction, organisms pass on their phenotypes to the next generation, unless a mutation occurs during reproduction. Mutations cause offspring to develop different phenotypes from their parents. This can be beneficial because it enables individual organisms in a population to adapt to an environmental threat. In addition, the organisms that initially acquired the beneficial mutations will proliferate more rapidly than their peers, eventually resulting in the population having adapted to this environmental threat.

Evolution can be studied through the Wright-Fisher model(180). In the bi-allelic Wright-Fisher model, we suppose there exist two alleles of a given gene, a' and a. We assume a' has a selective advantage over a - that is, individuals with the allele a' reproduce faster than individuals with allele a. Assuming no mutation, the probability that k individuals in a population of size N have allele a', assuming that k' individuals had allele a' in the last generation is given by

$$P(N_{a'} = k) = \binom{N}{k} (\frac{k'(1+s)}{N+k's})^k (\frac{N-k'}{N+k's})^{N-k}. \tag{1.1}$$

Here, the selective advantage is of size s. This distribution can also be extended to the continuous distribution and to include mutations:

$$\frac{df}{dt} = sf(1-f) + \mu(1-f) - \nu f + \sqrt{\frac{f(1-f)}{N}}\eta(t). \tag{1.2}$$

Here, f represents the frequency of allele a', μ represents the transition rate from allele a to a', ν represents the transition rate from a' to a, and $\eta(t)$ is Gaussian white noise.

The Wright-Fisher distribution can be used to study both real data and provide testable predictions about the trajectories of a species. For example, the Wright-Fisher distribution has been used to compute the minimum population size at which a beneficial mutation escapes the stochastic effects of frequency-dependent reproduction(73). In addition, the Wright Fisher model can be used to solve inverse problems, and identify the benefit of a mutation at a particular gene using real data(157). However, a core assumption of this work is that the benefit of a mutation is static compared to the timescale it takes for a mutation to fix or go extinct in a population.

Relaxing this assumption, however, yields exciting results. Some theoretical works have shown that if selection pressure fluctuates periodically, phenotypes that are entropically disfavored may become the predominant phenotype in a population(79). In addition, by connecting the Wright-Fisher diffusion approximation to the Fokker-Planck equation, it can be shown that the dynamics of a population in a fluctuating selection pressure resembles that of a heteropolymer in an external field. This connection, among many others, enables existing insights generated about physical systems to be applied to population genetics. Further, it has been shown that the effect of a mutation in a time-varying environment is not simply given by its average fitness effect across all environments(40). In fact, even deleterious-on-average mutations can sometimes fix if the environmental conditions are sufficient.

Some of these effects have been observed in natural settings. For example, the evolution of broadly neutralizing antibodies, typically disfavored for both entropic and efficacy reasons, against fast mutating viruses such as HIV has been observed in clinical data(84; 53; 25). Further experiments and computational studies demonstrated that in fact, static selection pressures fail to yield broadly neutralizing antibodies. Instead, by cycling through several strains of a virus, broadly neutralizing antibodies can be elicited(173).

Inspired by this work, in Chapter 2, we set out to engineer a time-varying landscape which robustly drives populations of antibodies towards being broadly neutralizing against a full family of viruses, rather than just one particular strain of a virus. We observe that both the entropic and energetic costs of being a broadly neutralizing antibody presents two disjoint challenges: driving a population towards a broadly neutralizing antibody and then stabilizing it in the broadly neutralizing antibody state. We demonstrate ways of navigating this tension, and eventually propose a robust time-varying landscape which could be translated into a vaccination protocol. This work also appears in (141).

We can also explore population adaptations through the lens of the efficient coding hypothesis(11). In the efficient coding hypothesis, it is assumed that over evolutionary timescales, sensory encoding schemes have evolved such that their response to an external stimuli is maximally informative of a relevant statistic subject to constraints originating from internal noise and resource constraints. A core prediction of the efficient coding hypothesis is that sensory encoding schemes are evolved to match the statistics of their natural environment. Evidence for this prediction has been observed in the visual system, where early visual processing systems are shown to reduce generalized redundancy(9). In general, however, a core challenge has been identifying the relevant statistic for a biological system. Several paradigms have been proposed, such as sparsity and predictiveness(29). For this thesis, we will consider prediction to be the primary relevant statistic.

The need to be predictive emerges from the delay between information being transmitted to the brain by sensory organs and the ability to then cue a motor response in reaction to the received information(18). Without prediction, some organisms may be unable to escape predation. In order to identify if a sensory organ is efficient for prediction, we utilize the information bottleneck method(158). In the information bottleneck method, we imagine there exists a system with coordinates given by the vector X evolving in time t. This system represents the external environment. Sensory organs respond to the dynamical system at

each time according to a sensory encoding scheme given as $p(z_t|x_t)$. Biological systems, over evolutionary timescales, have adapted their sensory encoding schemes to maximize the objective function:

$$\mathcal{L} = \max_{p(z_t|x_t)} I(X_{t+\Delta t}; Z_t) - \beta I(X_t; Z_t). \tag{1.3}$$

β represents a tradeoff parameter that encapsulates the effects of internal noise and resource constraints. Here, $I(A; B)$ is the mutual information between two random variables, A and B. It measures the reduction in uncertainty in one random variable caused by knowing the value of the other. It is computed as $\sum_{a\in A, b\in B} p(a, b) \log(\frac{p(a,b)}{p(a)p(b)})$.

Evidence for efficient predictive coding was observed in tiger salamander retina(125). In this experiment, a tiger salamander retina was placed on a microelectrode array and was exposed to a moving bar stimulus with predictable dynamics. The tiger salamander retina's firing pattern was recorded using the microelectrode array. The response was characterized by the neurons firing within some time window. The amount of predictive information in the retinal response was shown to be a maximum for the overall amount of information the retina could encode.

In Chapter 3, we present analytical predictions for the optimal sensory encoding schemes for a range of statistics expected in the visual world. This includes both stimuli with and without the heavy-tailed noise observed in the real world(139). Further, we propose the structure of an efficient predictive encoding for the adaptive immune system. We also propose a novel transferability metric that describes how well a particular sensory encoding scheme could be used for a stimulus with different underlying statistics. Using this transferability metric, we identify the optimal sensory encoding scheme if the visual scene's autocorrelation structure is rapidly varying. This work also appears in (142).

Not directly considered in either Chapters 2 or 3 is the question of how much data is needed to actually make predictions and how the inference method used impacts the amount of

data needed. While we do not explicitly treat how a biological system might make inferences, we consider some common inference methods that may be biologically plausible(103). Most available biological data exists in the strongly undersampled limit(74), and consequently, much care needs to be taken to avoid overfitting. In Chapter 4, we explore the impact of inference method on the number of samples needed to infer the generating distribution of a dataset. We compare global inference methods to local inference methods, and show that the efficacy of the method depends on whether the generator of a dataset is in an ordered phase or disordered phase. In particular, global inference methods are more effective for data generated in the ordered phase, as it is more sensitive to variations in the dataset, while local inference methods are better in the disordered phase. The work is presented as it appears in (116).

We extend upon the work in Chapter 3 in Chapter 5, focusing primarily on how an organism can make use of estimation memory to improve sensory encoding schemes. We argue that because biological systems are continuously making measurements of the external world throughout their lifetime, they may be able to use information from previous measurements to inform present estimates of the external world. The Bayes-optimal way of combining information from previous estimates of an external stimuli and a current measurement is given by the Kalman Filter(67). However, previous work on the Kalman Filter has assumed the measurement model, analogous to the sensory encoding scheme, is held fixed. However, biological systems are capable of changing their sensory encoding scheme, up to some resource constraints, to be more efficient, as predicted by the efficient coding hypothesis. We demonstrate that when sensory encoding schemes are optimized in this way, the biological organism decorrelates estimates based on previous measurements and estimates based on the current measurement. This decorrelation enables significant improvements in the ability of a biological system to estimate the stimulus with high precision.

CHAPTER 2

TUNING ENVIRONMENTAL TIMESCALES TO EVOLVE AND MAINTAIN GENERALISTS

This work was completed in collaboration with Kabir Husain, Jiming Sheng, Shenshen Wang, and Arvind Murugan.

2.1 Significance

Generalists, or jack-of-all-trades, that are fit across diverse environments can be difficult to evolve since they may not be as fit as a specialist in any particular environment. Such generalists are sought in immunology, where broadly neutralizing antibodies that can detect a broad variety of strains of a rapidly changing virus like HIV are often hard to evolve. Here we find that generalists are most easily evolved in the most poorly understood regime of evolution - where the environment changes are neither fast nor slow but on the same timescale as evolutionary response of the population. Our methods let us propose new temporal vaccination protocols, such as a chirp, that exploit this highly dynamic regime of evolution.

2.2 Abstract

Natural environments can present diverse challenges, but some genotypes remain fit across many environments. Such 'generalists' can be hard to evolve, out-competed by specialists fitter in any particular environment. Here, inspired by the search for broadly-neutralising antibodies during B-cell affinity maturation, we demonstrate that environmental changes on an intermediate timescale can reliably evolve generalists, even when faster or slower environmental changes are unable to do so. We find that changing environments on timescales comparable to evolutionary transients in a population enhances the rate of evolving generalists from specialists, without enhancing the reverse process. The yield of generalists is further

increased in more complex dynamic environments, such as a 'chirp' of increasing frequency. Our work offers design principles for how non-equilibrium fitness 'seascapes' can dynamically funnel populations to genotypes unobtainable in static environments.

2.3 Introduction

Evolutionary outcomes are driven by environmental pressures, but environments are rarely static(83). In a changing environment, some genotypes – termed generalists – maintain a uniformly high fitness over time, even if they are not globally fit at any particular instant. A striking example is that of broadly-neutralizing antibodies (bnAbs) against HIV and other viruses – these antibodies maintain potency against the large diversity of viral strains that may arise in an infected individual over time (27; 28; 184). It is desirable for the immune system to select for generalist antibodies during B-cell affinity maturation, a rapid evolutionary process(35), but generalists are often out-competed by specialists that only bind particular viral strains.

Recent work has suggested that sequential vaccination with different viral antigens, rather than a single cocktail of those antigens, can better select for generalist antibodies during affinity maturation (128; 89; 171; 173). This result is consistent with the broader idea that a time-varying environment can drive evolution out of equilibrium and into genotypes unevolvable in static environments (115; 114; 8; 78; 58). However, the broader principles underlying generalist selection by dynamic environments remain unknown. In particular, the interplay of environmental and evolutionary timescales and choices of correlated antigens generates a high-dimensional space of possible vaccination protocols. Hence guiding principles are needed to find optimal protocols for evolving generalist genotypes.

Here, we take a phenomenological approach to design dynamic environments that select generalists. We analyze situations in which generalists are entropically disfavoured or isolated by fitness valleys, and thus unevolvable in a static environment. We find that a dynamic

environmental protocol can maximize the yield of generalists if the environment changes on the same timescale as the evolutionary transients of the population, i.e., on the timescale for allele frequencies to reach steady state. Consequently, switching antigens before antibody populations have evolved to a steady state can dynamically funnel finite populations from specialists to generalists, even when faster or slower switching is unable to do so.

We understand these results in terms of a kinetic asymmetry between generalists and specialists. Environmental dynamics at the right timescale perturbs specialist populations while leaving generalists relatively undisturbed. This asymmetry favours evolution from specialists to generalists without enhancing the time-reversed process. In contrast, faster or slower environmental dynamics may be cast into effective static fitness landscapes (39), and are thus unable to maintain a strong kinetic asymmetry between specialists and generalists. In this sense, the intermediate cycling mechanism studied here exploits a truly non-equilibrium evolutionary 'seascape' (78; 114) with no static analog.

Our framework proposes novel protocols for evolving generalists, such as a 'chirp' where the environment is cycled at an increasing frequency, and predicts optimal correlations between antigens to be used. Since we use a sufficiently abstracted model of B-cell affinity maturation, our analysis might be adapted for other temporal evolution protocols, e.g., to avoid antibiotic resistance(163; 91; 42) and for cancer treatments (56; 71).

Numerous works have studied evolution in time-varying environments, including in the context of evolving generalists (70; 46; 164; 60; 69; 85; 185; 133; 75; 150). Relatively fewer works(39; 113; 79; 97) have analyzed the case of intermediate timescales where the environment changes before populations reach steady state, though these works do not consider the high dimensional genotypic space and correlated environments studied here. In this broader sense, our work is a step towards a theory of evolution in time-varying environments with no separation of timescale between the evolutionary response of populations and environmental changes.

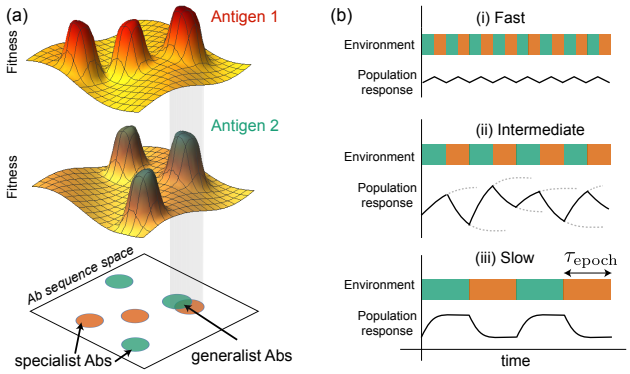

Fig 2.1: Time-varying environments on intermediate timescales can dynamically funnel specialists to generalists. (a) Generalist antibodies that bind multiple antigens can be hard to evolve during B-cell affinity maturation as compared to specialists that only bind one antigen. Specialists for an antigen can constitute a single (Fig.2.2) or multiple islands (Fig.2.4) in antibody sequence space. (b) We consider time-varying selection pressure on timescales (i) fast, (ii) intermediate or (iii) slow relative to evolutionary transients. In the intermediate regime, the selection pressure (e.g., antigen) changes before evolutionary transients (dashed lines) are complete and a steady state is reached.

2.4 Results

We study evolution in fitness landscapes with multiple fitness peaks in antibody sequence space as shown in Fig.2.1. During affinity maturation, each antigen defines a distinct 'environment' and thus a distinct fitness function with distinct fitness peaks. In general, 'specialist' fitness peaks for one antigen are not fitness peaks for other antigens. However, we assume one of these fitness peaks is approximately in the same location for all antigens. We first study evolution in the vicinity of this 'generalist' fitness peak and ignore the larger landscape. True generalists are found at the intersection of these peaks across environments; the challenge in evolving such generalists is primarily entropic. We then consider evolution on the full landscape with multiple fitness peaks; now, fitness valleys can prevent the evolution of generalists. By exploiting mathematical constructions from spin glass theory, we systematically study the impact of the relative placement of fitness peaks, or equivalently, correlation of features across antigens. In both cases, we model populations, e.g., the population of B-cells across all germinal centers in an organism. We explain our results in terms of the rate at which a

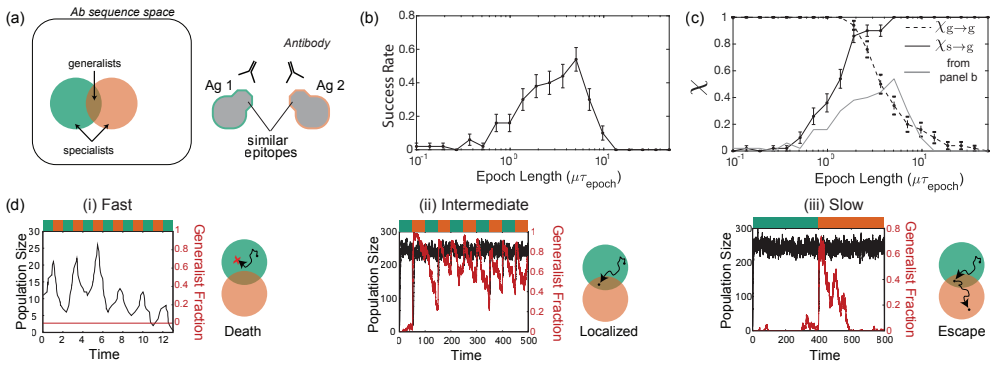

Fig 2.2: Intermediate timescale cycling of antigens strikes a balance between evolving and maintaining rare generalist antibodies. (a) We assume many specialist antibodies can bind each antigen at a partially conserved epitope; see text for model. Generalists and specialists have similar fitness. (b) Cycling antigens at an intermediate timescale τ_{epoch} most reliably yields generalists in repeated $K = 500$ population simulations. (c) An initially-specialist population is more likely to evolve generalists (higher $\chi_{\text{s}\to\text{g}}$) with slower cycling since (d,i) fast cycling typically leads to death of the entire population before any generalists are evolved. In contrast, slow cycling allows generalists to specialize; the probability of an initially-generalist population that remains generalists, $\chi_{\text{g}\to\text{g}}$, falls with τ_{epoch} (see (d,iii)). (d,ii) Intermediate timescale switching allows sufficient time for generalists to evolve from specialists without providing enough time for generalists to specialize.

population of specialists evolves generalists in time-varying environments relative to the rate of the time-reversed processes from generalists to specialists.

Both models here have been used in the context of affinity maturation((111; 30; 173; 127) and (34; 45)), corresponding to different molecular models of antigen-antibody binding. While more extensive antibody-antigen binding assays(84; 53; 25) can clarify the situation for a particular virus like HIV, we remain agnostic to the issue here and study both cases since they might be relevant in different evolutionary contexts.

2.4.1 Entropically disfavored generalists

A basic difficulty in evolving generalists is that generalists are often far fewer in number than specialists. This is schematically shown in Fig. 2.2a, where specialists in each environment form a connected set of genotypes of similar fitness. The relatively few generalists, found at the intersection of such sets, can easily mutate into the more numerous specialists in any

fixed environment.

We study the problem quantitatively in a simplified molecular model of antigen-antibody binding, as used for affinity maturation against HIV antigens. Antibodies bind to a single epitope, partially conserved across antigens $\eta = 1, 2$. A (binary) antibody sequence $\mathbf{x}$ binds to an epitope sequence $\mathbf{h}^\eta$ with an affinity given by an additive sum-of-sites model: $\mathbf{x} \cdot \mathbf{h}^\eta$. Antibodies that bind above a threshold T are assigned fitness $s(\epsilon - 1) > 0$, while those that bind weaker have fitness $-s < 0$. We take $1 < \epsilon < 2$, such that the average fitness of an antibody across antigens is negative.

Since the epitope is relatively but not entirely, conserved across antigens, $\mathbf{h}^\eta$ for different antigens are assumed to share a conserved region of length $L_c = 12$ but have a variable region of length $L_v = 7$ (173) (see Fig. 2.8 for other choices). While based on a simple model of molecular binding, our results below apply broadly to the phenomenological description of specialists as connected islands of relatively uniform fitness, with no fitness barriers separating the generalists.

We simulate a finite population ($N \sim 500$) of antibodies in an environment that switches between antigens 1 and 2 on a timescale τ_{epoch} using a birth-death model (see Section 2.5.1), working in the limit of frequent mutations ($\mu N > 1$). Initializing a monoclonal population in a random specialist state for antigen $\eta = 1$, we monitor the fraction of generalists in the populations at late times (Fig. 2.2d), systematically varying the timescale of switching τ_{epoch}. Averaging over many simulation, we find that neither fast nor slow cycling is able to reliably elicit generalists in the population; however, an intermediate timescale of switching is able to do so (Fig.2.2b).

We sought to understand the origin of this non-monotonic behaviour by examining population dynamics in the limits of fast and slow cycling. For fast cycling, (i.e., small τ_{epoch}), the initial specialist population is repeatedly confronted with an antigen it cannot bind to. Without enough time to mutate into a generalist, purifying selection drives the

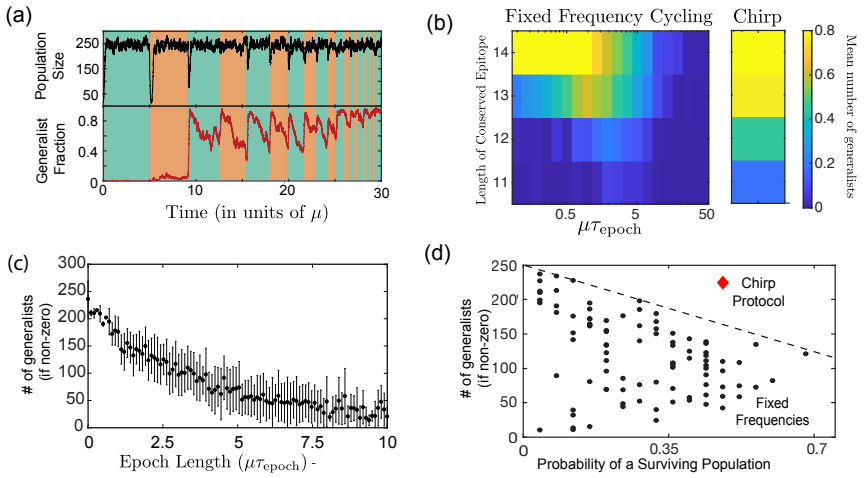

Fig 2.3: Chirped cycling yields generalist populations more robustly than fixed frequency cycling. (a) We consider chirped protocols, where the cycling frequency is increased over time, $\tau_{\text{epoch}} \to \frac{5}{6}\tau_{\text{epoch}}$ after each epoch. (b) When the number of generalists Ω_g is reduced by reducing the length of the conserved epitope, the range of τ_{epoch} that yields generalists decreases. Chirped cycling, however, continues to recover generalists with little parameter tuning. (c) Fixed frequency cycling results in a tension between high number of generalists if the population survive (high for fast cycling) and population survival (high for slow cycling), resulting (d) in a trade-off along a Pareto front. Chirped cycling breaks the trade-off since slow cycling initially ensures population survival and fast cycling later on ensures that a high fraction of the surviving population are generalists.

population to extinction (Fig. 2.2d,i). Consequently, the fraction of trials in which specialists evolve into generalists, $\chi_{\text{s}\to\text{g}}$, is low (Fig. 2.2c).

In fact, in this limit the dynamics of the population are effectively described by a static, average landscape, where the specialist has fitness $s(\epsilon - 2) < 0$. In this regime, we find that purifying selection drives the population to extinction when $s > \mu \log N$; see Section. 2.5.1 for derivation and discussion of alternative cases where purifying selection is reduced.

On the other hand, for very slow cycling (large τ_{epoch}), any generalists that arise have enough time to specialize again by mutational drift (Fig. 2.2d,iii). As a result, the fraction of an initially-generalist population that stay generalists over an environmental cycle, $\chi_{\text{g}\to\text{g}}$, falls with τ_{epoch}, as seen in Fig.2.2c.

Consequently, we find that intermediate timescale cycling strikes a balance: providing

enough time to for specialists to evolve into generalists (high $\chi_{\mathrm{s}\to\mathrm{g}}$), but not enough time for generalists to switch back to specialists again (high $\chi_{\mathrm{g}\to\mathrm{g}}$). In Section. 2.5.1, we determine this regime to be,

$$\tau_{\min} \sim \frac{1}{\mu} d_{\mathrm{init}\to g} < \tau_{\mathrm{epoch}} < \tau_{\max} \sim \frac{1}{\mu} \log\left(\Omega_g N\right) \tag{2.1}$$

where $d_{\mathrm{init}\to g}$ and Ω_g are the mutational distance of the initial naive repertoire from generalists, and the number of generalist genotypes, respectively; see Sections 2.5.1,2.5.1.

Notably, an intermediate regime – that is, a cycling time τ capable of eliciting generalists – only exists when the number of generalists, Ω_g, is sufficiently large: $\log \Omega_g N > d_{\mathrm{init}\to g}$. In contrast, when the number of specialists is large compared to the number of generalists, and population sizes are small, it takes longer for generalists to evolve from specialists than to specialize again. In this regime, the entropic bias in sequence space driving generalists to specialists is large and even fixed frequency cycling may not produce generalists.

Hence, we propose a new dynamic protocol - a 'chirp' - that can alleviate this tension between evolving generalists from specialists ($\chi_{\mathrm{s}\to\mathrm{g}}$), which requires slower cycling, and the ability to maintain a population of generalists ($\chi_{\mathrm{g}\to\mathrm{g}}$), which requires faster cycling. A chirp, shown in Fig. 2.3, starts with slow cycling and increases the cycling frequency over time. Such highly dynamic 'chirp' protocols outperform any fixed frequency cycling protocol; see Fig. 2.3c.

2.4.2 Generalists isolated by fitness valleys

We now consider a more general case where fitness valleys separate viable genotypes, and specialists and generalists form disconnected sets in sequence space. Such models have been used to describe antibodies for influenza and malaria (111; 30; 45; 34), as well as describing RNA molecular fitness landscapes (129; 24). Rugged landscapes are relevant whenever

mutations can act non-additively; that is, when epistasis is present. Indeed, epistasis has been broadly observed for molecular phenotypes and was quantified recently for antigen-antibody binding interactions(1). In the affinity maturation context, such a model with multiple fitness peaks naturally arises if each antigen has multiple epitopes, with one epitope shared across antigens (34).

Here, we take a phenomenological approach that is agnostic to molecular details. Exploiting Hopfield's (61) (or more generally, Gardner's (55)) construction, we construct fitness landscapes for each antigen with fitness islands around sequences corresponding to each epitope. In particular, consider P epitopes on each antigen $\eta = 1$, 2, that bind to antibody sequences $\mathbf{h}_\alpha^\eta$ ($\alpha = 1, \dots P$). The fitness of an antibody with sequence $\mathbf{x}$ confronted by antigen η is chosen to be $F^\eta \propto s \sum_\alpha \kappa_\alpha (\mathbf{x} \cdot \mathbf{h}_\alpha^\eta)^p$ where we set $p = 2$ (the Hopfield model). This minimal construction produces fitness landscapes with peaks at the specified epitopes $\mathbf{h}_\alpha^\eta$, provided P is sufficiently small compared to sequence length L (6). Larger p creates more sharply defined fitness peaks. Finally, the weights κ_α are used to reduce the fitness of generalists relative to specialists in any one environment.

By making different choices for the epitopes $\mathbf{h}_\alpha^\eta$, we may construct fitness landscapes with arbitrary amounts of correlation between them. We begin by studying the minimal case where 1 epitope is shared between the two antigens, $\mathbf{h}_1^1 = \mathbf{h}_1^2$, with the others epitopes being uncorrelated. Later, we relax this assumption. For our theoretical analysis, we assume selection is strong and beneficial mutations are rapidly fixed, $sN \gg \mu N$, $sN \gg 1$; hence fitness valleys between islands play a significant role.

We simulate a finite population of antibodies evolving via Moran dynamics. Initialising a monoclonal population at a specialist, we once again carried out simulations at different antigen switching times, τ_{epoch}, and quantified the fraction of generalists in the population at long times. As seen in Fig.2.4b, an intermediate timescale of switching elicits generalists in the population. This is reminiscent of the entropic model above, but for different underlying

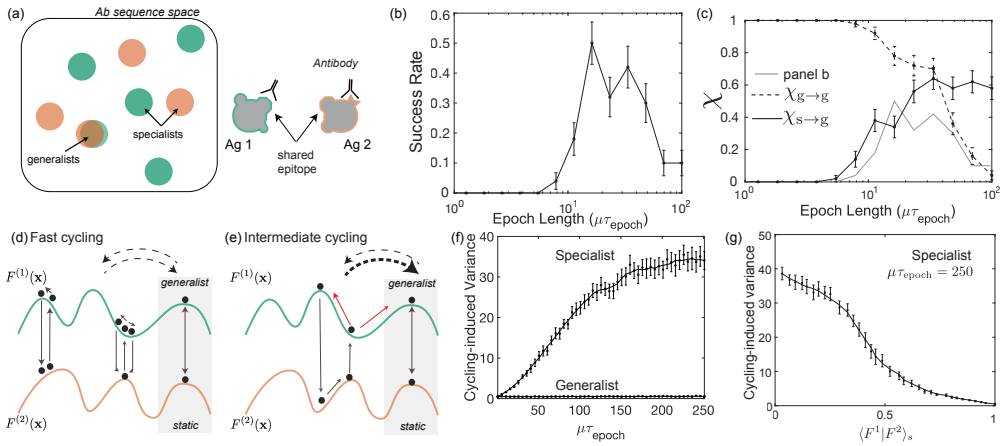

Fig 2.4: Intermediate timescale cycling enhances specialist-to-generalist conversions across fitness valleys without enhancing the time-reversed process. (a) Antibodies that bind distinct epitopes on antigens (right) form distinct specialist islands (left) in sequence space, separated by fitness valleys. Generalists bind an epitope shared by antigens. (b) Cycling at intermediate τ_{epoch} most reliably yields generalists in a finite population $N = 100$ simulation. (c) Specialist-to-generalist transitions, $\chi_{s\to g}$, grows with τ_{epoch}, while the ability to retain generalists $\chi_{g\to g}$ falls (both measured after $n = 30$ cycles). (d) Fast cycling traps populations at fitness peaks near where they are initialized. (e) But intermediate τ_{epoch} allows evolution between specialists. Such evolution introduces sequence variance even in initially monoclonal specialist populations (red arrows in (e), quantified in (f)) but not for generalists. Such higher variance for specialists enhances specialists-to-generalists transitions but not the reverse process. (g) Cycling-induced variance is largest when specialists in $F^{(1)}, F^{(2)}$ are uncorrelated (low $\langle F^{(1)}|F^{(2)}\rangle_s$).

reasons.

Here, fast switching fails to produce generalists because populations stay confined to their initial position(178) (Fig.2.4d). Rapid switching can be approximated by the averaged fitness landscape if the switching is fast enough and each individual has a fitness given by its fitness averaged over environments experienced in its lifetime. In such cases, new fitness peaks and valleys can be created as shown before for the spin glass-like fitness functions used here(6). Consequently, the population remains segregated away from the generalist genotypes by valleys of low-fitness, and generalist acquisition, $\chi_{\mathrm{s}\to\mathrm{g}}$, is small. In practice, such populations stuck in a specialist genotype for extended time can go extinct in the presence of multiple antigens (173).

In contrast, at slower switching times, evolution in each environment can shift the population away from its initial position in the prior environment (Fig.2.4d). As shown in Section. 2.5.3, this requires at least time $\tau_{\min} \sim d_{12}/\mu$, where d_{12} is the typical mutational distance separating specialists across environments. Consequently, the population is forced to continually traverse genotype space. This continual evolution is by necessity stochastic (Fig.2.4f), contingent on the random order of mutations that arise, as well as on any potential population variance. This cycling-induced mobility, augmented by stochasticity, allows the population to widely explore genotype space and find the generalist, and hence $\chi_{\mathrm{s}\to\mathrm{g}}$ rises (Fig.2.4d).

Importantly, upon evolving into generalists, environmental cycling no longer disturbs the population, as the fitness of generalist sequences does not appreciably change over time. Thus cycling breaks the symmetry between specialists and generalists and enhances $\chi_{\mathrm{s}\to\mathrm{g}}$ without enhancing $\chi_{\mathrm{g}\to\mathrm{s}}$. Intuitively, intermediate cycling selectively 'warms up' (i.e., increases stochasticity) specialist parts of sequence space, naturally leading the population to collect in 'cooler' generalist sequences.

Cycling significantly slower than $\tau_{\min}$ is counterproductive. The cycling-induced leaks

from specialists to generalists only occur due to environmental switches; hence unnecessarily long τ_{epoch} only adds dead time with no additional population divergence.

In the meantime, as shown in Section. 2.5.3, escape from generalists to specialists becomes significant on timescales of $(1/\mu)e^{\Delta F_g N}$ where ΔF_g is the fitness of the generalist relative to the fitness valley separating it from specialists; N is the population size. See (64; 166; 178) for calculations of valley crossing rates in other parameter regimes. These considerations limit intermediate timescales favorable for evolving generalists:

$$\tau_{\text{min}} \sim d_{12}/\mu \quad < \quad \tau_{\text{epoch}} \quad < \quad \tau_{\text{max}} \sim (1/\mu)e^{\Delta F_g N} \tag{2.2}$$

As in the earlier model, if $\tau_{\text{min}} > \tau_{\text{max}}$, fixed frequency cycling may fail. In SI Fig. 5, we find that chirped cycling can continue to recover generalists, even in these regimes. Chirp protocols produce generalists by alleviating the tension between $\chi_{\text{s}\to\text{g}}$ and $\chi_{\text{g}\to\text{g}}$ and do not require fine-tuning of parameters, as before in our models of entropically disfavored generalists.

Correlation between specialists

The effectiveness of this theoretical cycling mechanism depends on the correlation between specialists of $F^{(1)}$ and $F^{(2)}$, as demonstrated in a recent study of generalist evolution in tunably correlated landscapes (172): if specialists of $F^{(1)}$ and $F^{(2)}$ are similar or well within each other's attractors, cycling will primarily cycle the population between specialists with minimal divergence into generalists. In contrast, given that generalists exist, least similarity between specialists of $F^{(1)}$ and $F^{(2)}$ would best enable reliable evolution of generalists. As shown in Section. 2.5.3, we can quantify relevant correlations by

$$\langle F^{(1)} | F^{(2)} \rangle_s \equiv \frac{c_{1,2}}{\sqrt{c_{1,1}\, c_{2,2}}}$$

where $c_{\eta,\gamma} = \frac{1}{LP}\sum_{\alpha,\beta\neq 1}\mathbf{h}_\alpha^\eta \cdot \mathbf{h}_\beta^\gamma$ excludes the generalist pattern $\mathbf{h}_1^1 = \mathbf{h}_1^2$. When $\langle F^{(1)}|F^{(2)}\rangle_s$ is high, cycling-induced variance is low; see Fig.2.4g. Consequently, the small asymmetry between $\chi_{\mathrm{s}\to\mathrm{g}}$ and $\chi_{\mathrm{g}\to\mathrm{s}}$ created by a single environmental cycle must be compounded by cycling multiple times; however, in practice, other considerations might limit the number of such cycles. Hence, our proposal requires specialists of $F^{(1)}$ and $F^{(2)}$ to be sufficiently uncorrelated (low $\langle F^{(1)}|F^{(2)}\rangle_s$).

Is cycling a practical strategy given correlations between specialist antibodies found during HIV infection and physiological parameters for population dynamics?

We analyzed specialist and generalist antibody sequences collected from an HIV patient (84; 53; 25); see Fig.2.5a. We constructed landscapes $F^{(1)}, F^{(2)}$ with fitness peaks at these observed specialist and generalist sequences following Gardner's construction (55); as detailed in Section. 2.5.2, we repeated the analysis for multiple choices of fitness functions and restriction of sequence data to variable regions.

Simulations of cycling environments $F^{(1)}, F^{(2)}$ constructed from the above sequence data evolved generalist antibodies, while simultaneous presentation of both antigens, a practical alternative to fast cycling(173), fails to produce such generalists; see Fig.2.5b. We then artificially shuffled antigen labels for antibodies, so that CH105 was considered a Ag2 specialist and CH186, an Ag1 specialist and reconstructed $F^{(1)}, F^{(2)}$. This artificial shuffling significantly increased the correlation $\langle F^{(1)}|F^{(2)}\rangle_s = 0.78$ compared to the real data ($\langle F^{(1)}|F^{(2)}\rangle_s = 0.43$). Cycling is no longer effective in evolving generalists. We conclude that the low correlation between specialists in the real data is crucial for time-varying selection of generalists, in line with the result of Ref. (172).

While our model here did not explicitly account for extinction, simultaneous presentation or fast cycling can cause most specialist B-cells to perish, especially if many distinct antigens are used (see Section. 2.5.1). In this more realistic case, 'chirped' cycling at increasing frequency as in Fig.2.3 will alleviate the tension between $\chi_{\mathrm{s}\to\mathrm{g}}$ and $\chi_{\mathrm{g}\to\mathrm{g}}$ as demonstrated

in Figure 2.4c. That is, initial slow cycling allows the system to take advantage of cycling-induced stochasticity to find the generalist (the regime of high $\chi_{s\to g}$), while fast cycling towards the end forces the localization of the population to the generalist (high $\chi_{g\to g}$).

2.5 Supporting Information

2.5.1 Entropically Disfavored Generalists

Model

To construct landscapes with entropically disfavored generalists, we model antibodies and antigens in a manner similar to the one described by by Wang et. al(173). In this model, the sequence of each antibody, $\mathbf{x}$ is a sequence of length L with entries ± 1. Each antigen, indexed by η, is assumed to have an epitope of sequence, $\mathbf{h}^{\eta}$. Each epitope is length L with entries ± 1. The binding energy of a given antigen to an antibody is given by an additive sum-over-sites model:

$$E\left(\mathbf{x}, \mathbf{h}^{\eta}\right) = -\frac{1}{L}\sum_{i}^{L} h_i^{\eta} x_i \tag{2.3}$$

The fitness of each antibody $\mathbf{x}$ in the presence with antigen η is given by thresholding its binding affinity, as follows:

$$F^{(\eta)}\left(\mathbf{x}\right) = s\epsilon\Theta\left(-\left(E(\mathbf{x}, \mathbf{h}^{\eta}) + \frac{T}{L}\right)\right) - s \tag{2.4}$$

Here, $\Theta(x)$ is the Heaviside function. $\frac{T}{L}$ is the binding energy threshold that an antibody must overcome before reaping a fitness benefit. The binding energy threshold, $\frac{T}{L}$ can be translated into minimum number of binding interactions needed to reap a fitness benefit, T_{sites}, by $T_{\text{sites}} = (L+T)/2$. By construction, all fit individuals have the same fitness $s(\epsilon - 1) > 0$ and all unfit individuals are equally unfit to an extent $-s$, resulting in a degeneracy of fit

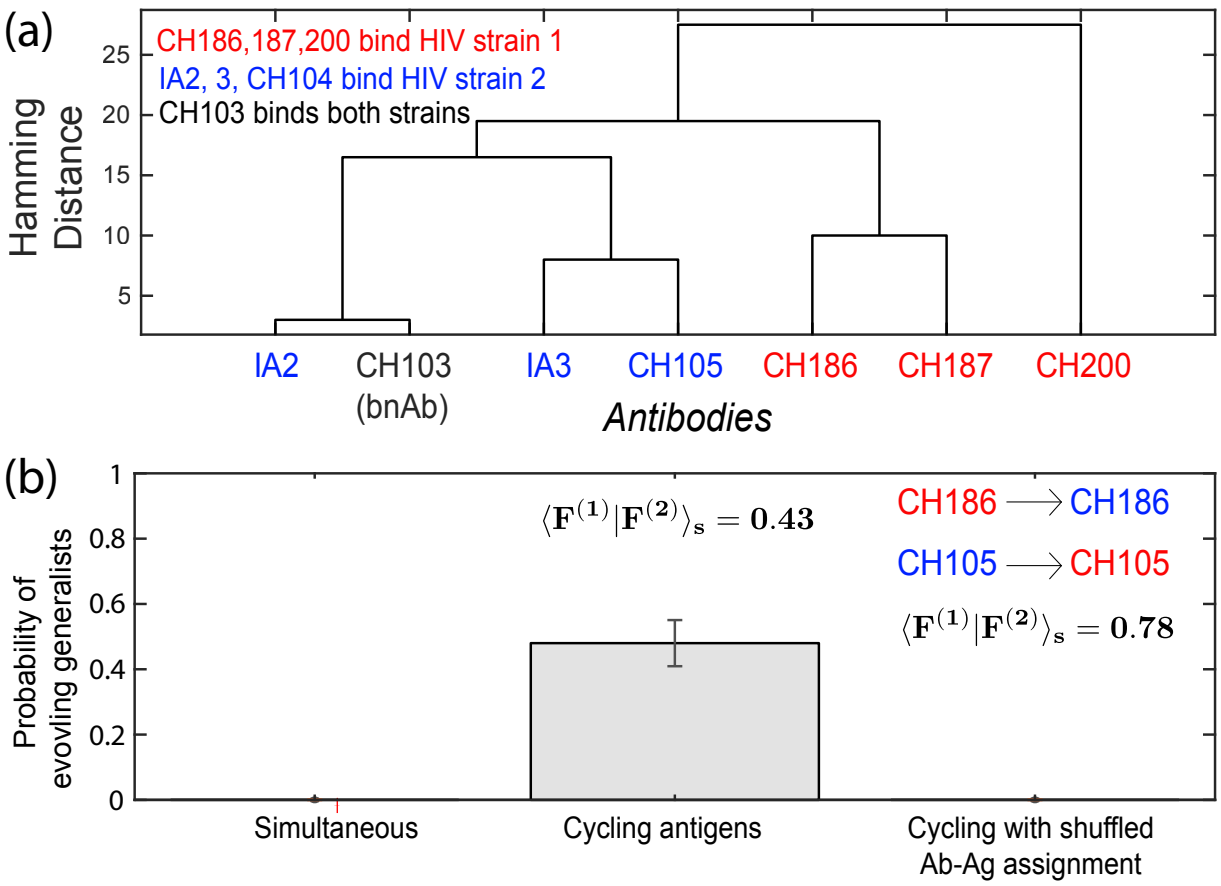

Fig 2.5: Cycling between fitness landscapes constructed using antibody sequences from HIV patients yields generalists; however, cycling is less effective for artificially shuffled data with higher specialist correlation. (a) Sequence divergence of antibodies that bind two distinct strains (red, blue) of HIV. See SI Antibody Sequence Data for sequence and binding affinity data, reproduced from (84; 53; 25). (b) Following Gardner(55), we constructed two fitness landscapes $F^{(1)}, F^{(2)}$ with peaks at red, blue sequences resp. and simulated evolution with realistic parameters (see Section. 2.5.3. Generalists are evolved only if antigens are cycled. Cycling is less effective if we shuffle antibody-antigen assignment: CH105 now considered specialized for strain 2 (i.e., now red), CH186 for strain 1 (i.e., now blue). Shuffling artificially increases specialist correlation $\langle F^{(1)} | F^{(2)} \rangle_s$ from 0.43 to 0.78.

and unfit genotypes.

Specialists and Generalists

Here, we demonstrate the fraction of antibodies that are generalists for two antigens. Biological constraints impose that parts of an antigen's epitope is conserved, while other parts are variable as the viral strain evolves. As such, we suppose L_c sites of the L total sites are fixed across η, while the others are unique to each antigen. This constraint results in the possibility that some antibody sequences have a positive fitness for all antigens. Such antibodies are called generalists. The number of generalists is a function of the threshold T and the lengths of the conserved L_c and variable $L_v = L - L_c$ regions. In particular, if $T_{\text{sites}} > L_c + \frac{1}{2}L_v$, there are no generalists. A simple equation to compute the fraction of antibodies with positive fitness for an antigen that are generalists is given as follows:

$$\frac{\Omega_g}{\Omega} = \frac{\sum_{j=0}^{L} \binom{L_c}{j}\binom{L_v}{T_{\text{sites}}-j}}{\sum_{k=T_{\text{sites}}}^{L} \binom{L}{k}} \tag{2.5}$$

where, $\binom{N}{m}$ is the combinatorial function, Ω_g is the number of generalists, and Ω is the number of antibodies with positive fitness. By rule, this function is zero if $m > N$ or $m < 0$. Here, j indices the number of sites matched in the conserved portion of the antibody and k indices the number of overall sites matched along the string.

Here, we considered $L = 19$, $L_c = 12$, and $T = 11$. The proportion of antibodies with positive fitness that are generalists for these parameter choices is $\approx 1.3\%$. This choice is qualitatively similar to the analysis developed in (173) based on experiments there. The analysis can be repeated for longer sequences and the results are qualitatively unchanged; the primary effect of changing L is explained by the change in entropy, as predicted by SI Equation 2.5.

Finite population simulation

Affinity maturation is an evolutionary process for antibodies with complex population dynamics(132). Here, we first model this process using a simplified canonical birth-death-mutation model - a 'Yule' process (92) - commonly used to study evolutionary dynamics. The 'Yule' process ignores many of the molecular details of affinity maturation while still enabling us to develop a minimal model of evolutionary dynamics in time-varying environments. We then verify our results with an independent simulation that accounts for population dynamics and complexities inherent to affinity maturation.

Yule Process The key ingredients in a Yule process are:

- **Mutation** with rate μ per individual, in which a single site on the genome is mutated (i.e. a single bit-flip of $\mathbf{x}$).
- **Birth-death** with rate λ per individual.
- The population size is maintained at a carrying capacity K by modulating the probability of replication by a factor $(1 - N/K)$. With the particular fitness function of this model, we have $\Pr(\mathbf{x} \text{ reproduces}) = \Theta\left(E(\mathbf{x}, h^\eta) - T\right)(1 - N/K)$. If the individual does not reproduce, it is removed from the population. Here, N is the current population size, $K = 500$ is a carrying capacity that prevents the population from growing indefinitely, and $F^{(\eta)}$ is the fitness of that individual in the (current) landscape η.

At each event, time is advanced by the usual exponentially distributed amount. The environment η is taken to alternate between $\eta = 1$ and $\eta = 2$ every τ_{epoch}.

Choosing units of time by setting the birth-death rate $\lambda = 1$, we set the mutation rate to $\mu = 0.05$. Finally, we set the carrying capacity $K = 500$. We can infer from these choices that $s = 1, \epsilon = 2 - \frac{2N}{K}$.

We evolved an initially monoclonal population of size $N = 10$ (initialised with $\mathbf{x} = \mathbf{h}^\eta$ for all individuals). Simulations were run for either a fixed time $t = 100$ (in units of λ) or for

$t = 10\tau_{\text{epoch}}$, whichever is longer, and we performed 25 replicates for each value of τ_{epoch}. For each run, we saved the number of generalists and the overall population size at the end of the simulation. In Fig. 2b of the main text, we plotted the proportion of trials that had more than 10 generalist antibodies at the end of the simulation, finding that the proportion was high for an intermediate rate of cycling.

Affinity Maturation Inspired Model To more directly model affinity maturation, we also simulate a model that mirrors the known dynamics of B-cells in germinal centers (173; 35; 34). The steps are as follows:

- B-cell clones expand without significant mutation in the first week after vaccination. We model such a formation of germinal centers by taking a B-cell with an antibody that meets the binding affinity threshold for one antigen and replicate it to a size of 1500 B-cells.

- We model the reproduction and somatic hypermutation phase of affinity maturation in the dark zone of the germinal center by allowing each B-cell to duplicate twice with a mutation rate of 0.00625 per replication per base pair.

- We then model the selection phase in the light zone by determining if each B-cell in our B-cell population can internalize antigens it encounters on a follicular dendritic cell (FDC). We say that a B-cell can internalize an antigen if its antibody's binding affinity for that antigen, as given by Equation 2.3 meets a threshold, T.

- B-cells receive T-cell help to avoid an apoptosis signal as a function of whether or not they internalized antigen. As in earlier work(173; 35; 34), we assume the probability B-cells do not receive help increases with binding energy is proportional to $\exp(\alpha(F^{(\eta)}(\mathbf{x}) - F_{\text{threshold}}))$.

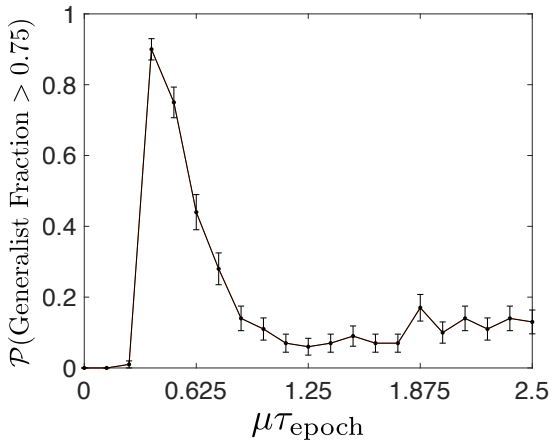

Fig 2.6: Evolving generalists using a detailed model of affinity maturation. We simulated cycling antigens in a model with known details of the population dynamics of B-cells in germinal centers. Here, we plot the probability that the population at the end of affinity maturation has at least 75% of its population in a generalist genotype. We observe a resonant peak in this probability, similar to results presented in the main paper for the simpler population dynamics model based on a Yule process.

- The surviving B-cells are recycled into the dark zone. We repeat the steps above until the B-cell population grows to be larger than 2000 or the process has cycled 100 times. These choices model antigen depletion on the follicular dendritic cells (FDCs) in the germinal center and antigen decay.

We present the results of this simulation in Fig. 2.6.

We find that the results from using this affinity maturation-specific evolutionary scheme are qualitatively similar to the minimal Yule process model. Note that this affinity maturation model incorporates numerous ingredients, particular to affinity maturation, that are not captured by the Yule process used in the earlier section. E.g., the specifics of birth and death, carrying capacity and details of how the affinity maturation process terminates differ. And yet we obtain qualitatively similar results, showing that our results are primarily tied to the broad topology of the fitnesss landscape and the ratio of broadly relevant timescales and not to particular details of evolutionary population dynamics.

Evolution between Specialists and Generalists: $\chi_{\mathrm{s}\to\mathrm{g}}$ and $\chi_{\mathrm{g}\to\mathrm{g}}$

As discussed in the main text, there are two possible failure modes: (1) at cycling rates too fast, the population does not have the time to evolve a generalist before adverse selection result in population extinction, (2) at cycling rates too slow, the population loses its ability to maintain generalists. In order to illustrate the tension between cycling too fast and cycling too slow, we compute two quantities:

- $\chi_{\mathrm{g}\to\mathrm{g}}$, the fraction of trials starting from an initially generalist population maintaining at least 20% generalists after one epoch.

- $\chi_{\mathrm{s}\to\mathrm{g}}$, the fraction of trials in which a monoclonal population of specialists, initialized at $\mathbf{x} = \mathbf{h}^{(1)}$, evolve a single generalist within an epoch.

Both $\chi_{\mathrm{g}\to\mathrm{g}}$, $\chi_{\mathrm{s}\to\mathrm{g}}$ are computed from 50 replicates for each τ_{epoch}. As plotted in Fig. 2c, we observe that $\chi_{\mathrm{s}\to\mathrm{g}}$ starts initially at 0 and rises with τ_{epoch}, while $\chi_{\mathrm{g}\to\mathrm{g}}$ starts at 1 and falls with τ_{epoch}.

Population traces: Fig. 2d shows population traces in single runs. As in Section 2.5.1, we initialized a population at $\mathbf{x} = \mathbf{h}^{(1)}$ with parameters as above. Generalist fraction is defined as $\frac{\text{number of generalists}}{\text{population size}}$.

We use the following values of τ_{epoch} for Fig.2(d): Fast cycling $\tau_{\mathrm{epoch}} = 1$ (Fig. 2d(i)), Intermediate cycling $\tau_{\mathrm{epoch}} = 60$ (Fig.2d(ii)), Slow cycling $\tau_{\mathrm{epoch}} = 400$ (Fig.2d(iii))

Timescale Analysis

Our numerical study found that an intermediate timescale of environmental cycling, $\tau_{\min} < \tau_{\mathrm{epoch}} < \tau_{\max}$, was most effective at obtaining generalists. Here we estimate the bounds, $\tau_{\min}$and $\tau_{\max}$, in terms of the mutation rate μ, the population size N, the length of the genotype L, and the distance from the initial ancestral genotype to the generalist, $d_{i\to g}$.

Finding the generalist: $\tau_{\min}$ Consider a population initialised as a specialist for antigen 1. For sufficiently strong selection pressure ($s > \mu \log N$), purifying selection drives the population to extinction if a generalist has not been discovered before the environment switches to antigen 2. Thus we demand that τ_{epoch} is long enough for the population to evolve a generalist in a single epoch.

As fitness is uniform across the specialist region, the population must discover the generalist by diffusion. An initially monoclonal population of size N diffuses out from the initial genotype. If the population size is much smaller than the number of possible genotypes ($N \ll 2^L$), there are two possible regimes of the diffusive search:

1. The initial genotype is far from the generalist: more precisely, the population size N is smaller than the set of sequences between the initial genotype and the generalist. In terms of the Hamming distance between the initial genotype to the generalist, $d_{i\to g}$: $\sum_{d=0}^{d_{i\to g}} \binom{L}{d} > N$

   In this regime, finding the generalist is a rare event, requiring time $\mu\tau_{\min} \sim 2^L$, i.e. the time taken to explore all of genotype space. It is therefore extremely improbable that the generalist will be found, and population extinction is likely.

2. The initial genotype is close to the generalist: that is, the generalist is sufficiently close to the initial condition that it may be found by the diffusing population of antibodies:$\sum_{d=0}^{d_{i\to g}} \binom{L}{d} < N$

   For $L = 19$ and $N = 500$, as used in the simulation, this suggests that for $d_{i\to g} \leq 4$ the generalist may be reasonably found by diffusion.

Assuming that we are in the latter regime, we may estimate $\tau_{\min}$ from $\langle d(t)\rangle$, the average distance away from the initial condition that an individual has diffused in time t, by solving:

$$\langle d(\tau_{\min})\rangle \sim d_{i\to g} \tag{2.6}$$

We compute $\langle d(t) \rangle$ as follows: the probability that a diffusing individual may be found at (Hamming) distance d from its initial genotype is:

$$P(d,t) = \binom{L}{d} \mathrm{e}^{-\mu t} \sinh^d\left(\frac{\mu t}{L}\right) \cosh^{L-d}\left(\frac{\mu t}{L}\right) \tag{2.7}$$

Thus, $\langle d(t) \rangle = \sum_{d=0}^{L} d\, P(d,t) = L\mathrm{e}^{-\frac{\mu t}{L}} \sinh\left(\frac{\mu t}{L}\right)$. Inserting into Eq. 2.6 and solving for $\tau_{\min}$:

$$\tau_{\min} = \frac{L}{2\mu} \log\left(\frac{L}{L - 2d_{i\to g}}\right) \approx \frac{d_{i\to g}}{\mu} \tag{2.8}$$

where the approximation is valid for $d_{i\to g}/L \ll 1$. For values used in the simulation ($L = 19$, $d_{i\to g} = 3$), we obtain $\mu\tau_{\min} \approx 3.5$, which is consistent with our numerical results.

Maintaining a generalist: $\tau_{\max}$ We may similarly estimate the upper bound for effective cycling, $\tau_{\max}$. Supposing that an evolving population has found the generalist, it must now remain localised there. For this to happen, the environment must switch rapidly enough to prune, by purifying selection, those individuals that diffuse away from the generalist.

Consider a monoclonal population of size N at a generalist at time $t = 0$. For simplicity, let us suppose that the generalist has to accrue $d_{g\to s}$ mutations to become a specialist. The number of generalists in sequence space, Ω_g, is then approximated by the volume of the Hamming ball of radius $d_{g\to s}$

$$\Omega_g = \sum_{k=0}^{d_{g\to s}} \binom{L}{k} \approx \binom{L}{d_{g\to s}} \tag{2.9}$$

where we have replaced the sum by its dominant term, valid when the genome length L is large and $d_{g\to s} \ll L$.

From Eq. 2.7, the number of generalists remaining at time t is:

$$
\begin{aligned}
P_{\text{generalists}}(t) &= \mathrm{e}^{-\mu t} \sum_{k=0}^{d_{g\to s}} \sinh^k\left(\frac{\mu t}{L}\right) \cosh^{L-k}\left(\frac{\mu t}{L}\right) \binom{L}{k} \\
&\approx \mathrm{e}^{-\mu t} \left(\frac{\mu t}{L}\right)^{d_{g\to s}} \Omega_g
\end{aligned} \tag{2.10}
$$

where we have once again replaced the sum by its dominant term, assumed that $\mu t/L$ is small (valid when genome length L is large), and used Eq. 2.9 to write $\binom{L}{d_{g\to s}} \approx \Omega_g$.

Then, $\tau_{\max}$ is defined as the time taken for the occupancy of the generalist region to fall below 1 individual, $P_{\text{generalists}}(\tau_{\max}) = 1/N$, i.e. the solution of

$$
\mathrm{e}^{-\mu t} \left(\frac{\mu t}{L}\right)^{d_{g\to s}} = \frac{1}{N\Omega_g} \tag{2.11}
$$

For $t > 1\mu$, the expression on the left hand side is dominated by the exponential; we thereby solve for t to obtain:

$$
\tau_{\max} \sim \frac{1}{\mu} \log \Omega_g \, N \tag{2.12}
$$

For the parameters used in the simulations ($N = 500$, $\Omega_g \approx 7 \times 10^5$, as computed from Eq. 2.5), we have $\mu\tau_{\max} \approx 20$, which is consistent with the numerical results (see $\chi_{g\to g}$ in Fig. 4c in the main text).

Existence of an intermediate timescale The existence of an intermediate timescale τ_{epoch} that produces generalists requires $\tau_{\min} < \tau_{\max}$. However, the above expressions make it clear that $\tau_{\min} < \tau_{\max}$ only if (a) the initial condition is close enough to generalists (small $d_{i\to g}$), (b) the fraction of generalists relative to specialists is large enough (large Ω_g).

Fig SI 2.7a shows how the yield of generalists at intermediate timescales disappears as the initial conditions are made less favorable. This panel was constructed using the simulation method described in Section 2.5.1 with variations on the initial condition for $\mathbf{x}$ such that

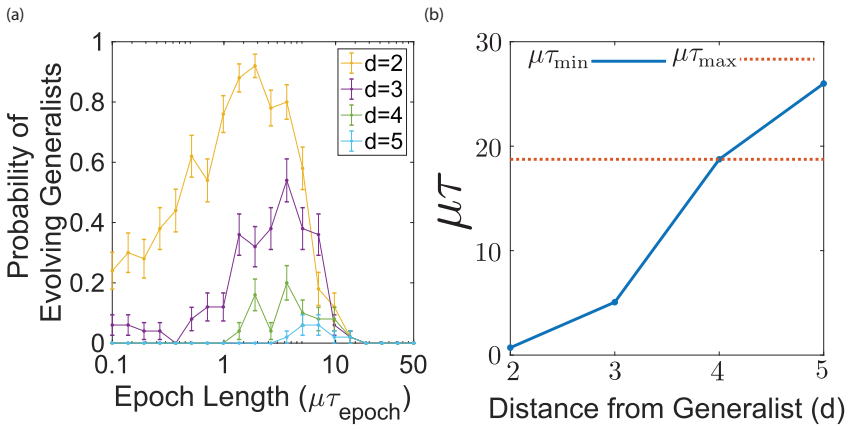

Fig 2.7: (a) We compute the probability of evolving generalists for varying Hamming distances from the generalist, given by d, using the method described in Section 2.5.1. We note that as d increases, the time at which generalists begins to be evolved increases, though the time at which they probability decays remains fixed. (b) We compute $\chi_{s\to g}$ for each initial condition and take $\tau_{\min}$ to be the smallest time where $\chi_{s\to g} > 0.6$. We find that this rises, eventually rising above $\tau_{\max}$ for large enough d. By construction, $\tau_{\max}$ is independent of d.

the Hamming distance between $\mathbf{x}$ and the generalist varied by some distance d. Fig SI 2.7b shows that $\tau_{\min}$ rises in this limit and exceeds $\tau_{\max}$. We approximated $\tau_{\min}$ by computing the smallest time for which $\chi_{s\to g} > 0.6$ and $\tau_{\max}$ by computing the largest time for which $\chi_{g\to g} > 0.6$.

When $\tau_{\min} < \tau_{\max}$ is not satisfied, there is no intermediate timescale. The time needed for generalists to specialize is shorter than the time needed to evolve generalists from specialists. In this case, the chirp protocol described below is still successful at producing generalists.

Simultaneous presentation of antigens

The cycling strategy explored in this paper may not be practical in the fast limit in the context of B-cell affinity maturation. A common practical alternative is vaccination with a cocktail of antigens, i.e., simultaneous exposure to multiple antigens.

Such simultaneous exposure to multiple antigens is mathematically equivalent to fast cycling of those antigens if specific microscopic assumptions about antibody-antigen interactions

in germinal centers hold(173). During the affinity maturation process, folicular dendritic cells (FDCs) host antigens on their surface for B-cells to interact with, and if antibodies expressed on B-cells bind the antigens with high enough affinity, B-cells internalize those antigens. This process enables those B-cells to avoid apoptosis, and thus proliferate and continue affinity maturation.

There are two currently experimentally unresolved hypotheses about antigen presentation by FDCs:

1. Antibodies are fit only if they can bind ALL presented antigens: In this hypothesis, FDCs only present a single antigen or present antigens in a spatially heterogeneous manner. Consequently, each B-cell is randomly exposed to a single antigen at the selection stage of the affinity maturation process. A B-cell must be able to bind ALL presented antigens to survive selection.

Such simultaneous presentation is qualitatively similar to the fast cycling limit studied in this paper. When presented with such a cocktail vaccine, antibodies starting from a naive repertoire are expected to go extinct since such antibodies typically cannot bind all antigens with high affinity, as seen in the experiments of (173).

2. Antibodies are fit if they can bind ANY one of the presented antigens: If the FDCs present the antigens in a homogeneous manner, each B-cell only needs to bind ANY single one of the presented antigens to avoid apoptosis.

In this case, specialists are fit enough to survive early rounds of selection and evolve generalists. But generalists cannot be maintained in preference over specialists unless selection pressures are fine tuned (e.g., specialists are strongly out-competed by generalists once generalists evolve, despite specialists having significant fitness to begin with).

Death and fast cycling

In the fast cycling limit, $\tau_{\text{epoch}} \to 0$, the fitness of specialist antibodies is the average of their fitness in different environments; as seen Eqn. 2.4, this fitness is $s(\epsilon - 2)$.

As discussed above for simultaneous presentation, we only consider the case where fast cycling corresponds to hypothesis (1), where antibodies need to bind all antigens to survive. Hence, we need $s(\epsilon - 2)$ to be sufficiently negative, so that a specialist population of size N typically dies out before reaching the generalists in this fast cycling limit. Since the latter process takes time $\tau_{\text{min}} \sim d_{\text{i}\to\text{g}}/\mu$ as derived earlier and the initial population size is N, the condition for a specialist population to go extinct in the fast limit is $N \exp(s(\epsilon - 2)\tau_{\text{min}}) \sim N \exp(s(\epsilon - 2)/\mu) < 1$. Assuming that $1 < \epsilon < 2$, we find $s > \mu \log N$ as a conservative criterion independent of ϵ.

Thus, cycling is necessitated because of population extinction in the fast cycling limit (or equivalently, in the averaged environment). Population extinction does appear to be relevant in affinity maturation (173). However, given this reliance on population extinction, one can ask whether the cycling strategies proposed here are relevant to other problems. For example, in other evolutionary contexts, can one evolve generalists easily in the fast cycling or averaged environment limit by violating the $1 < \epsilon < 2$ condition?

In fact, reducing death in this manner reduces purifying selection and hence does not always make it easier to evolve generalists. To see this, note that without death, the average fitness of specialists is positive. Consequently, the purifying selection needed to proliferate generalists over specialists is much weaker. Such reduced purifying selection is especially relevant in evolutionary processes that terminate at finite population sizes.

As a concrete example, in Sec I of the SI, we have run simulations of an affinity maturation process that terminates once the population exceeds a threshold, a realistic termination criterion(173). There, we find that if population death is removed (i.e., $\epsilon > 2$), we still fail to find generalists in the fast limit because the germinal centers are filled with a large number

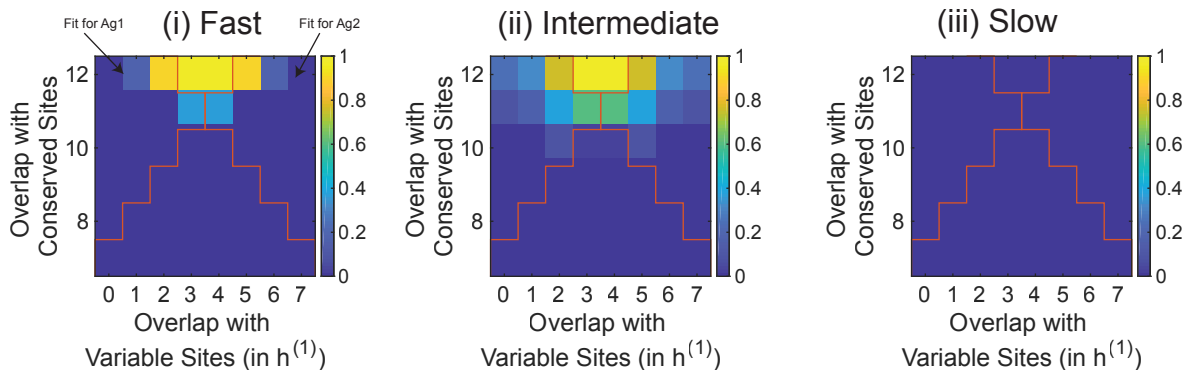

Fig 2.8: Here, the color in the color map represents the proportion of trials leading to a generalist, where yellow corresponds to all trials, while blue corresponds to no trials leading to a generalist. We initialize our population at size $N = 10$ with a carrying capacity $K = 500$. We maintain $L = 19$, $L_c = 12$, and $T = 11$. We ran our simulations for a total of ten epochs. We ran each ordered pair of conserved and variable matches for 50 replicates in environment $\eta = 1$ and assumed symmetry over environments. (i) Fast cycling ($\mu\tau_{\text{epoch}} = 0.05$) can only evolve generalists when initial conditions are already very close to generalists. (ii) Intermediate cycling ($\mu\tau_{\text{epoch}} = 3$) increases the number of viable initial conditions. (iii) Slow cycling ($\mu\tau_{\text{epoch}} = 15$) does not lead to generalists for any initial condition since this limit is unable to maintain generalists.

of specialists which typically terminates the process. Thus, the principles developed here have larger relevance to any context of evolving generalists where there is sufficient purifying selection.

Dependence on initial repertoire

In the main paper, we initialize our population from a genotype that exactly binds the antigen characterized by genotype $\mathbf{h}^{(1)}$. In SI Fig 2.8, we show that not only does intermediate cycling increase the likelihood of evolving generalists from a given initial condition, it also increases the number of initial conditions (e.g., initial B-cell repertoire) that can lead to generalists. Hence, intermediate cycling increases the effective attractor size of the fitness peak associated with a generalist. In SI Fig 2.8, we consider cycling fast ($\mu\tau_{\text{epoch}} = 0.05 < \mu\tau_{\text{min}}$), cycling at an intermediate rate ($\mu\tau_{\text{epoch}} = 3$), and cycling slowly ($\mu\tau_{\text{epoch}} = 15 > \tau_{\text{max}}$). Blue regions in the heatmap correspond to initial conditions that led to few surviving populations after

ten epochs. Yellow regions in the heatmap correspond to initial conditions that led to many surviving populations after ten epochs. We ran each ordered pair of conserved matches and variable matches for 50 replicates in environment $\eta = 1$ and symmetrized over environments.

Chirp protocol

Trade-off in fixed frequency cycling: The anticorrelated behavior of $\chi_{s\to g}$ and $\chi_{g\to g}$ is indicative of a trade-off between evolving generalists and maintaining them in the population.

We first assess this by only considering simulation runs used in Fig 2b that did not result in extinction and computed the number of generalists at the end of such simulations. This number is plotted in Fig 3a. We note that as epoch length increases, the number of generalists remaining in the population decreases, but the probability of a population evolving a single generalist increases.

To illustrate this point we compute two quantities for each simulation for a given τ_{epoch}:

- The number of generalists (if non-zero): The number of generalists is simply the average number of generalists in the population for simulations where the population survived an evolutionary run. This is plotted on the y-axis of Fig 3c.

- The probability of a surviving population: The proportion of trials for a given τ_{epoch} that a population does not go extinct during the evolutionary run. This is plotted on the x-axis of Fig 3c.

By plotting these two quantities against each other, as in Fig 3c in black dots, we observe a tradeoff front.

Chirp Cycling Breaks the Trade-off: This trade-off leads us to proposing a 'chirp' protocol. In a 'chirp' cycling protocol, we decrease the length of the epoch using a multiplicative factor after each cycle, enabling us to take advantage of high probability of population survival (favored by slow cycling) and still obtain high yield (favored by fast cycling). We

update τ_{epoch} according to the following rule:

$$\tau_{\text{epoch}} \leftarrow k\tau_{\text{epoch}} \tag{2.13}$$

where k is some number smaller than 1. We continue evolving the population until $\tau_{\text{epoch}} << \lambda$. Plotted in Fig 3a is a time trace of the population size and the fraction of the population that is of a generalist genotype. Generalist fraction is $\frac{\text{Number of Generalists}}{\text{Population Size}}$. We note that as the length of each epoch decreases, the generalist fraction decreases less in time, until eventually, it remains stabilized at ≈ 1. Additionally, fluctuations in population size are suppressed. We ran the chirp protocol for 25 replicates and computed the number of generalists at the end of each run, if the population survived the run, and the probability that the population survived a chirp protocol. Plotting this in Fig 3d. demonstrates that the tradeoff boundary has been broken. Finally, we compared the chirp protocol to fixed frequency cycling for L_c ranging from 11 to 14. We computed the mean number of generalists observed over 50 replicates in fixed frequency cycling and plot the results in Fig 3b. We compare this to the mean number of generalists discovered under chirp cycling. For this particular chirp, we set $\kappa = \frac{1}{6}$ and initial $\mu\tau_{\text{epoch}} = 5$. We demonstrate that even in regimes where entropic cost is high, chirped cycling can yield generalists robustly.

We note that a similar tradeoff can be observed in Figure 4c of the main text. This indicates that the chirp protocol will work for models with other fitness landscape topologies, so long as there exists a tension between $\chi_{\text{s}\to\text{g}}$ and $\chi_{\text{g}\to\text{g}}$.

Chirped Cycling Evolves Generalists when $\tau_{\text{max}} < \tau_{\text{min}}$ By implementing the same chirped strategy for initial conditions where $\tau_{\text{max}} < \tau_{\text{min}}$, we find that we have high generalist yield at rates match are near the maximum probability of evolving generalists of the fixed frequency evolutionary runs, as demonstrated in Fig 3b.

2.5.2 Generalists Separated by Valleys

Model

We model the fitness landscape of an antibody binding to an antigen with multiple epitopes through a phenomenological construction, inspired by Hopfield's spin glass landscape.

Consider an antigen η with P_η epitopes (i.e., sets of residues on the antigen that form binding locations for antibodies). Suppose that for each epitope, an antibody with sequence $\mathbf{h}_\alpha^{(\eta)}$ of length L with entries ± 1 binds with high affinity (with $\alpha \in \{1, ..., P_\eta\}$ indexing epitopes). Then, the overall binding affinity to antigen η of any antibody with sequence $\mathbf{x}$, of length L with entries ± 1 is taken to be:

$$F^{(\eta)}(\mathbf{x}) = s \sum_\alpha \kappa_\alpha^{(\eta)} \left(\frac{\mathbf{h}_\alpha^{(\eta)} \cdot \mathbf{x}}{L} \right)^p \tag{2.14}$$

This construction naturally produces islands of high fitness around the epitope-binding antibodies $\mathbf{h}^{(\eta)}$, separated by regions of low fitness. Our results are tied to the topology of fitness islands and not to details of the functions used to achieve them, provided genotypic space is of sufficiently high dimension. The specific mathematical choice of p has a limitation set by capacity; in a sequence space of dimension L, this method only allows us to program fewer than $\alpha L^{(p-1)}$ fitness islands. Beyond this 'capacity', there is a spin glass transition and the mathematical function above actually models a glassy landscape with many other fitness peaks. We can increase this capacity by increasing the non-linearity p of the model. In what follows, we choose p large enough to stay under this spin glass transition. Note that in the mathematical construction of this landscape , $-\mathbf{x}$ is as equally fit as $\mathbf{x}$. Such a degeneracy can be lifted by adding a linear term, $\sum_{i=1}^{L} h_i x_i$ to the fitness function. However, we found we did not need such a term since the degenerate pairs of fitness peaks, $\mathbf{x}$ and$-\mathbf{x}\}$, are far from each other in sequence space. Here, $\kappa_\alpha^{(\eta)}$ is the binding affinity of the ideal antibody $\mathbf{h}^{(\eta)}$ to its cognate epitope. We shall take epitope $\alpha = 1$ to be present for all antigens, thus

defining the generalist. In keeping with the assumption that the generalist is less fit in any landscape, we take its binding affinity $\kappa_1^{(1)} = \kappa_1^{(2)} = 0.8$ and all other $\kappa_\alpha^{(\eta)} = 1$.

Fitness penalty for generalists: We impose that the height of the fitness peak associated with the generalist is lower than the peaks of the specialist ($\frac{\kappa_1^{(\eta)}}{\kappa_{\alpha\neq 1}^{(\eta)}} = 0.8$), reflecting fitness costs associated with being a generalist relative to specialists in a fixed environment.

Population Simulations

We simulate a population of antibodies evolving in these landscapes by implementing a Moran process(108) with three events:

- **Environment shifts** with a deterministic rate, $\frac{1}{\tau_{\text{epoch}}}$
- **Mutation** with a rate, μ per individual, where a single site on $\mathbf{x}$ is bit-flipped
- **Reproduction** with a rate, λ per individual, where an individual is selected from the population with probability proportional to $\exp(F^{(\eta)})$, with $F^{(\eta)}$ defined above with $p = 2$ (Hopfield model), to be duplicated and another individual from the population to be removed with uniform probability

and a population size of N.

In our population simulations, we impose the following values for our simulation parameters. We fix the total number of epitopes for each landscape, $P_\eta = 11$ across all η, keeping just one generalist. We impose sequence length to be $L = 100$, generating each optimal epitope-binding antibody randomly. We initialize our simulations from a monoclonal specialist initial condition of size $N = 100$ unless otherwise specified. We impose a per individual mutation rate of $\mu = 0.25$ and a reproduction rate of $\lambda = 1$. The overall selection strength is set to $s = 0.1$.

Fig 4b: Resonance Peak in Generalist Discovery as a function of τ_{epoch}

We ran our simulation for 50 replicates from random monoclonal initial conditions of size $N = 100$ with sequences of length $L = 100$. We initialize the landscapes with 10 specialist antibodies and 1 generalist antibody (i.e. $P_\eta = 11$ for all η). Our simulations were run for a total of 30 τ_{epoch}s. We swept over τ_{epoch}s. We note that during the simulations, regardless of the frequency of environmental shifts, the population remains tightly clustered as it evolves in time. This behavior corresponds to evolution in the strong selection and weak mutation limit. We consider a generalist to have been discovered if, after a run, there exists at least one individual whose overlap with the generalist antibody is 90%. Using these simulations, we demonstrate that an intermediate regime of switching enhances the discovery of generalists. We plot the results of these simulations in Fig. 4b.

Time traces of the population from its initial condition: To identify the reason for this resonant peak, we run the simulation for a 100 cycles and plot in SI Fig 2.9 how far the population is from its initial condition in hamming distance after a fixed number of epochs We initialize a population at a specialist antibody and a generalist antibody, comparing the behavior for fast cycling and slow cycling. We show that for generalist initial conditions, regardless of cycling rate, the population remains in the generalist. There is some fluctuation out but the population returns to the generalists often. Given enough time, the population will escape though this is a slow process. However, for specialist initial condition, slow cycling enables the population to escape its initial condition, while fast cycling does not allow such escape.

2.5.3 Timescale Analysis

Computing the minimum epoch length, τ_{min}:

Cycling benefits evolution of generalists in fitness landscapes with valleys by enabling the population to escape specialist peaks in one landscape by evolving subject to a different

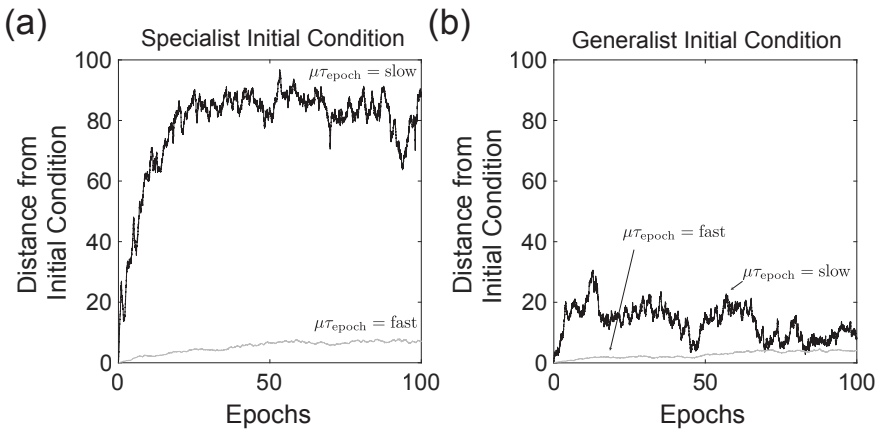

Fig 2.9: Specialist populations evolve significantly through sequence space for intermediate timescale cycling but not fast cycling; generalists do not evolve significantly for any timescale cycling. (a) An initially specialist population does not evolve away from the initial genotype for fast cycling. However, with slower cycling, the population evolves to a significantly different genotype(s). (b) An initially-generalist population does not significantly evolve away from the initial genotype for fast or slow cycling.

fitness landscape. To escape from a specialist peak in environment 1, the population must accrue enough mutations when subject to environment 2, such that when environment 1 returns again, the population is not likely to return to the original specialist peak.

We first consider the strong selection limit $sN \gg 1$. Consider a monoclonal population at a specialist peak $\mathbf{h^{(1)}}_\alpha$ in fitness landscape $F^{(1)}$. When such a population is now subject to landscape $F^{(2)}$ for a time τ_{epoch}, $\mathbf{h^{(1)}}_\alpha$ serves as an initial condition of typical low fitness and will evolve towards a fitness peak $\mathbf{h^{(2)}}_\beta$ in $F^{(2)}$. If we switch back to $F^{(1)}$ after a sufficiently long time, the population genotype $\mathbf{x}$ will be sufficiently mutated compared to $\mathbf{h^{(1)}}_\alpha$ that the population will likely fix to an alternative fitness peak $\mathbf{h^{(1)}}_\beta$ in $F^{(1)}$. Let us assume that the number of such mutations needed is d_{12}.

Since the population in genotype $\mathbf{h^{(1)}}_\alpha$ is typically of low fitness in landscape $F^{(2)}$, most mutations are beneficial. Then, in the strong selection limit, the time needed to acquire d_{12} beneficial mutations is by the mutation rate,

$\tau_{\min} \approx \frac{d_{12}}{\mu}$.

This minimal number of mutations d_{12} to escape the 'attractor basin' of a fitness peak $\mathbf{h^{(1)}}_\alpha$ is model dependent. - d_{12} depends on the size of the attractor basin around $\mathbf{h^{(1)}}_\alpha$ and the correlations between $F^{(1)}$ and $F^{(2)}$. In our Hopfield-inspired model of fitness landscapes $F^{(i)}$, if the fitness peaks are randomly distributed in a sequence space of length L, then the empirical value of $d_{12} \sim \frac{1}{4}L$. In real fitness landscapes, this distance d_{12} can vary widely for different specialists which can have attractor regions of different size.

Computing the maximum epoch length, $\tau_{\mathbf{max}}$:

Unnecessarily long times $\tau_{\text{epoch}} > \tau_{\text{min}}$ spent in each environment is counter-productive. To see this, note that specialists are most likely to evolve to generalists in a short duration of time after an environmental switch. Any extra time spent $\tau_{\text{epoch}} > \tau_{\text{min}}$ in the same environment is simply 'dead time' that does not increase the yield of generalists further. Hence the effective rate of evolving generalists from specialists falls as $1/\tau_{\text{epoch}}$ for $\tau_{\text{epoch}} > \tau_{\text{min}}$.

Meanwhile, existing generalists can specialize again. Let the rate of this process be $r_{g\to s}$. The yield of generalists is reduced when this escape rate $r_{g\to s}$ from generalists to specialists is larger than the switching-induced rate from specialists to generalists $1/\tau_{\text{epoch}}$. Hence, $\tau_{\text{max}} \sim 1/r_{g\to s}$.

The rate $r_{g\to s}$ at which generalists specialize is easily estimated since the fitness landscape does not change in time for sequences near the generalist. Hence this process is the well-studied process of an asexual population crossing a fitness valley by picking up a sequence of deleterious mutations. This process has been studied in numerous regimes with different assumptions about population sizes, selection pressure (178; 166; 64). Here, we assume strong selection and weak mutation, allowing us to use the simple result $r_{gs} \sim \mu\, e^{-N\Delta F_g}$ result obtained from the analogy of statistical physics and population dynamics; population size N plays the role of temperature and ΔF_g, the fitness difference between the generalist peak

and the fitness valley, plays the role of an energy barrier. Hence,

$$\tau_{\max} \approx \frac{\exp(N\Delta F_g)}{\mu}.$$

Real populations can often violate these assumptions; in that case, any other relevant result(64; 166; 178) for valley crossing rates can be used in place of $r_{g\to s}$.

Fig 4c: Transitions Amongst Specialists and Generalists: $\chi_{\text{s}\to\text{g}}$ and $\chi_{\text{g}\to\text{g}}$

The presence of a resonant peak in Fig. 4b is suggestive an underlying tension between discovering the generalist and escaping the generalist, similar to that in the earlier model of entropically disfavored generalists. As such, we re-introduce the quantities $\chi_{\text{s}\to\text{g}}$ and $\chi_{\text{g}\to\text{g}}$:

- $\chi_{\text{s}\to\text{g}}$ is the proportion of trials initialized from a monoclonal specialist initial condition that evolve a generalist (i.e. a single member of the population matches 90% of the generalist) antibody within 30 epochs of an evolutionary run for a given τ_{epoch}
- $\chi_{\text{g}\to\text{g}}$ is the number of trials in which a population, initialized from the generalist initial condition, maintains 20% of its population in the generalist (i.e. a given antibody maintains 90% overlap with $\mathbf{h}_1^{(\eta)}$) after 30 epochs for a given τ_{epoch}.

We chose 30 epochs in the definitions above as 30 epochs are needed to give the population enough time to accrue enough cycling-induced stochasticity to explore genotype space. This extension was not necessary for entropically disfavored generalists because multiple epochs are not needed to induce cycling induced stochasticity in such landscapes. We discuss cycling induced stochasticity in more detail in SI Section 2.5.3. Plots for $\chi_{\text{g}\to\text{g}}$ and $\chi_{\text{s}\to g}$ are shown in Fig. 4c and illustrate the same behavior as in the entropically disfavored models.

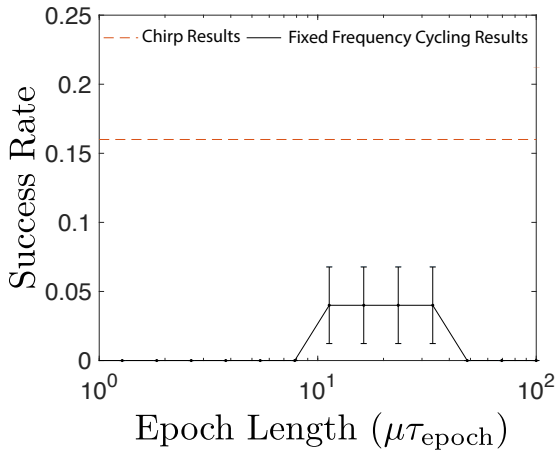

Fig 2.10: We compute the success rate of finding a generalist using fixed frequency cycling and chirped strategies when the fitness of binding a generalist is 70% of binding a specialist site. We see, in black, that fixed frequency cycling finds generalists with very low probability. However, in orange, we see that chirped protocols, where $\tau_{\text{epoch}}^{(n+1)} \leftarrow \frac{15}{16}\tau_{\text{epoch}}^{(n)}$ and $\mu\tau_{\text{epoch}}^{(0)} = 100$, find the generalists at a probability 0.16 ± 0.05, which is higher than the best fixed frequency cycling strategy. This demonstrates the success of chirped strategies
.

Chirping in Rugged Landscapes

Here, we demonstrate that chirp cycling provides benefits over fixed frequency cycling in fitness landscapes where peaks are separated by fitness valleys. We begin by changing the binding affinity of the antibody to the generalist site to $\kappa_1^{(1)} = \kappa_1^{(2)} = 0.7$. This results in the performance of fixed frequency cycling decreasing dramatically. Chirping, however, continues to provide generalists in a robust manner, as demonstrated in SI Figure 2.10.

Cycling-induced variance and correlations between environments

To illustrate how cycling enables the discovery of generalists, we consider population trajectories during cycling. We consider the impact of the initial condition of the population on these trajectories and the impact of the correlation structure between the different environments. To this end, we introduce a measure of correlation and introduce a new simulation to capture the effective behavior of the population.

Definition of Correlations between Environments We measure the correlation between landscapes, denoted $\langle F^{(1)}|F^{(2)}\rangle_s$ using the following equation:

$$\langle F^{(1)}|F^{(2)}\rangle_s \equiv \frac{c(\mathbf{h}^{(1)},\mathbf{h}^{(1)})}{c(\mathbf{h}^{(1)},\mathbf{h}^{(2)})c(\mathbf{h}^{(2)},\mathbf{h}^{(2)})} \tag{2.15}$$

where we have defined the function $c(\mathbf{h}^{(1)},\mathbf{h}^{(2)}) = \sum_{\alpha_1,\alpha_2=2}^{P_1,P_2} \mathbf{h}^{(1)}_{\alpha_1} \cdot \mathbf{h}^{(2)}_{\alpha_2}/(L\sqrt{P_1P_2})$.

This measure of correlation is high if the specialist genotypes for different antigens are highly similar in pairs; e.g., if each specialist for antigen 1 is similar to a specialist for antigen 2. As seen below, a high correlation by this measure implies that specialist antibodies do not evolve significantly due to cycling and thus generalists are not easily evolved.

Note that this measure is normalized so the measure is unaffected by the diversity $c(\mathbf{h}^{(\eta)},\mathbf{h}^{(\eta)})$ of specialist genotypes for a single antigen η.

Modeling Population Trajectories with Single Walkers To measure the role of cycling between landscapes and the correlation structure of the landscape, we studied the dynamics of single walkers. This is justified as the population is shown to be roughly monoclonal in its evolutionary trajectories. Single walkers were simulated via the well-known Metropolis-Hasting algorithm(101). We preserve definitions of $\mathbf{x}$, $F^{(\eta)}(\mathbf{x})$, and all related quantities from before. The process is as follows:

- Randomly select a single site to mutate to create new variant $\mathbf{x'}$ from original $\mathbf{x}$
- Compute fitness of new variant
- Accept new variant with probability $\exp(\beta(F^{(\eta)}(\mathbf{x'}) - F^{(\eta)}(\mathbf{x})))$ and repeat.

Because of the differences between single walker dynamics and population dynamics, we include an overall scale for the landscapes, β. β is chosen to be $\beta = 4$.

Fig 4f: Cycling-Induced stochasticity We begin by considering antigens with uncorrelated specialists (ie, $\langle F^{(1)}|F^{(2)}\rangle_s \approx 0$). Starting from two initial conditions, a generalist antibody and a specialist antibody for antigen $\eta = 1$, we evolve the walker for k proposals in the presence of antigen 2, and then allowed the walker enough proposals to relax to a stable solution in the presence of antigen 1. By computing the final state for 20 different walkers in a given landscape, and averaging over 20 random landscapes, we can compute the variance in the final positions of the walkers. This is accomplished by computing the average pairwise distance between walkers in the same landscape, and then averaging over landscapes. To demonstrate the importance of the number of proposals, k, which is serving as a proxy for τ_{epoch}, we swept over k. The result is plotted in Fig. 4f.

We see that when starting from a specialist initial condition, cycling-induced variance rises when τ_{epoch} is sufficiently large. Generalists, as predicted, are unaffected by cycling, as those genotypes are fit in both environments.

Fig 4g: Impact of Correlation Structures Between Cycled Environments We then repeated the same simulations as above with increasing correlation structure between the landscapes. We enforced that $k = 250$ was large enough to ensure high stochasticity in uncorrelated environments. We see that as correlation between the landscapes rises, cycling-induced stochasticity decreases. This indicates that generalist discovery hinges on the landscapes being sufficiently uncorrelated. The results are plotted in Fig. 4g.

More than 2 Antigens

Throughout the main text, we only consider the evolution of generalists in the presence of two antigens. Here, we demonstrate that when we increase the number of antigens, discovery of the generalist becomes easier. We demonstrate this in the Moran simulations by increasing the number of randomly constructed landscapes with a single generalist peak shared across

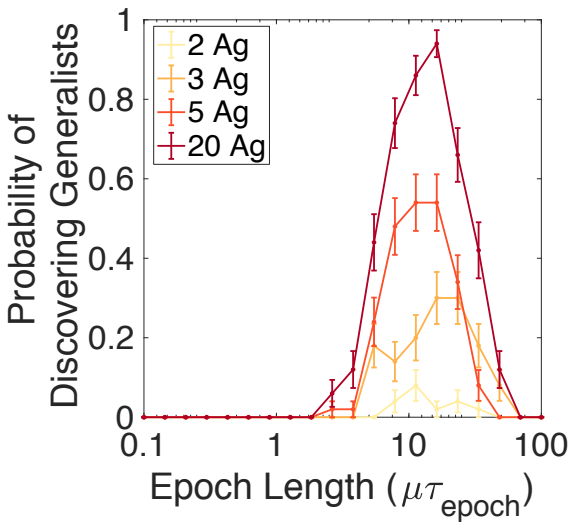

Fig 2.11: An increased number of distinct antigens makes it easier to evolve generalists through cycling. We increased the number of distinct antigens in stepsfrom 2 through 20, each with 10 randomly chosen specialist epitopes, and one generalist epitope common to all of them. We cycle between these landscapes using the Moran simulation described in Section 2.5.2 for 30 τ_{epoch}. We consider the evolutionary run to have evolved a generalist if at least one antibody has an overlap of 90% with the generalist. We find that increasing the number presented increases the likelihood that generalist genotypes are discovered.

the landscapes. In order to probe a dynamic range of antigen number, we weight the generalist to be smaller than in previous trials, setting $\kappa_1^{(\eta)} = 0.07$, rather than 0.08. We ensure each antigen is presented at least once, cycling for at least 30 epochs. We maintain the parameters used before.

We find that increasing the number of antigens increases the discovery-likelihood of generalists, as plotted in SI Fig 2.11. We interpret these results as due to effectively reduced correlation between landscapes since correlations shared across a subset of the antigens may not be shared across all antigens. For example, the population can settle into a limit cycle between specialists with only 3 antigens but evolution in the presence of other antigens allows escape from such cycles. As a result, we increase the rate of evolution from specialists to generalists without enhancing the reverse process.

We further consider the number of cycles needed for the first generalist to appear as a function of the number of antigens. This can be interpreted as the number of vaccine doses

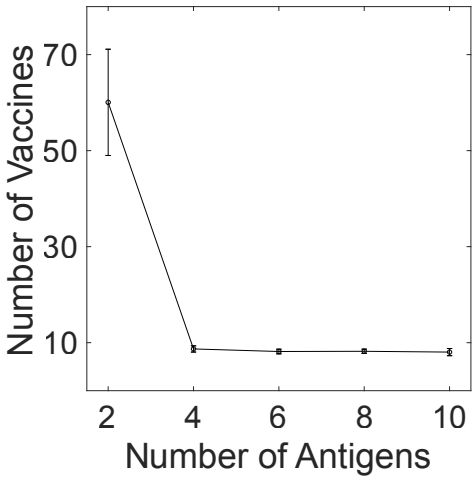

Fig 2.12: The number of vaccine shots required to evolve generalists is reduced if distinct antigens are used. We repeat the simulation described in SI Figure 2.11, except we run until a generalist is discovered for a fixed $\mu\tau_{\text{epoch}} = 40$. We compute the average first arrival time across 50 replicates and the standard error. We find that increasing the number of distinct antigens used decreases the number of doses needed up.

needed when using a particular number of strains in the vaccine course. We probe this by running Moran dynamics at a fixed $\mu\tau_{\text{epoch}} = 40$ with $\kappa_1^{(\eta)} = 0.08$, as in other simulations, for as many epochs as needed to discover a generalist. We run 50 replicates of this simulation, reporting the average time at which a generalist first appeared across those replicates. Our choice of epoch length is the epoch length at probability of evolving generalists appears to maximize. We find that the number of vaccine doses decreases as the number of antigen strains increases, as plotted in Fig. 2.12.

Molecular specificity

A critical requirement of antibodies, including broadly neutralizing antibodies, is molecular specificity. This is, antibodies must show higher binding affinity for their particular target and low binding affinities for all other antigens. We quantify molecular specificity of antibodies in our models by comparing the binding energies of antibodies to antigens featuring the conserved epitope to the binding energies of antibodies to antigens without the conserved

epitope. We use these comparisons to identify parameter regimes in which antibodies show molecular specificity; all analyses in the paper are carried out in such regimes.

Molecular Specificity in the Entropic Model We consider antibodies of length $L = 20$. We impose that the first 15 sites of this antibody bind to a conserved region on some set of antigens. We further impose that for the antibody to be considered to bind to an antigen, its binding energy, as given by Equation 2.3 to be below $\frac{T}{L} = -\frac{1}{2}$, which is to say that $-\frac{1}{L}\sum_i^L h_i x_i + \frac{T}{L}$ must be negative. We randomly generate 1000 antigens featuring the conserved eptiope and compute the binding energies of the antibody to the antigens. We then compare these binding energies to the binding energy of the same antibody against 1000 randomly generated antigens that are not obligated to feature the conserved epitope. We present the results of this in Fig. 2.13a, with the antigens featuring the conserved epitope in blue and the antigens without in black. We observe that while the antibody strongly binds all antigens featuring the conserved epitope, it only binds a small fraction of random antigens, showing molecular specificity in this parameter regime.

To ensure molecular specificity is achieved, we must ensure that the number of antigens to which an antibody binds must be small compared to the space of all antigens. We determine the choice of binding energy thresholds T that enforces molecular specificity by first stating that the fraction of antigens bound by a particular antibody is given by $\frac{\sum_{i=T_{\text{sites}}}^{L}\binom{L}{i}}{2^L}$, where the numerator represents the volume of the Hamming ball associated with the antigens that the antibody binds and the denominator represents the space of all antigens. We note that for $T \sim O(1)$, the volume of the Hamming ball is similar to that of the whole space, and for this choice of binding energy threshold, molecular specificity is not achieved. For $T \sim L$, the volume of the Hamming ball can be upper bounded by $\frac{L^L}{T^T}\left(\frac{1}{(L-T)^{L-T}}\right)$. This results in a vanishingly small fraction antigens being bound by our antibody when $T \sim L$. We work only in this regime.

Molecular Specificity in Landscapes with Barriers We begin by initializing an antibody of length $L = 100$ to bind to a conserved epitope. We then randomly construct 1000 antigens using the prescription described in Section 2.5.2 with 11 epitopes, 1 fixed across all antigens, and the remaining 10 random. We compute the binding energy of the antibody against these antigens. We then compare these binding energies to the binding energy of the antibody to 1000 antigens, each with 11 randomly generated epitopes. The results are plotted in Fig. 2.13b, with the binding energies associated with the conserved epitope in blue and the binding energies with random epitopes in black. We observe a large separation between the binding energies of the the antigens featuring the conserved epitope and the antigens without. In general, we expect that for conditions where the number of epitopes is below the Hopfield capacity(6), the probability that a random antigen is bound by an antibody that does not bind one of its epitopes to be vanishingly small.

HIV Antibody Data

The success of our proposed cycling strategy depends on specific assumptions about correlations between different antigens. In particular, antigens need to be sufficiently correlated in that the same generalist antibodies can bind them (e.g., the antigens share a epitope). And yet antigens need to be sufficiently *uncorrelated*: i.e., specialist antibodies that bind different antigens must be sufficiently distinct as measured by $\langle F^{(1)}|F^{(2)}\rangle_s$ (e.g., the specialist epitopes on antigens must be sufficiently distinct).

We sought to test whether these correlation conditions are met by antibodies evolved in response to real HIV strains.

Antibody sequences and Binding Affinity Data Several works have studied observed antibodies from individuals afflicted with different strains of HIV(25; 53; 84). These works sequenced the observed antibodies, studied their binding affinities to different strains, and

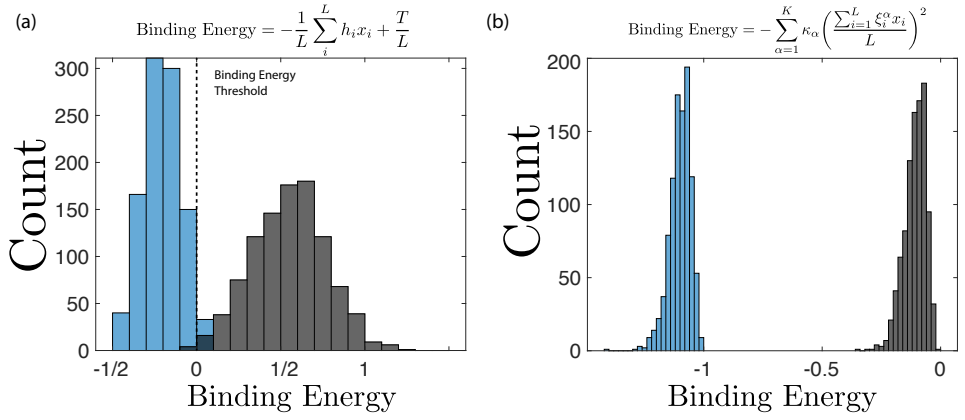

Fig 2.13: (a) We fix an antibody of length $L = 20$. We construct 1000 antigens, each of which with a conserved portion of length 15. We compute the binding energies of the antibody with each of these antigens and plot them in blue. We note that the binding energy for each of these falls below the binding energy threshold. We compare this to the binding energy of the same antibody to 1000 random antigens, plotted in black, and find it to be unlikely that the binding energy to fall below the threshold, indicating the antibody is unlikely to bind random antigens. (b) We fix an antibody of length $L = 100$ to bind to some conserved epitope. We generate 1000 antigens, each with 11 epitopes, 1 of which is conserved. We compute the binding energy of the antibody to these antigens and plot them in blue. We compare these binding energies to the binding energy of the same antibody to 1000 antigens, each with their own random 11 epitopes, plotted in black. We observe that a large separation in the binding energies, suggesting it is unlikely that an antibody will spontaneously bind a random antigen.

proposed intermediate antibodies in between the germline antibody and the discovered broadly neutralizing antibody. They evaluated the binding affinities of each of these antibodies using the ELISA assay. The binding affinity data is presented in SI Table 2.1. The mutational distance between each antibody is given in SI Table 2.2.

Antibody Sequence Data: Two classes of antibodies are presented here: mature antibodies observed in patients during their course with HIV and antibody sequences inferred to be (25; 53; 84) intermediate between the germline and the mature broadly neutralizing antibody. The natural antibodies appear with the prefix 'CH', and the inferred antibodies, which were synthesized, appear with the prefix 'IA'.

Antibody Binding Data: The binding affinity of each antibody to two different strains of HIV, 31D8gp120/293F and 11D8gp120, is evaluated using the ELISA assay. Particular values for binding are presented in table 2.1. We impose a cutoff of 10 log(AUC) to indicate when an antibody has bound an HIV strain. By this rule,

- 31D8gp120/293F is bound by antibodies IA2, IA3, CH105, and CH103.
- 11D8gp120 is bound by antibodies CH186, CH187, CH200, and CH103.

Constructing landscapes $F^{(1)}$ and $F^{(2)}$ Let $F^{(1)}$ and $F^{(2)}$ define the fitness landscape of antibody space corresponding to 31D8gp120/293F and 11D8gp120 respectively. In each landscape, the experimentally discovered and synthetically produced antibodies will define disconnected neighborhoods of antibodies that are fit for that landscape.

We begin constructing these landscapes by converting the sequence of each antibody into a binary vector with entries ± 1, noting that each antibody is length $L = 121$. We accomplish this randomly generating a binary vector with entries ± 1 of length L to represent the unmutated common ancestor. Then, using the sequence data given by (25; 84; 53), we determine where each antibody differs from the unmutated common ancestor and introduce a binary spin vector for each that preserve the differences from the unmutated common

HIV strain, Ab	CH105	CH186	CH187	CH200	IA2	IA3	CH103
31D8gp120/293F	13.52	1.13	0.00	5.80	13.34	13.01	13.63
11D8gp120	8.97	13.59	10.21	10.92	9.12	6.82	10.92

Table 2.1: Binding affinity of different antibodies (columns) to two different HIV strains (rows), measured via the ELISA assay (units of the logarithm of the area under the curve (logAUC) of the absorbance of the sample)(25; 53; 84). Higher values reflect stronger affinity. We consider an antibody to be a specialist for a strain using a cutoff of 10 logAUC. Note that only CH103 is a generalist in this dataset.

ancestor as presented in the real data. We define the sequences associated with $F^{(\eta)}$ as $\mathbf{h}_\alpha^{(\eta)}$. We set $\mathbf{h}_1^{(1)} = \mathbf{h}_1^{(2)}$ to the sequence of the generalist antibody CH103.

We take the fitness of each antibody, represented by $\mathbf{x}$ with entries ± 1 and length L, to be:

$$F^{(\eta)}(\mathbf{x}) = s \sum_\alpha \left(\frac{\mathbf{h}_\alpha^{(\eta)} \cdot \mathbf{x}}{L} \right)^p . \tag{2.16}$$

In the main text, we take $p = 10$ to stay below the spin glass transition for the sequences under consideration. s is a scalar that controls overall magnitude of fitness, which we take to be $s = 200$.

Simulations We simulated evolution using the technique described in Section 2.5.3. Given that mutation rates in B-cells undergoing somatic hypermutation are taken to be 10^{-3} per base pair per division(173), we choose our epoch length to be long enough that the population accumulates 100 mutations. This corresponds to an epoch length that allows 800 total divisions. The initial condition for these simulations was set to be the unmutated common ancestor (UCA). We note that if the population is not started from the UCA, the simulation fails to find successful antibodies.

Antibody	UCA	CH105	CH186	CH187	CH200	IA2	IA3	CH103
UCA	0	27	8	16	20	27	19	28
CH105	27	0	25	24	38	21	9	24
CH186	8	25	0	11	20	25	25	26
CH187	16	24	11	0	27	23	19	24
CH200	20	38	20	27	0	37	31	38
IA2	27	21	25	23	37	0	15	4
IA3	19	9	25	19	31	15	0	19
CH103	28	24	26	24	38	4	19	0

Table 2.2: Mutational distances (Hamming Distance) between antibody sequences for antibodies observed in an HIV patient who eventually developed bnAbs. Sequences for these antibodies are found in (25; 84; 53). Using the raw sequence data and the binding energy presented in 2.1, we can construct fitness landscapes $F^{(1)}$ and $F^{(2)}$ with fitness peaks that reflect these mutational distances.

Shuffled assignment We earlier demonstrated that the correlation structure of the landscape impacted the ability of the landscape to effectively cycle its way to a generalist. Here, we find that

$$\langle F^{(1)}|F^{(2)}\rangle_s = 0.43 \tag{2.17}$$

which reflects the distance between specialist sequences for the two strains in the data of (25; 53; 84) (see Table 2.2). With such low correlations between specialists, the simulations discover generalists around 60% of the time when cycling in this landscape, as shown in Fig. 5b.

To understand how increasing the correlation structure can impact generalist discovery in real data, we artificially shuffled the antibody binding data. In particular, we treated CH186 as a specialist antibody for 11D8gp120 and CH105 as a specialist antibody for 31D8gp120/293F and then followed the same construction of landscapes described in 2.5.3. In the new constrution, we find that the two things are substantially more correlated.

$$\langle F^{(1)}|F^{(2)}\rangle_s = 0.78 \tag{2.18}$$

Then, after running simulations with changing environments, we find that recovery rates of the generalist drops significantly, as shown in Fig. 5b. This demonstrates that our results are relevant when the fitness landscapes are sufficiently uncorrelated.

In Fig. 2.14, we repeat the simulation described above, but construct the patterns using only the antibody sites that are variable across the antibodies considered. This restriction reduces the length of the antibody sequences from $L = 121$ to $L = 47$. As a result, the correlation measure for the unshuffled landscape drops from 0.43 to 0.07, while the correlation measure for the shuffled landscape drops from 0.78 to 0.54. We set the epoch length to be longer than before, as a result of the differences in the magnitudes of correlations and the sequence length. The results of the simulation are shown in Fig.2.14.

Despite the resulting quantitative differences, qualitatively, these results are similar to those displayed in Figure 5 of the main text for the full $L = 121$ length sequences.

Thus, our conclusions primarily depend on the correlations between the fitness landscape and not the mathematical details of how we construct the fitness landscape. In particular, the dimensionality of sequence space affects our results to the extent that the dimensionality changes correlation structure across environments. We expect the effects of cycling to be weaker in lower dimensions, such as the case explored in (79) where there are fewer paths from specialists to generalists. For example, in 1 dimension, cycling can be entirely unproductive if cycling-induced evolution repeatedly traps the population at specialist peaks adjacent to the generalist genotype.

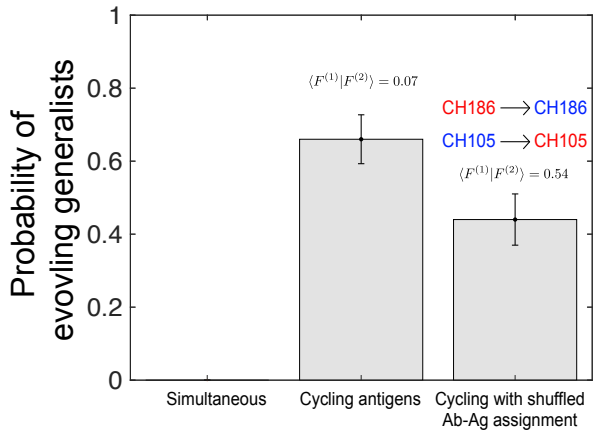

Fig 2.14: We construct a landscape only using sites along the antibodies which feature genotypic diversity, reducing the overall length of the antibodies from $L = 121$ to $L = 47$. We set $p = 10$ and $s = 200$ and repeat the simulation described in Section 2.5.3 and observe qualitatively similar results to those observed in Figure 5 of the main text.

CHAPTER 3

OPTIMAL PREDICTION WITH RESOURCE CONSTRAINTS USING THE INFORMATION BOTTLENECK

This work was done with Thierry Mora, Aleksandra Walczak, and Stephanie Palmer.

3.1 Abstract

Responding to stimuli requires that organisms encode information about the external world. Not all parts of the input are important for behavior, and resource limitations demand that signals be compressed. Prediction of the future input is widely beneficial in many biological systems. We compute the trade-offs between representing the past faithfully and predicting the future using the information bottleneck approach, for input dynamics with different levels of complexity. For motion prediction, we show that, depending on the parameters in the input dynamics, velocity or position information is more useful for accurate prediction. We show which motion representations are easiest to re-use for accurate prediction in other motion contexts, and identify and quantify those with the highest transferability. For non-Markovian dynamics, we explore the role of long-term memory in shaping the internal representation. Lastly, we show that prediction in evolutionary population dynamics is linked to clustering allele frequencies into non-overlapping memories.

3.2 Author summary

From catching a ball to building immunity, we rely on the ability of biological systems to incorporate past observations to make predictions about the future state of the environment. However, the success of these predictions is limited by environmental parameters and encoding capacities of the predictors. We explore these trade-offs in three systems: simple intertial motion, more complex motion with long-tailed temporal correlations, and mutating viral

strains. We show that the velocity and position of a moving object should not be equally well-remembered in the biological systems internal representation, and identify the flexible "best-compromise" representations that are not optimal but remain predictable in a wide range of parameters regimes. In the evolutionary context, we find that the optimal predictive representations are discrete, reminiscent of immune strategies that cover the space of potential viruses.

3.3 Introduction

How biological systems represent external stimuli is critical to their behavior. The efficient coding hypothesis, which states that neural systems extract as much information as possible from the external world, given basic encoding capacity constraints, has been successful in explaining some early sensory representations in the brain. Barlow suggested sensory circuits may reduce redundancy in the neural code and minimize metabolic costs for signal transmission (11; 80; 44; 123). However, not all external stimuli are as important to an organism, and behavioral and environmental constraints need to be integrated into this picture to more broadly characterize biological encoding. Delays in signal transduction in biological systems mean that predicting external stimuli efficiently can confer benefits to biological systems (18; 81; 134; 151), making prediction a general goal in biological sensing.

Evidence that representations constructed by sensory systems efficiently encode predictive information has been found in the visual and olfactory systems (135; 125; 188). Molecular networks have also been shown to be predictive of future states, suggesting prediction may be one of the fundamental principles of biological computation (98; 175). However, the coding capacity of biological systems is limited because they cannot provide arbitrarily high precision about their inputs: limited metabolic resources and other sources of internal noise impose finite-precision signal encoding. Given these trade-offs, one way to efficiently encode the history of an external stimulus is to keep only the information relevant for the prediction of

the future input (158; 3; 175). Here, we explore how optimal predictions might be encoded by neural and molecular systems using a variety of dynamical inputs that explore a range of temporal correlation structures. We solve the 'information bottleneck' problem in each of these scenarios and describe the optimal encoding structure in each case (158).

The information bottleneck framework , introduced by Tishby and colleagues (158; 31; 52; 148), allows us to define a 'relevance' variable in the encoded sensory stream. We take the relevant piece to be the future behavior of that input. Solving the bottleneck problem allows us to optimally estimate the future state of the external stimulus, given a certain amount of information retained about the past. In general, predicting the future coordinates of a system, $X_{t+\Delta t}$ reduces to knowing the precise historical coordinates of the stimulus X_t and an exact knowledge of the temporal correlations in the system. These rules and temporal correlations can be thought of as arising from two parts: a deterministic portion, described by a function of the previous coordinates, $\mathcal{H}(X_t)$, and the noise internal to the system, $\xi(t)$. Knowing the actual realization of the noise $\xi(t)$ reduces the prediction problem to simply integrating the stochastic equations of motion forward in time. If the exact realization of the noise if not known, we can still perform a probabalistic prediction by calculating the future form of the probability distribution of the variable X_t or its moments (54; 165). The higher-order moments yield an estimate of X_t and the uncertainty in the estimate. However, biological systems cannot precisely know X_t due to inherently limited readout precision (14; 16), creating a trade-off between representing the past and predicting the future.

We briefly summarize the information bottleneck method to quantify this trade-off here, and provide a more thorough explanation of the case with Gaussian statistics (reproduced from (31)) in S1 Text. The method assumes that the input variable, in our case the signal $X_{t-t_0:t}$, which considers measurements between times $t - t_0$ and t. We will call the past. This can be used to make inferences about the relevance variable, in our case the future

signal $X_{t+\Delta t:t+\Delta t+t_0}$, which considers measurements between times $t+\Delta t$ and $t+\Delta t+t_0$. We will call this the future. For convenience, in this introduction, we will take the past as a single point in time, X_t and the future as $X_{t+\Delta t}$. The resource constraints are introduced via a representation variable, $\tilde{X}$, which can have a varying amount of information about the input signal, X_t. This $\tilde{X}$, which has a dependence on the input, $\mathcal{P}(\tilde{X}|X_t)$, is constrained to be maximally informative of the future signal, subject to a constraint on $I(X_t;\tilde{X})$, the information it has about the past (Fig 3.1).

Formally, this representation is constructed by optimizing the objective function,

$$\min_{\mathcal{P}(\tilde{X}|X_t)} \mathcal{L}[\mathcal{P}(\tilde{X}|X_t)] = I(X_t;\tilde{X}) - \beta I(\tilde{X};X_{t+\Delta t}). \tag{3.1}$$

Each term is the mutual information between two variables: the first between the X_t and estimate of X_t given our representation model, $\tilde{X}$, and the second between $\tilde{X}$ and future input. The tradeoff parameter, β, controls how much future information we want $\tilde{X}$ to retain as it is maximally compressed. For large β, $\tilde{X}$ must be maximally informative about $X_{t+\Delta t}$, and will have, in general, the lowest compression. Small β means less information is retained about the future and high, lossy compression is allowed.

The causal relationship between X_t and $X_{t+\Delta t}$ results in a data processing inequality, $I(X_t;\tilde{X}) \geq I(X_{t+\Delta t};\tilde{X})$, meaning that the information generated about $X_{t+\Delta t}$ cannot exceed the amount encoded about X_t (13). Additionally, the information about X_t that the representation can extract is bounded by the amount of information X_t, itself, contains about the $X_{t+\Delta t}$, $I(\tilde{X};X_{t+\Delta t}) \leq I(X_t;X_{t+\Delta t})$.

We use this framework to study how biological systems can optimally encode external stimuli for downstream decoding, but without any explicit constraints on or specification of that decoder. Here, we assume that the compressed representation variable has a one-time-step output and only has access to a fixed amount of historical information about the stimulus. Here, we assume that the compressed representation variable has a single 'present'

time-step output and only has access to a fixed amount of historical information about the stimulus. This reflects, for example, the instantaneous neural output from a retinal ganglion cell population that is passed downstream to the cortex for further processing and readout. We start with a one-time-step past input and then extend this to a longer temporal window into the past. We begin by assuming a one-time-step past input and then later extend it to a more extended temporal window in the past. The optimal predictive encoder does in general favor some aspects of this past information (position information) over others (velocity information). A downstream decoder may be able to recover some of the lower priority information by combining measurements and predictions across time to reduce variance post hoc, but the gain in precision comes at the cost of additional constraints on the size and complexity of the encoded representation variable. In addition, the gained information about the stimulus that was originally discarded may not provide significant predictive advantages. We do, however, provide a comparison between our information bottleneck framework and the results of a model that performs this kind of prediction combined with measurement and error estimates across time in Section D in S2 Text. There we demonstrate that for a given level of $I(X_t; \tilde{X})$, a Kalman filter achieves lower $I(\tilde{X}; X_{t+\Delta t})$. A question we do not explore here is how to, practically, read out the optimally encoded representation. It has been shown previously that simple perceptrons can read out predictive information from the retinal code(147), which makes biologically plausible readout possible and is a direction of future work.

We use information bottleneck to compute the optimal predictive encoding in two well-studied dynamical systems with ties to biological data: the stochastically-driven damped harmonic oscillator (SDDHO) and the Wright-Fisher model. We look at these two different systems to gain intuition about how different types of dynamics influence the ability of a finite and noisy system to make accurate predictions. We further consider two types of SDDHO processes to study the effects of noise correlations on prediction. Our exploration of

the SDDHO system has a two-fold motivation: it is a physical system that describes motion that a visual system might need to process and predict to catch prey or evade predators. It is also the simplest possible continuous stochastic system whose full dynamics can be solved exactly. Previous studies used the SDDHO process to create moving bar stimuli and quantify retinal prediction (125; 147; 143). Prediction of a time series with Markovian dynamics is not limited to physical motion, of course. The Wright-Fisher model (180) is a canonical model of evolution (157) which has been used to consider how the adaptive immune system predicts the future state of the pathogenic environment (96; 98). Resource constraints also create trade-offs between representation precision and prediction in the immune system, and finding the general principles that connect prediction in these two contexts can reveal common principles across biological systems and scales.

The results of these information bottleneck calculations in these different dynamical contexts will reveal the form and content of optimally predictive features. These features are matched both to the input parameters and to the level of resource constraints that compress the input. Our results form expectations about what to find in biological systems when the internal representation can be measured (e.g. as in (125)), and the input statistics match the kinds of dynamics studied here. While our results will show what types of feature extraction are expected in systems predicting their inputs optimally, not all systems may be optimized for a broad range of input dynamics. In fact, we assume that natural selection favors encodings that confer just enough predictive capacity to support the organism's behavioral repertoire. That might mean flexibly predicting in many different environments either over an individual or group migratory lifespan. To help quantify the 'transferability' of any optimally predictive encoding scheme, we will develop a metric, Q, that tracks how well one representation performs under other input dynamics, where it might not be the absolute optimal, but still performs well. Of course, we only expect our maximally predictive encodings to match biological filters when the system has an intrinsic behavioral goal that requires prediction.

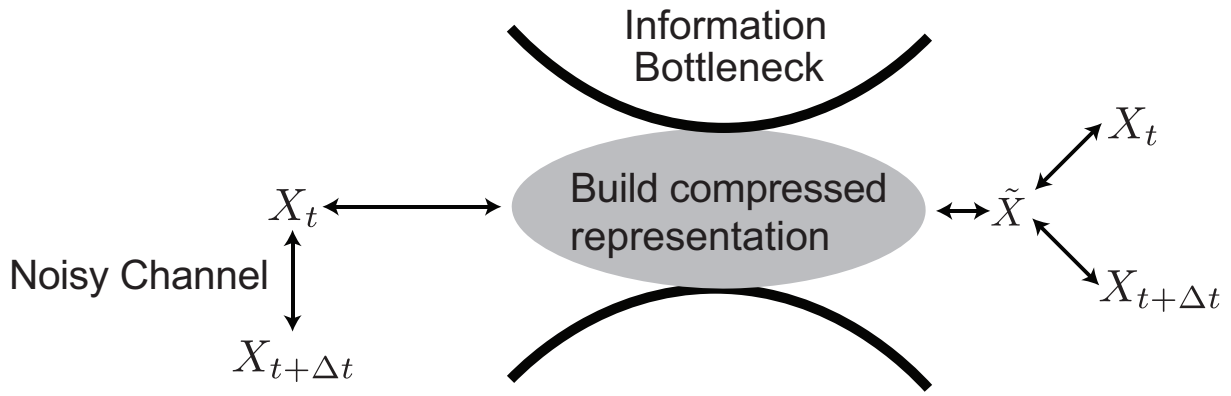

Fig 3.1: A schematic representation our predictive information bottleneck. On the left hand side, we have coordinates X_t evolving in time, subject to noise to give $X_{t+\Delta t}$. We construct a representation, $\tilde{X}$, that compresses the X_t (minimizes $I(X_t;\tilde{X})$) while retaining as much information about $X_{t+\Delta t}$ (maximizes $I(\tilde{X};X_{t+\Delta t})$) up to the weighting of the prediction compared to the compression set by β.

There are computations that do not require prediction, and would presumably result from constraints that prioritize other types of information in the input.

3.4 Results

3.4.1 The Stochastically Driven Damped Harmonic Oscillator

Previous work explored the ability of the retina to construct an optimally predictive internal representation of a dynamic stimulus. Palmer et al (125) recorded the response of a salamander retina to a moving bar stimulus with SDDHO dynamics. In this case, the spike trains in the retina encode information about the past stimuli in a near-optimally predictive way (125). In order for optimal prediction to be possible, the retina should encode the position and velocity as dictated by the information bottleneck solution to the problem, for the retina's given level of compression of the visual input. In that study, the SDDHO was set near critical damping, and only one set of parameters in the model was shown to the

retina. Inspired by this experiment, we explore the optimal predictive encoding schemes as a function of the parameters in the dynamics, and we describe the optimal solution across the entire parameter space of the model, over a wide range of desired prediction timescales.

We consider the dynamics of a mass m in a viscous medium attached to a spring receiving noisy velocity kicks generated by a temporally uncorrelated Gaussian process, as depicted in Fig 3.2A. The dynamics of this model were solved previously(119). See Section A in S2 Text for details. Equations of motion are introduced in terms of physical variables $\bar{x}$, $\bar{v}$, and $\bar{t}$ (bars will be dropped later when referring to rescaled variables), which evolve according to

$$\begin{aligned} m\frac{d\bar{v}}{d\bar{t}} &= -k\bar{x} - \Gamma\bar{v} + (2k_BT\Gamma)^{1/2}\xi(\bar{t}), \\ \frac{d\bar{x}}{d\bar{t}} &= \bar{v}, \end{aligned} \tag{3.2}$$

where k is the spring constant, Γ the damping parameter, k_B the Boltzmann constant, T temperature, $\langle\xi(\bar{t})\rangle = 0$, and $\langle\xi(\bar{t})\xi(\bar{t}')\rangle = \delta(\bar{t} - \bar{t}')$. We rewrite the equation with $\omega_0 = \sqrt{\frac{k}{m}}$, $\tau = \frac{m}{\Gamma}$, and $D = \frac{k_BT}{\Gamma}$.We also introduce a dimensionless parameter, the damping coefficient, $\zeta = 1/(2\omega_0\tau)$. When $\zeta < 1$, the system is underdamped and the motion of the mass will be oscillatory. When $\zeta \geq 1$, the system is overdamped and the motion will be non-oscillatory. Additionally, the equipartition theorem tells us that $\langle\bar{x}(\bar{t})^2\rangle \equiv x_0^2 = k_BT/k = D/(\tau\omega_0^2)$. Putting this all together, we obtain

$$\frac{d\bar{v}}{d\bar{t}} = -\frac{\bar{x}}{4\tau^2\zeta^2} - \frac{\bar{v}}{\tau} + \frac{x_0}{\sqrt{2\tau^3}\zeta}\xi(\bar{t}) \tag{3.3}$$

We make two changes of variable to further simplify our expressions. We set $t = \frac{\bar{t}}{\tau}$ and $x = \frac{\bar{x}}{x_0}$. We also define a rescaled velocity, $\frac{dx}{dt} = v$, so that our equation of motion now reads

$$\frac{dv}{dt} = -\frac{x}{4\zeta^2} - v + \frac{\xi(t)}{\sqrt{2}\zeta}. \tag{3.4}$$

There are now just two parameters that govern a particular solution to our information bottleneck problem: ζ and Δt, the timescale on which we want to retain optimal information about the future. We define $X_t = (x(t), v(t))$ and $X_{t+\Delta t} = (x(t+\Delta t), v(t+\Delta t))$ and seek a representation, $\tilde{X}(\zeta, \Delta t)$, that can provide a maximum amount of information about $X_{t+\Delta t}$ for a fixed amount of information about X_t. By considering position and velocity, our system is Markovian, so extended temporal windows provide no additional information. If we were to ignore velocity in this model, estimates of the future would become suboptimal to the information bottleneck bound. We explore models where extended temporal windows are relevant in Section 3.4.2. To construct the information bottleneck solution in the case with Gaussian variables, we follow the construction given in (31). We note that due to the Gaussian structure of the joint distribution of X_t and $X_{t+\Delta t}$ for the SDDHO, the problem can be solved analytically. The optimal compressed representation is a noisy linear transform of X_t (see S1 Text) (31),

$$\tilde{X} = A_\beta X_t + \xi. \tag{3.5}$$

A_β is a matrix whose elements are a function of β, the tradeoff parameter in the information bottleneck objective function, and the statistics of the input and output variables. The added noise term, ξ, has the same dimensions as X_t and is a Gaussian variable with zero mean and unit variance.

We calculate the optimal compression, $\tilde{X}$, and its predictive information (see Section B in S2 Text). The coordinates at time t and time $t + \Delta t$ in the SDDHO bottleneck problem are jointly Gaussian, which means that the optimal compression can be fully described by its first and second-order statistics. We generalize analytically the results that were numerically obtained in Ref. (125) and explore the full parameter space of this dynamical model and examine all predictive bottleneck solutions, including different desired prediction timescales.

We quantify the efficiency of the representation $\tilde{X}$ in terms of the variance of the following four probability distributions: the prior distribution, $\mathcal{P}(X_t)$, the distribution of

X_t conditioned on the compression, $\mathcal{P}(X_t|\tilde{X})$, the distribution of $X_{t+\Delta t}$ conditioned on the compressed variable $\mathcal{P}(X_{t+\Delta t}|\tilde{X})$, and the distribution of $X_{t+\Delta t}$ conditioned on X_t $\mathcal{P}(X_{t+\Delta t}|X_t)$. We represent the uncertainty reduction, or the mutual information between these two variables, using two dimensional contour plots that depict the variances of the distributions in the $((x - \langle x \rangle)/\sigma_x, (v - \langle v \rangle)/\sigma_v)$ plane, where σ_x and σ_v are the standard deviations of the signal distribution $\mathcal{P}(X_t)$. We present example distributions of $\mathcal{P}(X_t|\tilde{X})$ and $\mathcal{P}(X_{t+\Delta t}|\tilde{X})$ in Fig 3.2B (left, right, respectively).

The representation, $\tilde{X}$, will be at most two-dimensional, with each of its components corresponding to linear combinations of position and velocity. It may be lower dimensional for certain values of β. The smallest critical β for which the representation remains two-dimensional is given in terms of the smallest eigenvalue of the matrix $\Sigma_{X_t|X_{t+\Delta t}}\Sigma_{X_t}^{-1}$ as $\beta_c = 1/(1 - \min\{\lambda_1, \lambda_2\})$ (see Section B in S2 Text). $\Sigma_{X_t|X_{t+\Delta t}}$ is the covariance matrix of the probability distribution of $\mathcal{P}(X_t|X_{t+\Delta t})$ and Σ_{X_t} is the input variance. Below this critical β, the compressed representation is one dimensional, $\tilde{X} = k_1 x + k_2 v + \text{noise}$, but it is still a combination of position and velocity.

Limiting cases along the information bottleneck curve help build intuition about the optimal compression. If $\tilde{X}$ provides no information about the stimulus (e.g. $\beta = 0$), the variances of both of the conditional distributions match that of the prior distribution, $\mathcal{P}(X_t)$, which is depicted as a circle of radius 1 (blue circle in Fig 3.2C). However, if the encoding contains information about X_t, the variance of $\mathcal{P}(X_t|\tilde{X})$ will be reduced compared to the prior. The maximal amount of predictive information, which is reached when $\beta \to \infty$, can be visualized by examining the variance of $\mathcal{P}(X_{t+\Delta t}|X_t)$ (e.g. the purple contour in Fig 3.2C), which quantifies the correlations in X, itself, with no compression. Regardless of how precisely the current state of the stimulus is measured, the uncertainty about the future stimulus cannot be reduced below this minimal variance, because of the noise in the equation of motion.

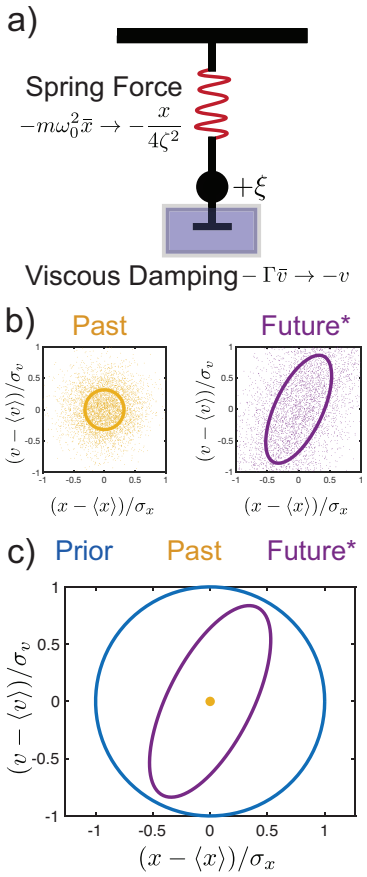

Fig 3.2: Schematic of the stochastically driven damped harmonic oscillator (SDDHO). (a) The SDDHO consists of a mass attached to a spring undergoing viscous damping and experiencing Gaussian thermal noise of magnitude. There are two parameters to be explored in this model: $\zeta = \frac{1}{2\omega_0\tau}$ and $\Delta t = \frac{\Delta t}{\tau}$. (b) $\zeta = \frac{1}{2}$, $\Delta t = 1$. Here, we show an example distribution of the history (yellow, left) and show its time evolution (purple, right). We take 5000 samples from the distribution, at random, and let these points evolve in time according to the SDDHO equation of motion. We visualize the evolution of the distribution of points in time via an ellipse representing the 1-Σ confidence region of the rescaled position and velocity. (c) We illustrate the limiting case of the information bottleneck method when $\beta \to \infty$. Representations of the past and how that constrains an estimate of the future position and velocity of the object can be compared to the prior be examining the relative size and shape of their respective ellipses. The blue circle represents the prior and its 1-Σ confidence region. In yellow, we plot the inferred 1-Σ confidence interval associated with the estimate of past, X_t, given by the encoding distribution when $\beta \to \infty$. In this limit, the distribution is reduced to a single point. In purple, we plot the 1-Σ confidence region of $X_{t+\Delta t}$ given our knowledge of X_t. Precise knowledge of the past coordinates reduces the our uncertainty about the future position and velocity (as compared to the prior), as depicted by the smaller area of the purple ellipse.

From Fig 3.2B, we see that the conditional distribution $\mathcal{P}(X_{t+\Delta t}|X_t)$ is strongly compressed in the position coordinate with some compression in the velocity coordinate. The information bottleneck solution at a fixed compression level (e.g. $I(X_t;\tilde{X}) = 1$), shown in Fig 3.3A (left), gives an optimal encoding strategy for prediction (yellow curve) that reduces uncertainty in the position variable. This yields as much predictive information, $I(X_{t+\Delta t};\tilde{X})$, as possible for this value of $I(X_t;\tilde{X})$. The uncertainty of the prediction is illustrated by the purple curve. We can explore the full range of compression levels, tracing out an information bottleneck curve for this damping coefficient and desired prediction timescale, as shown in Fig 3.3. Velocity uncertainty in the compressed representation is only reduced (i.e. predictive information that uses past velocity estimates is only useful) as we allow for less compression, as shown in Fig 3.3A (right). For both of the cases represented in Fig 3.3A, the illustrated encoding strategy yields a maximal amount of mutual information between the compressed representation, $\tilde{X}$, and the future for the given level of compression, as indicated by the red dots in Fig 3.3B.

As noted above, there is a phase transition along the information bottleneck curve, where the optimal, predictive compression of X_t changes from a one-dimensional representation to a two-dimensional one. This phase transition can be pinpointed in β for each choice of ζ and Δt, and can be determined using the procedure described in is given in the S1 Text. To understand which directions are most important to represent at high levels of compression, we derive the analytic form of the leading eigenvector, w_1, of the matrix $\Sigma_{X_t|X_{t+\Delta t}}\Sigma_{X_t}^{-1}$. We have defined $\omega^2 = \frac{1}{4\zeta^2} - \frac{1}{4}$ such that

$$w_1 = \begin{bmatrix} \omega\cot(\omega\Delta t) + \frac{|\csc(\omega\Delta t)|}{2\sqrt{2}\zeta}\sqrt{2-\zeta^2-\zeta^2\cos(2\omega\Delta t)} \\ 1 \end{bmatrix}. \tag{3.6}$$

The angle of the encoding vector from the position direction is then given by

$$\phi = \arctan\left(\left(\omega\cot(\omega\Delta t) + \frac{|\csc(\omega\Delta t)|}{2\sqrt{2}\zeta}\sqrt{2-\zeta^2-\zeta^2\cos(2\omega\Delta t)}\right)^{-1}\right). \tag{3.7}$$

We consider ϕ in three limits: (I) the small Δt limit, (II) the strongly overdamped limit ($\zeta \to \infty$), and (III) the strongly underdamped limit ($\zeta \to 0$).

(I): When $\omega\Delta t \ll 1$, the angle can be expressed as

$$\phi = \arctan\left(\frac{\Delta t}{1+\omega^2}\right). \tag{3.8}$$

This suggests that for small $\omega\Delta t$, the optimal encoding scheme favors position information over velocity information. The change in angle of the orientation from the position axis in this limit goes as $O(\Delta t)$.

(II): The strongly overdamped limit. In this limit, ϕ becomes

$$\phi = \arctan\left(\frac{2\sinh(\frac{\Delta t}{2})}{\cosh(\frac{\Delta t}{2}) + \sqrt{\frac{1+\cosh(\Delta t)}{2}}}\right). \tag{3.9}$$

In the large Δt limit, $\phi \to \frac{\pi}{4}$. In the small Δt limit, $\phi \to \arctan(\Delta t)$. Position information is the best predictor of the future input at short lags, which velocity and position require equally fine representation for prediction at longer lags.

(III) The strongly underdamped limit. In this limit, ϕ can be written as

$$\phi = \arctan\left(\frac{2\zeta\sin(\frac{\Delta t}{2\zeta})}{\cos(\frac{\Delta t}{2\zeta}) + \sqrt{2-\zeta^2-\zeta^2\cos(\frac{\Delta t}{\zeta})}}\right). \tag{3.10}$$

We observe periodicity in the optimal encoding angle between position and velocity. This means that the optimal tradeoff between representing position or velocity depends on the timescale of prediction. However, the denominator never approaches 0, so the encoding scheme

never favors pure velocity encoding. It returns to position-only encoding when $\Delta t/2\zeta = n\pi$.

At large compression values, i.e. small amounts of information about X_t, the information bottleneck curve is approximately linear. The slope of the information bottleneck curve at small $I(X_t;\tilde{X})$ is given by $1-\lambda_1$, where λ_1 is the smallest eigenvalue of the matrix, $\Sigma_{X_t|X_{t+\Delta t}}\Sigma_{X_t}^{-1}$. The value of the slope is

$$1-\lambda_1 = \exp(-\Delta t)(\frac{1}{4\omega^2\zeta^2} + \frac{\cos(2\omega\Delta t)}{4\omega^2} + \frac{|\sin(\omega\Delta t)|}{2\sqrt{2}\omega^2\zeta}\sqrt{2-\zeta^2-\zeta^2\cos(2\omega\Delta t)}). \quad (3.11)$$

For large Δt, it is clear that the slope will be constrained by the exponential term, and the information will fall as $\exp(-\Delta t)$ as we attempt to predict farther into the future. For small Δt, however, we see that the slope goes as $1-\Delta t^2$, and our predictive information decays more slowly.

For vanishingly small compression, i.e. $\beta \to \infty$, the predictive information that can be extracted by $\tilde{X}$ approaches the limit set by the temporal correlations in X, itself, given by

$$I(X_t; X_{t+\Delta t}) = \frac{1}{2}\log(|\Sigma_{X_t}|) - \frac{1}{2}\log(|\Sigma_{X_t|X_{t+\Delta t}}|). \quad (3.12)$$

For large Δt, this expression becomes

$$I(X_t; X_{t+\Delta t}) \propto \exp(-\Delta t). \quad (3.13)$$

For small Δt,

$$I(X_t; X_{t+\Delta t}) \propto \Delta t - \frac{1}{2}\log(\Delta t). \quad (3.14)$$

The constant term emerges from the physical parameters of the input dynamics.

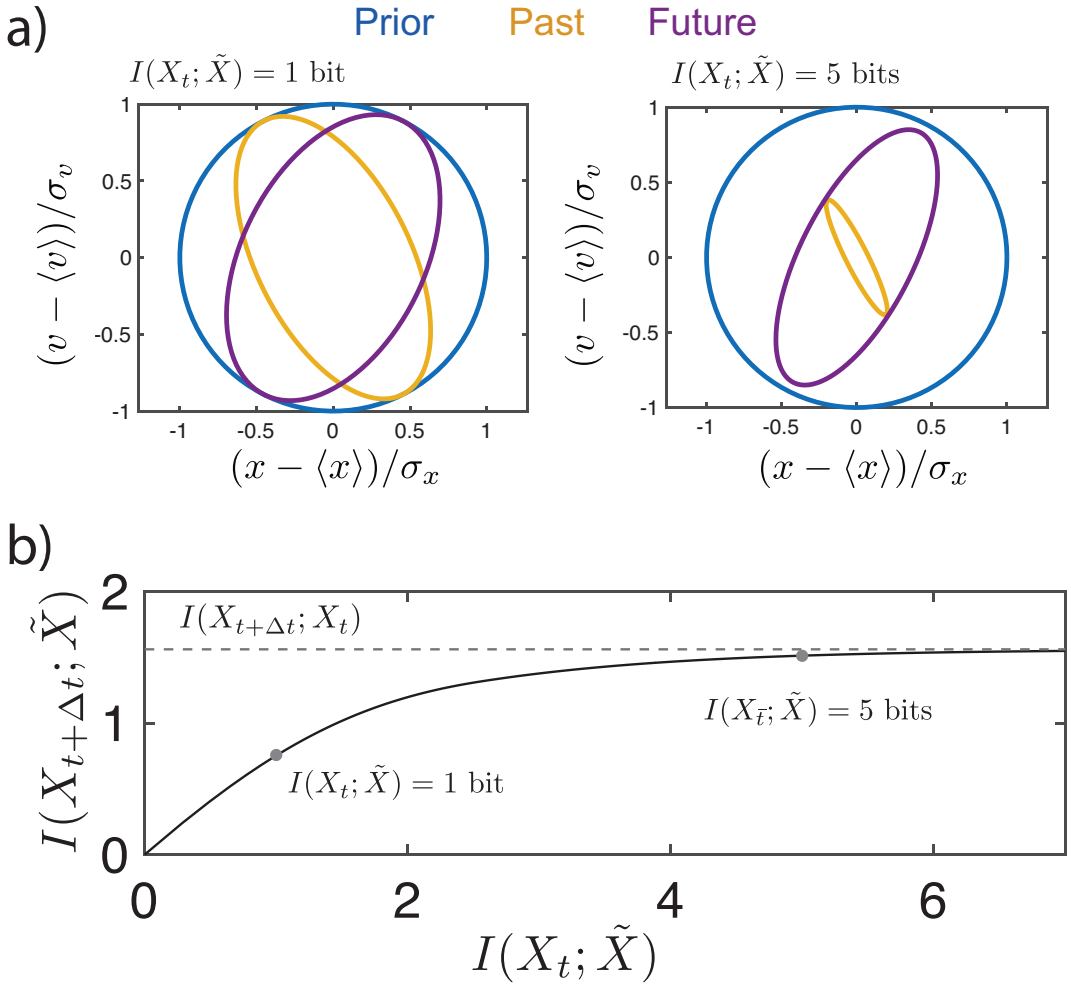

Fig 3.3: We consider the task of predicting the path of an SDDHO with $\zeta = \frac{1}{2}$ and $\Delta t = 1$. (a) (left) We encode the history of the stimulus, X_t, with a representation generated by the information bottleneck, $\tilde{X}$, that can store 1 bit of information. Knowledge of the coordinates in the compressed representation space enables us reduce our uncertainty about the bar's position and velocity, with a confidence interval given by ellipse in yellow. This particular choice of encoding scheme enables us to predict the future, $X_{t+\Delta t}$ with a confidence interval given by the purple ellipse. The information bottleneck guarantees this uncertainty in future prediction is minimal for a given level of encoding. (right) The uncertainty in the prediction of the future can be reduced by reducing the overall level of uncertainty in the encoding of the history, as demonstrated by increasing the amount of information $\tilde{X}$ can store about X_t. However, the uncertainty in the future prediction cannot be reduced below the variance of the propagator function. (b) We show how the information with $X_{t+\Delta t}$scales with the information about X_t, highlighting the points represented in panel A.

Optimal representations in all parameter regimes for fixed $I(X_t; \tilde{X})$

We sweep over all possible parameter regimes of the SDDHO keeping $I(X_t; \tilde{X})$ fixed at 5 bits and find the optimal representation for a variety of timescales (Fig 3.4), keeping a fixed amount of information encoded about X_t for each realization of the stimulus and prediction. More information can be transmitted for shorter delays (Fig 3.4A, 3.4D, and 3.4G) between the X_t and $X_{t+\Delta t}$ signal than for longer delays (Fig 3.4C, 3.4F, and 3.4I). In addition, at shorter prediction timescales more information about X_t is needed to reach the upper bound, as more information can be gleaned about the future. In particular, for an overdamped SDDHO at short timescales (Fig 3.4A), the evolution of the equations of motion are well approximated by integrating Eq. 3.3 with the left hand side set to zero, and the optimal representation encodes mostly position information. This can be visualized by noting that the encoding ellipse remains on-axis and mostly compressed along the position dimension. For the underdamped case, in short time predictions (Fig 3.4G), a similar strategy is effective. However, for longer predictions (Fig 3.4H and 3.4I), inertial effects cause position at one time to be strongly predictive of future velocity and vice versa. As a result, the encoding distribution has to take advantage of these correlations to be optimally predictive. These effects can be observed in the rotation of the encoding ellipse, as it indicates that the uncertainty in position-velocity correlated directions are being reduced, at some cost to position and velocity encoding. The critically damped SDDHO (Fig 3.4D-F) demonstrates rapid loss of information about the future, like that observed in the underdamped case. The critically damped case displays a bias towards encoding position over velocity information at both long and intermediate timescales, as in the overdamped case. At long timescales, Fig 3.4F, the optimal encoding is non-predictive.

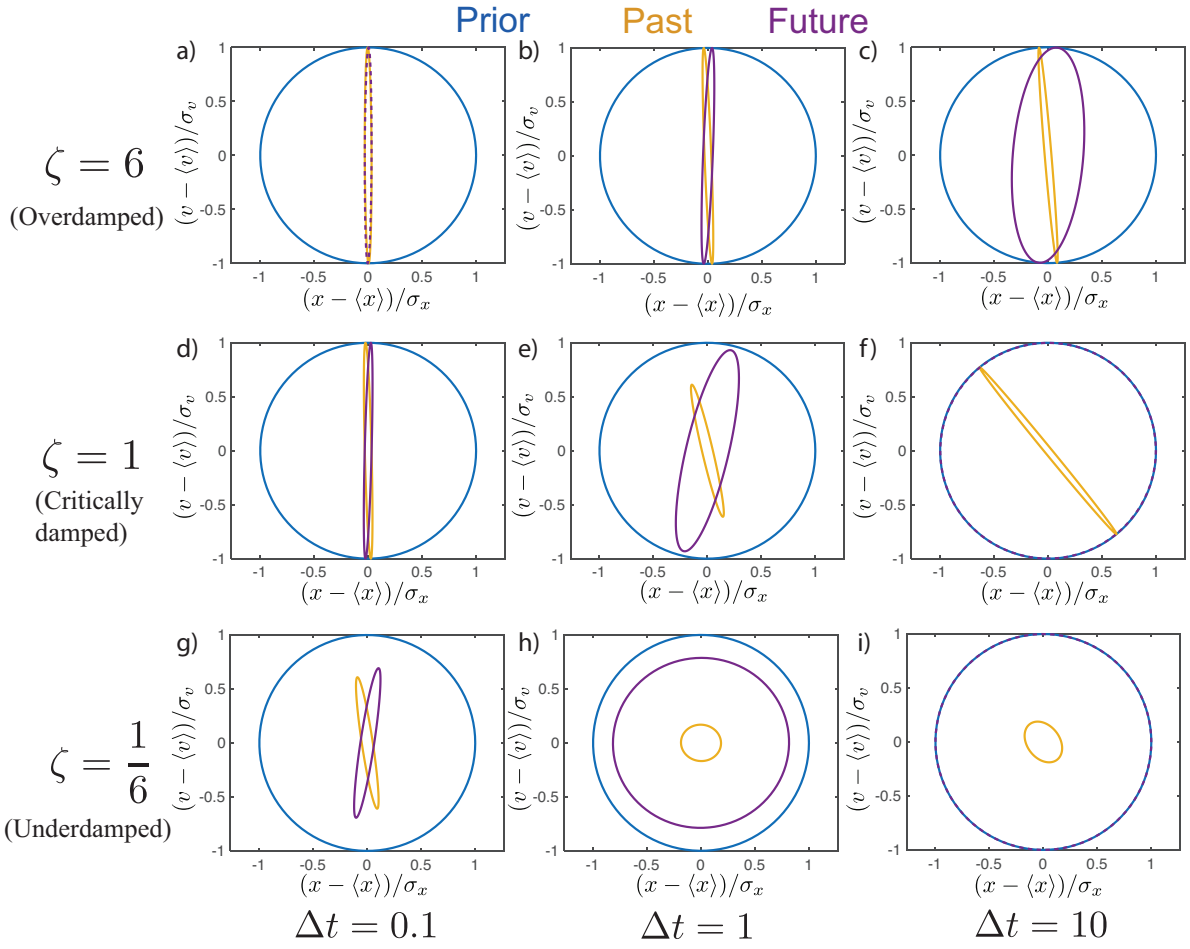

Fig 3.4: Possible behaviors associated for the SDDHO for a variety of timescales with a fixed $I(X_t; \tilde{X})$ of 5 bits. For an overdamped SDDHO, panel a-c, the optimal representation continues to encode mostly position information, as velocity is hard to predict. For the underdamped case, panels g-i, as the timescale of prediction increases, the optimal representation changes from being mostly position information to being a mix of position and velocity information. Optimal representations for critically damped input motion are shown in panels d-f. Comparatively, overdamped stimuli do not require precise velocity measurements, even at long timescales. Optimal predictive representations of overdamped input dynamics have higher amounts of predictive information for longer timescales, when compared to underdamped and critically damped cases.

Suboptimal representations

Biological systems might not adapt to each input regime perfectly, nor may they be optimally efficient for every possible kind of input dynamics. We consider what happens when an optimal representation is changed, necessarily making it suboptimal for predicting the future stimulus. We construct a new representation by rotating the optimal solution in the position, velocity plane. We examine the conditional distributions for this suboptimal representation, both about X_t, $\mathcal{P}(X_t|\tilde{X}_{\text{suboptimal}})$, and the future, $\mathcal{P}(X_{t+\Delta t}|\tilde{X}_{\text{suboptimal}})$. For a fixed amount of information about X_t, $I(X_t;\tilde{X}_{\text{optimal}}) = I(X_t,\tilde{X}_{\text{suboptimal}})$, we compare the predictive information in the optimal (Fig 3.5A) and the suboptimal representations (Fig 3.5B). We examine the choice of parameters in the stimulus dynamics for which encoding position alone is an optimal strategy. We note that encoding velocity with high certainty provides very little predictive power, indicating that encoding velocity and position is not equally important, even for equal compression levels. While the nature of the suboptimal and optimal representations depend on the input dynamics, we see that the encoding schemes discovered by the information bottleneck are, indeed, optimally predictive.

Transferability of a representation

So far, we have described the form that optimal predictive compressions take along the information bottleneck curve for a given ζ and Δt. How do these representations translate when applied to other prediction timescales (i.e. can the optimal predictive scheme for near-term predictions help generate long-term predictions, too?) or other parameter regimes of the model? This may be important if the underlying parameters in the external stimulus are changing rapidly in comparison to the adaptation timescales in the encoder, which we imagine to be a biological network. For example, a salamander may, on one hand, need to be able to predict at a timescale relevant for prey catching and predict the dynamics of its prey, while on the other, be able to make predictions at different timescales to avoid predators, and

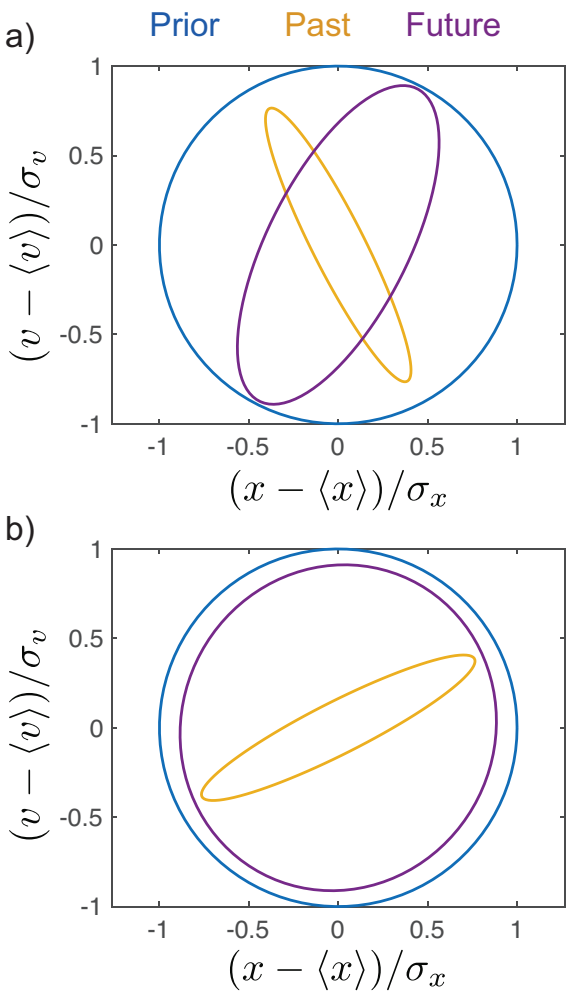

Fig 3.5: Example of a sub-optimal compression. An optimally predictive, compressed representation, in panel (a) compared to a suboptimal representation, in panel (b) for a prediction at $\Delta t = 1$ in the future, within the underdamped regime ($\zeta = 1/2$). We fix the mutual information between the representations and X_t ($I(X_t; \tilde{X}) = 3$ bits), but find that, as expected, the suboptimal representation contains significantly less information about the future.

predators may have a different dynamical regime(143; 139). One possible solution is for the encoder to employ a representation that is useful across a wide range of input statistics. This requires that the predictive power of a given representation is, to some extent, transferrable to other input regimes. To quantify how 'transferrable' different representations are, we take an optimal representation from one $(\zeta, \Delta t)$ and ask how efficiently it captures predictive information for a different parameter regime, $(\zeta', \Delta t')$.

We identify these global strategies by finding the optimal encoder for a stimulus with parameters $(\zeta, \Delta t)$ that generates a representation, $\mathcal{P}(\tilde{X}|X_t)$, at some given compression level, I_{past}. We will label the predictive information captured by this representation $I^{\text{future}}_{\text{optimal}}((\zeta, \Delta t), I_{\text{past}})$. We hold the representation fixed and apply it to a stimulus with different underlying parameters $(\zeta', \Delta t')$ and compute the amount of predictive information the previous representation yields for this stimulus. We call this the transferred predictive information $I^{\text{future}}_{\text{transfer}}((\zeta, \Delta t), I_{\text{past}} \rightarrow (\zeta', \Delta t'))$. We note that $I^{\text{future}}_{\text{transfer}}((\zeta, \Delta t), I_{\text{past}} \rightarrow (\zeta', \Delta t'))$ may sometimes be larger than $I^{\text{future}}_{\text{optimal}}((\zeta, \Delta t), I_{\text{past}})$, because changing $(\zeta, \Delta t)$ may increase both I_{past} and I_{future} (see e.g. Fig 3.6A).

For every fixed $(\zeta, \Delta t)$ and I_{past}, we can take the optimal $\tilde{X}$ and transfer it to a wide range of new ζ''s and timescales, $\Delta t'$. For a particular example $(\zeta, \Delta t)$, this is shown in Fig 3.6B. The representation optimized for critical damping is finer-grained than what's required in the overdamped regime. We can sweep over all combinations of the new ζ''s and $\Delta t'$s. What we get, then, is a mapping of $I^{\text{future}}_{\text{transfer}}$ for this representation that was optimized for one particular $(\zeta, \Delta t)$ pair across all new $(\zeta', \Delta t')$'s. This is shown in Fig 3.6C, (Fig 3.6B are just two slices through this surface). This surface gives a qualitative picture the transferability of this particular representation.

To get a quantitative summary of this behavior that we can then compare across different starting points $(\zeta, \Delta t)$, we integrate this surface over $1/3 < \zeta' < 3$, $0.1 < \Delta t' < 10$, and then normalize by the integral of $I^{\text{future}}_{\text{optimal}}((\zeta', \Delta t'), I_{\text{past}})$ over the same surface. This

yields an overall transferability measure, $Q^{\text{transfer}}(\zeta, \Delta t)$. We report these results in Fig 3.6D. Representations that are optimal for underdamped systems at late times are the most transferable. This is because generating a predictive mapping for underdamped motion requires some measurement of velocity, which is generally useful for many late-time predictions. Additionally, prediction of underdamped motion requires high precision measurement of position, and that information is broadly useful across all parameters.

3.4.2 History-dependent Gaussian Stimuli

In the above analysis, we considered stimuli with temporal correlations that fall off exponentially. However, natural scenes, such as leaves blowing in the wind or bees moving in their hives, are shown to have heavy-tailed statistics (22; 143; 140). To extend our results to such stimuli, we consider prediction where the statistics of the motion model may feature long-ranged temporal correlations and by increasing the dimensionality of the input and output to the information bottleneck, we demonstrate that the information bottleneck continues to provide useful predictive encoding schemes for such stimuli. We show this through the use of the Generalized Langevin equation (144; 88; 66):

$$\frac{dv}{dt} = -\int_0^t \frac{\gamma v}{|t-t'|^\alpha} dt - \omega_0^2 x + \xi(t) \tag{3.15}$$

$$\frac{dx}{dt} = v \tag{3.16}$$

Here, we have returned to unscaled definitions of v, and t. The damping force has a power-law kernel. In order for the system to obey the fluctuation-dissipation theorem, we note that $\langle \xi(t) \rangle = 0$, and $\langle \xi(t')\xi(t) \rangle \propto \frac{1}{|t-t'|^\alpha}$. In this dynamical system, position autocorrelation $\langle x(t)x(t') \rangle \sim t^{-\alpha}$ and velocity autocorrelation $\langle v(t)v(t') \rangle \sim t^{-\alpha-1}$ for large t.

The prediction problem is similar to the prediction problem for the memoryless SD-

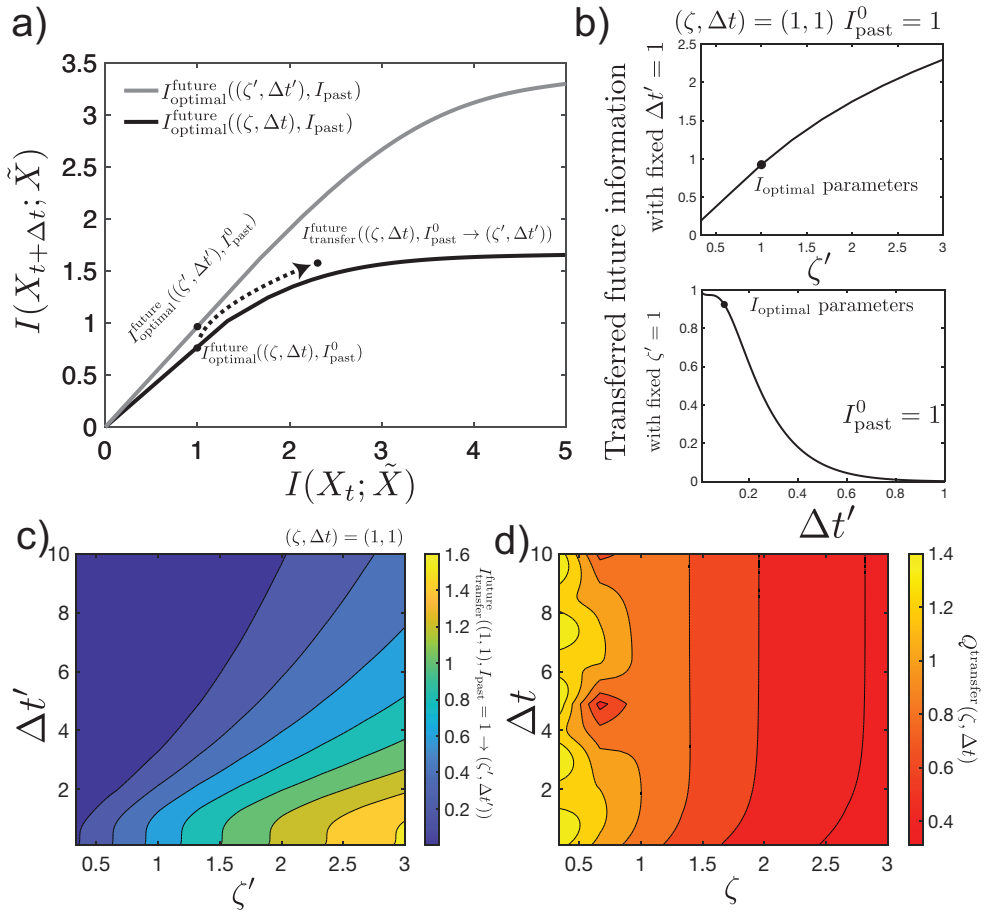

Fig 3.6: Representations learned on underdamped systems can be transferred to other types of motion, while representations learned on overdamped systems cannot be easily transferred. (a) Here, we consider the information bottleneck bound curve (black) for a stimulus with underlying parameters, $(\zeta, \Delta t)$. For some particular level of $I_{\text{past}} = I^0_{\text{past}}$, we obtain a mapping, $\mathcal{P}(\tilde{X}|X_t)$ that extracts some predictive information, denoted $I^{\text{future}}_{\text{optimal}}((\zeta, \Delta t), I^0_{\text{past}})$, about a stimulus with parameters $(\zeta, \Delta t)$. Keeping that mapping fixed, we determine the amount of predictive information for dynamics with new parameters $(\zeta', \Delta t')$, denoted by $I^{\text{future}}_{\text{transfer}}((\zeta, \Delta t), I^0_{\text{past}} \to (\zeta', \Delta t'))$. (b) One-dimensional slices of $I^{\text{future}}_{\text{transfer}}$ in the $(\zeta', \Delta t')$ plane: $I^{\text{future}}_{\text{transfer}}$ versus ζ' for $\Delta t' = 1$. $I^0_{\text{past}} = 1$ (top), and versus $\Delta t'$ for $\zeta' = 1$. Parameters are set to $(\zeta = 1, \Delta t = 1)$, $I^0_{\text{past}} = 1$. (c) Two-dimensional map of $I^{\text{future}}_{\text{transfer}}$ versus $(\zeta', \Delta t')$ (same parameters as b). (d) Overall transferability of the mapping. The heatmap of (c) is integrated over ζ' and $\Delta t'$ and normalized by the integral of $I^{\text{future}}_{\text{optimal}}((\zeta', \Delta t'), I_{\text{past}})$. We see that mappings learned from underdamped systems at late times yield high levels of predictive information for a wide range of parameters, while mappings learned from overdamped systems are not generally useful.

DHO, but we now take an extended past, $X_{t-t_0:t}$, for prediction of an extended future, $X_{t+\Delta t:t+\Delta t+t_0}$, where t_0 sets the size of the window into the past we consider and the future we predict (Fig 3.7A). Using the approach described in S1 Text, we compute the optimal representation and determine how informative the past is about the future. The objective function for this extended information bottleneck problem is,

$$\mathcal{L} = \min_{\mathcal{P}(\tilde{X}|X_{t-t_0:t})} I(X_{t-t_0:t};\tilde{X}) - \beta I(X_{t+\Delta t:t+\Delta t+t_0};\tilde{X}). \tag{3.17}$$

We demonstrate the impacts of the discretization of time in S2. The information bottleneck curves show more predictive information as the prediction process uses more past information (larger t_0 in Fig 3.7B). Not including any history results in an inability to extract the predictive information. However, for low compression, large β, we find that the amount of predictive information that can be extracted saturates quickly as we increase the amount of history, t_0. This implies diminishing returns in prediction for encoding history. Despite the diverging autocorrelation timescale, prediction only functions on a limited timescale and the maximum available prediction information always saturates as a function of t_0 (Fig 3.7C). These results indicate that efficient coding strategies can enable prediction even in complex temporally correlated environments.

3.4.3 Evolutionary dynamics

Exploiting temporal correlations to make predictions is not limited to vision. Another aspect of the prediction problem appears in the adaptive immune system, where temporal correlations in pathogen evolution may be exploited to help an organism build and maintain immunity in a changing environment. Exploiting these correlations can be done at a population level, in terms of vaccine design (86; 47; 173; 141), and has been postulated as a means for the immune system to adapt to future threats (98; 120). Here, we present efficient predictive coding strategies for the Wright-Fisher model, which is commonly used to describe viral

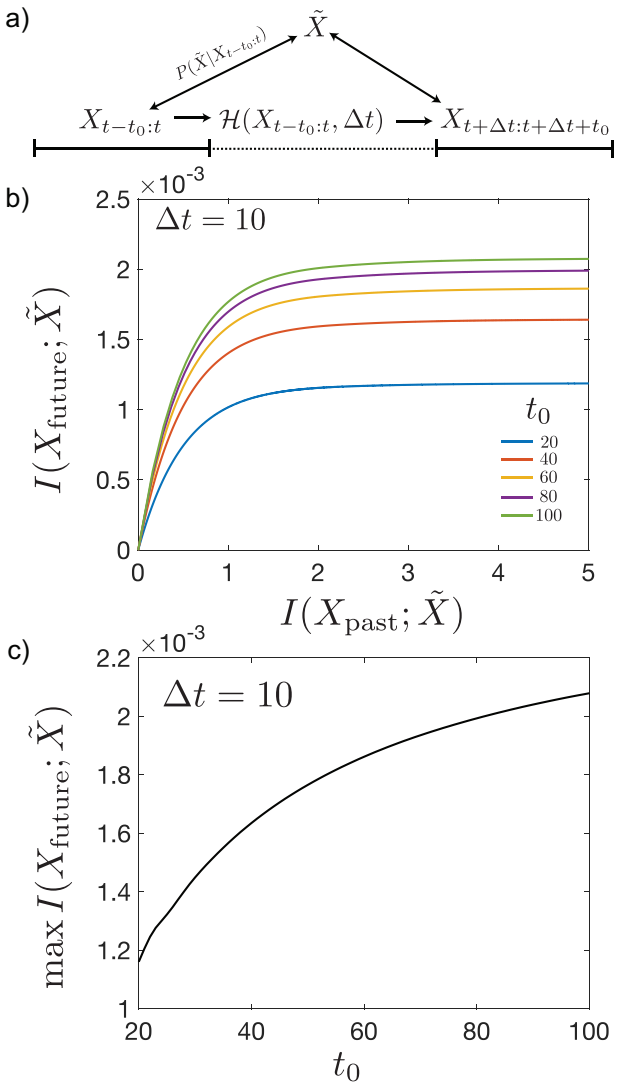

Fig 3.7: The ability of the information bottleneck Method to predict history-dependent stimuli. (a) The prediction problem, using an extended history and a future. This problem is largely similar to the one set up for the SDDHO but the past and the future are larger composites of observations within a window of time $t - t_0 : t$, expressed as X_{past} for the past and $t + \Delta t : t+\Delta t+t_0$, expressed as X_{future} for the future. (b) Predictive information $I(X_{t+\Delta t:t+\Delta t+t_0}, \tilde{X})$ with lag Δt. (c) The maximum available predictive information saturates as a function of the historical information used t_0.

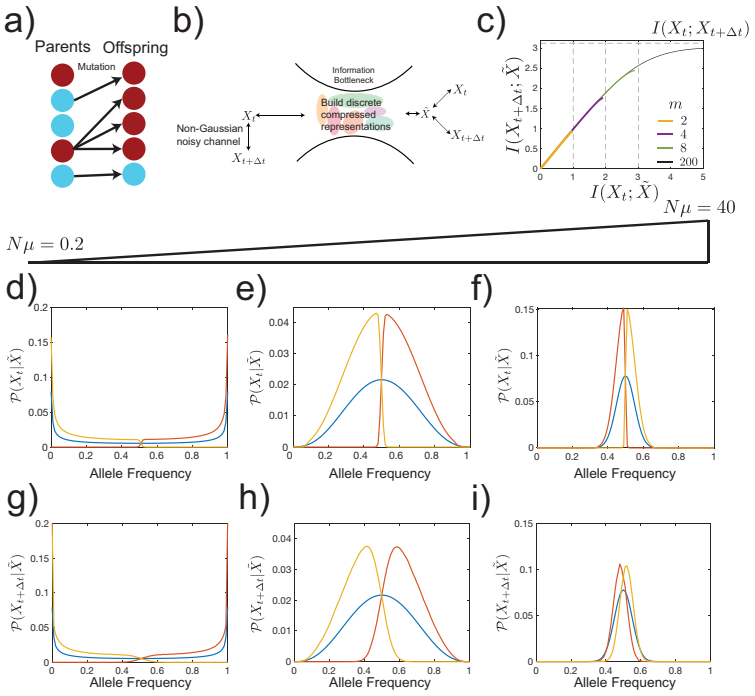

Fig 3.8: The information bottleneck solution for a Wright Fisher process. (a) The Wright-Fisher model of evolution can be visualized as a population of N parents giving rise to a population of N offspring. Genotypes of the offspring are selected as a function of the parents' generation genotypes subject to mutation rates, μ, and selective pressures s. (b) Information bottleneck schematic with a discrete (rather than continuous) representation variable, $\tilde{X}$. (c) Predictive information as a function of compression level. Predictive information increases with the cardinality, m, of the representation variable. The amount of predictive information is limited by $\log(m)$ (vertical dashed lines) for small m, and the mutual information between allele frequencies at time $t + \Delta t$ and time t, $I(X_{t+\Delta t}; X_t)$ (horizontal dashed line), for large m. Bifurcations occur in the amount of predictive information. For small $I(X_t; \tilde{X})$, the encoding strategies for different m are degenerate and the degeneracy is lifted as $I(X_t; \tilde{X})$ increases, with large m schemes accessing higher $I(X_t; \tilde{X})$ ranges. Parameters: $N = 100$, $N\mu = 0.2$, $N\mu = 0.2$, $Ns = 0.001$, $\Delta t = 1$. (d-i) We explore information bottleneck solutions to Wright-Fisher dynamics under the condition that the cardinality of $\tilde{X}$, m, is 2 and take β to be large enough that $I(X_t; \tilde{X}) \approx 1$, $\beta \approx 4$. Parameters: $N = 100$, $Ns = 0.001$, $\Delta t = 1$, and $N\mu = 0.2$, $N\mu = 2$, and $N\mu = 40$ (from left to right). (d-f) In blue, we plot the steady state distribution. In yellow and red, we show the inferred historical distribution of alleles based on the observed value of $\tilde{X}$. Note that each distribution is corresponds to roughly non-overlapping portions of allele frequency space. (g-i) Predicted distribution of alleles based on the value of $\tilde{X}$. We observe that as mutation rate increases, the timescale of relaxation to steady state decreases, so historical information is less useful and the predictions becomes more degenerate with the steady state distribution.

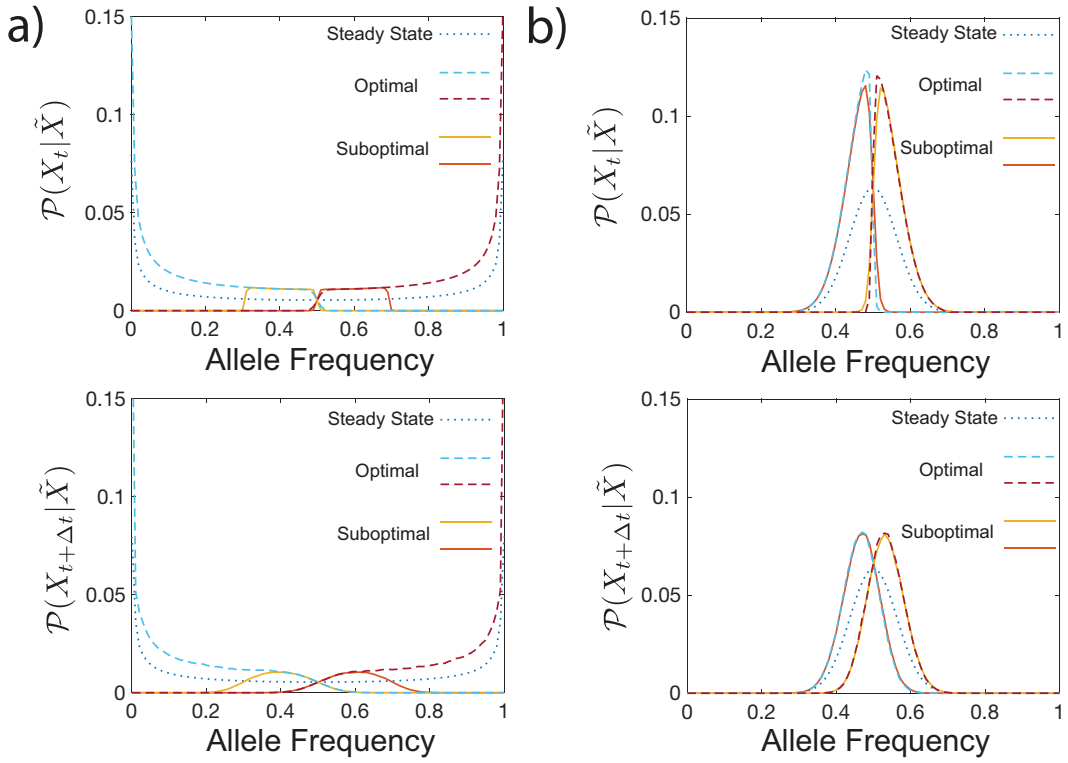

Fig 3.9: Transferability of prediction schemes in Wright-Fisher dynamics. We transfer a mapping, $\mathcal{P}(\tilde{X}|X_t)$, trained on one set of parameters and apply it to another. We consider transfers between two choices of mutability, $N\mu_1 = 0.2$ (low) and $N\mu_2 = 20$ (high), with $N = 100$, $Ns = 0.001$, $\Delta t = 1$. The dotted line is the steady state allele frequency distribution, the solid lines are the transferred representations, and the dashed lines are the optimal solutions. The top panels correspond to the distributions of X_t and the bottom panels correspond to distributions of $X_{t+\Delta t}$. (a) Transfer from high to low mutability. Optimal information values: $I^{\text{past}}_{\text{optimal}} = 0.98$ and $I^{\text{future}}_{\text{optimal}} = 0.93$; transferred information values: $I^{\text{past}}_{\text{transfer}}((N\mu_2), I_{\text{past}} = 0.92 \to (N\mu_1)) = 0.14$ and $I^{\text{future}}_{\text{transfer}}((N\mu_2), I_{\text{past}} = 0.92 \to (N\mu_1)) = 0.05$. Representations learned on high mutation rates are not predictive in the low mutation regime. (b) Transfer from low to high mutability. Optimal information values: $I^{\text{past}}_{\text{optimal}} = 0.92$ and $I^{\text{future}}_{\text{optimal}} = 0.92$ and $I^{\text{future}}_{\text{optimal}} = 0.28$. Transferred information values: $I^{\text{past}}_{\text{transfer}}((N\mu_1), I_{\text{past}} = 0.98 \to (N\mu_2)) = 0.79$ and $I^{\text{future}}_{\text{transfer}}((N\mu_1), I_{\text{past}} = 0.98 \to (N\mu_2)) = 0.27$. Transfer in this direction yields good predictive informations.

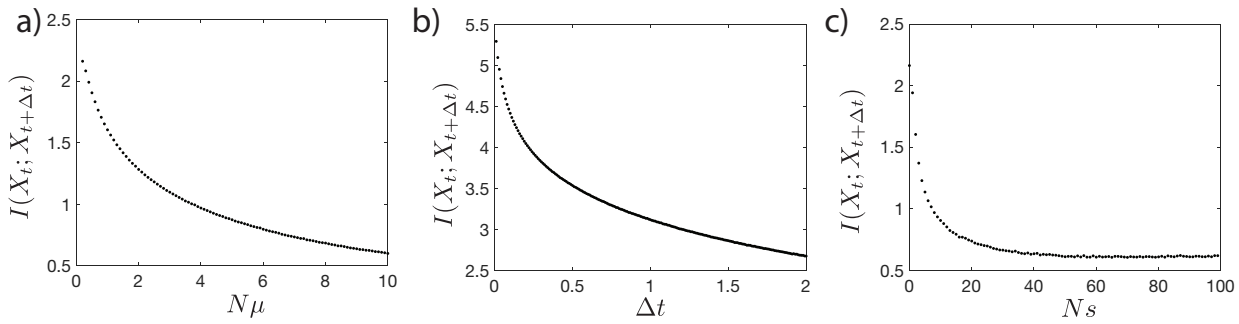

Fig 3.10: Amount of predictive information in the Wright Fisher dynamics as a function of model parameters. (a-c), Value of the asymptote of the information bottleneck curve, $I(X_t; X_{t+\Delta t})$ with: (a) $N = 100$, $Ns = 0.001$, $\Delta t = 1$ as a function of μ; (b) $N = 100$, $N\mu = 0.2$, $Ns = 0.001$ as a function of Δt; and (c) $N = 100$, $N\mu = 0.2$, and $\Delta t = 1$ as a function of s.

evolution (138). In contrast to the two models studied so far, Wright-Fisher dynamics are not Gaussian, though they are still Markovian. This implies that predictive information can reside in higher-order moments of the joint distribution, thus the optimal compressed representation variable can no longer be Gaussian. The Wright-Fisher model allows us to explore how the results obtained in the previous sections generalize to non-Gaussian statistics of the past and future distributions. To make this computationally tractable, we will take the representation variable to be discrete, though later allow its cardinality to be large to approximate the continuous solution. There exist methods to approximate continuous compressed representations directly(29; 124; 76), though we do not use those here.

Wright-Fisher models of evolution assume a constant population size of N. We consider a single mutating site with each individual in the population having either a wild-type or a mutant allele at this site. The allele choice of subsequent generations depends on the frequency of the mutant allele in the ancestral generation at time t, X_t, the selection pressure on the mutant allele, s, and the mutation rate from the wild-type to the mutant allele and back, μ, as depicted as Fig 3.8A. For large enough N, the update rule of the allele frequencies

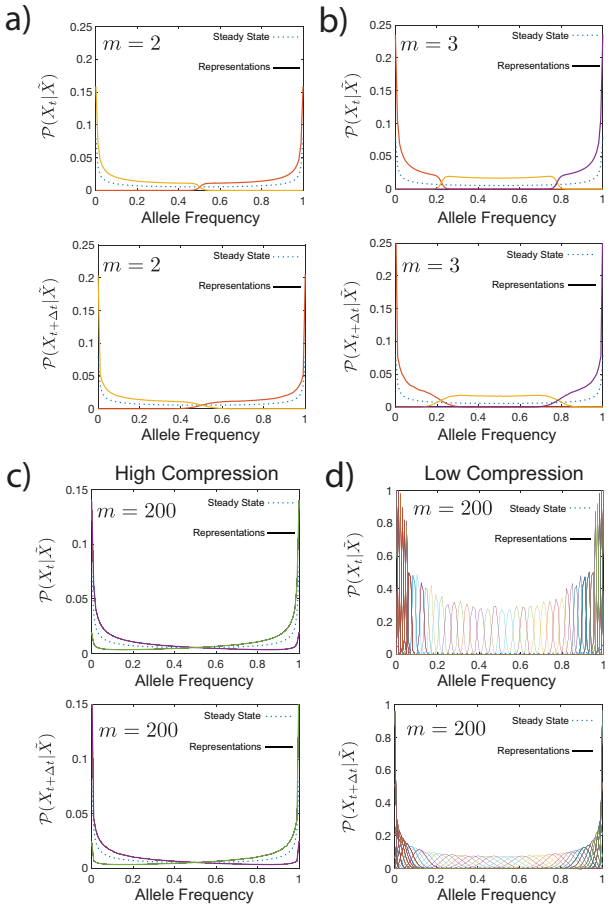

Fig 3.11: Encoding schemes with $m > 2$ representation variables. The steady state is plotted as a dotted line and the representation for each realization of the value of $\tilde{X}$ are plotted as solid lines. The representations which carry maximum predictive information for (a) $m = 2$ at $I(X_t; \tilde{X}) \approx \log(m) = 1$ bit, and (b) $m = 3$ at $I(X_t; \tilde{X}) \approx \log(m) \approx 1.5$ bits. The optimal representations at large m tile space more finely and have higher predictive information. The optimal representations for $m = 200$ at fixed $\beta = 1.01$ ($I(X_t; \tilde{X}) = 0.28$, $I(X_{t+\Delta t}; \tilde{X}) = 0.27$) (c) and $\beta = 20$ ($I(X_t; \tilde{X}) = 2.77$, $I(X_{t+\Delta t}; \tilde{X}) = 2.34$). (d) At low $I(X_t; \tilde{X})$, many of the representations are redundant and do not confer more predictive information than the $m = 2$ scheme. A more explicit comparison is given in S3 Fig. At high $I(X_t; \tilde{X})$, the degeneracy is lifted. All computations done at $N = 100$, $N\mu = 0.2$, $Ns = 0.001$, $\Delta t = 1$.

is given through the diffusion approximation interpreted with the Ito convention (72):

$$\frac{dX_t}{dt} = sX_t(1 - X_t) + \mu(1 - 2X_t) + \sqrt{X_t(1 - X_t)/N}\eta(t), \tag{3.18}$$

where $\langle \eta(t) \rangle = 0$, $\langle \eta(t)\eta(t') \rangle = \delta(t - t')$. We note that this model is Markovian, so as we did with the SDDHO, we will take the historical variable to be X_t and the future variable to be $X_{t+\Delta t}$. Details are given in S3 Text. Extending the timescale of the representation of the past will not confer additional predictive information.

For this model, defining the representation $\tilde{X}$ as a noisy linear transformation of X_t, the allele frequency at time t, as we did for the Gaussian case in S1 Text Eq.1 does not capture all of the dependences between the past and future allele frequencies, because correlations exist beyond second order. This arises because of the non-linear form of Eq. 3.18. Instead, we determine the mapping of X_t to $\tilde{X}$ numerically using the Blahut-Arimoto algorithm (7; 23). In general, for a discrete representation variable $\tilde{X}$, the true cardinality of $\tilde{X}$ is unknown for a given β. Our approach is to first fix the cardinality of $\tilde{X}$ to a given value m (Fig 3.8C) and compute the information curve for the given m by sweeping over β. We then repeat this for larger values of m. We note that for small β, the solutions for different values of m are degenerate, while at higher values of β, bifurcations emerge between encoding schemes for the solutions with cardinality m and $m - 1$. This is because the true cardinality of the optimal solution undergoes transitions to higher and higher values as β increases (158). The discreteness of $\tilde{X}$ results in each realization of the representation tiling a distinct part of frequency space. This encoding scheme can be thought of a different types of immune defenses: innate, adaptive, and different lymphocyte phenotypes acting at different stages or for different types of immune responses (112). Accordingly, m would correspond to the number of distinct cell types mobilized against pathogens of various frequencies. The concept of discrete tiling of space is also analogous to ideas of immune coverage, whereby a finite number of distinct antigen receptors cover the entire "shape space" of possible antigens(126).

However, to make this analogy more precise would require to study an effective theory of phenotypic evolution(121).

We first consider the example with $m = 2$ representations. In the weak-mutation, weak-selection limit ($N\mu, Ns \ll 1$), the steady state probability distribution of allele frequencies,

$$P_s(X) \propto [X(1-X)]^{N\mu-1} e^{NsX} \tag{3.19}$$

(blue line in Fig 3.8D) is peaked around the frequency boundaries, indicating that at long times, an allele either fixes or goes extinct. In this case, one value of the representation variable corresponds to the range of high allele frequencies and the other corresponds to low allele frequencies (Fig 3.8D, yellow and red lines). These encoding schemes can be used to make predictions, whether it be by an observer or the immune system, via determining the future probability distribution of the alleles conditioned on the value of the representation variables, $\mathcal{P}(X_{t+\Delta t}|\tilde{X})$. We present these predictions in Fig 3.8G. The predictive information conferred by the representation variable is limited by the information it has about X_t as in the Gaussian case (Fig 3.8C.)

For larger mutation rates, the steady state distribution becomes centered around the equal probability of observing either one of the two alleles, but the two representation variables still cover the frequency domain in way that minimizes overlap (Fig 3.8E and 3.8F). We observe a sharp drop in $P(X_t|\tilde{X})$ at the boundary between the two representations. The future distribution of allele frequencies in this region (Fig 3.8H and 3.8I), however, displays large overlap. The degree of this overlap increases as the mutation rate gets larger, suggesting prediction is harder in the strong mutation limit. The optimal encoding of the distribution of X_t biases the representation variable towards frequency space regions with larger steady state probability mass.

In Fig 3.9, we explore the consequence of transferring a mapping, $\mathcal{P}(\tilde{X}|X_t)$, from a high mutation model to a low mutation model and vice versa. We observe that the weak mutation

representation is more transferrable than the strong mutation representation. One reason for this is that the strong mutation limit provides little predictive information, as seen in Fig 3.10A. In addition, high mutation representations focus on $X = 1/2$, while the population more frequently occupies allele frequencies near 0 and 1 in other regimes. Comparatively, representations learned on weak mutation models can provide predictive information, because they cover more evenly the spectrum of allele frequencies.

We can extend the observations in Fig 3.8 to see how the predictive information depends on the strength of the selection and mutation rates (Fig 3.10A and 3.10C). Prediction is easiest in the weak mutation and selection limit, as population genotype change occur slowly and the steady state distribution is localized in one regime of the frequency domain. For evolutionary forces acting on faster timescales, prediction becomes harder since the relaxation to the steady state is fast. Although the mutation result might be expected, the loss of predictive information in the high selection regime seems counterintuitive: due to a large bias between one of the two alleles evolution appears reproducible and "predictable" in the high selection limit. This bias renders the allele state easier to guess but this is not due to information about the initial state. The mutual information-based measure of predictive information used here captures a reduction of entropy in the estimation of the future distribution of allele frequencies due to conditioning on the representation variable. When the entropy of the future distribution of alleles $H(X_{t+\Delta t})$ is small, the reduction is small and predictive information is also small. As expected, predictive information decreases with time Δt, since the state X_t and $X_{t+\Delta t}$ decorrelate due to noise (Fig 3.10B).

So far we have discussed the results for $m = 2$ representations. As we increase the tradeoff parameter, β in Eq. 3.1, the amount of predictive information increases, since we retain more information about the the allele frequency at time t. However, at high β values the amount of information the representation variable can hold saturates, and the predictive information reaches a maximum value (1 bit for the $m = 2$ yellow line in Fig 3.10A). Increasing the number

of representations m to 3 increases the range of accessible information the representation variable has about the past $I(X_t; X)$, increasing the range of predictive information (purple line in Fig 3.8C)). Comparing the $m = 2$ and $m = 3$ representations for maximum values of β for each of them (Fig 3.11A and 3.11B), shows that larger numbers of representations tile allele frequency space more finely, allowing for more precise encodings of the past and future distributions. The maximum amount of information about the past goes as $\log(m)$ (Fig 3.8C). The predictive information curves for different m values are the same, until the branching point $\lesssim \log(m)$ for each m (Fig 3.8C).

We analyze the nature of this branching by taking $m \gg 1$, $m = 200$ (Fig 3.11C and 3.11D). At small β (and corresponding small $I(X_t; \tilde{X})$) the optimal encoding scheme is the same if we had imposed a small m (Fig 3.11C), with additional degenerate representations (S3 Fig). By increasing β (and $I(X_t; \tilde{X})$), the degeneracy is lifted and additional representation cover non-overlapping regimes of allele frequency space. This demonstrates the existence of a critical β for each predictive coding scheme, above which m needs to be increased to extract more predictive information and below which additional values of the representation variable encode redundant portions of allele frequency space. While we do not estimate the critical β, approaches to estimating them are presented in (183; 182).

The $m = 200$ encoding approximates the continuous $\tilde{X}$ representation. In the high $I(X_t; \tilde{X})$ limit, the $m = 200$ encoding gives precise representations (i.e. with low variability in $\mathcal{P}(X_t|\tilde{X})$) in regions of allele frequency space with high steady state distribution values, and less precise representations elsewhere (Fig 3.11D top panel and S4). This dependence differs from the Gaussian case, where the uncertainty of the representation is independent of the encoded value. The decoding distributions $\mathcal{P}(X_t|\tilde{X})$ are also not Gaussian. This encoding builds a mapping of internal response to external stimuli, by tiling the internal representation space of external stimuli in a non-uniform manner. These non-uniform frequency tilings are similar to Laughlin's predictions for maximally informative coding in vision (80), but with

the added constraint of choosing the tiling to enable the most informative predictions.

3.5 Discussion

We have demonstrated that the information bottleneck method can be used to construct predictive encoding schemes for a variety of biologically-relevant dynamic stimuli. The approach described in this paper can be used to make predictions about the underlying encoding schemes used by biological systems that are compelled by their behavioral and fitness constraints to make predictions. These results thus provide experimentally testable hypotheses. The key principle is that not all input dimensions are equally relevant for prediction; information encoding systems must be able to parse which dimensions are relevant when coding capacity is small relative to the available predictive information. Hence, the biological (or engineered) system must navigate a tradeoff between reducing the overall uncertainty in its prediction while only being able to make measurements with some fixed uncertainty.

It may not always be the case, experimentally, that a system uses an optimal encoding for prediction of a particular motion stimulus. When the stimulus nonetheless falls within the natural scene input repertoire for the organism, we hypothesize that biological systems may use a best-compromise predictive encoding of their inputs because that need to operate flexibly across a wide range of different input statistics. We provide a transferability metric, Q, which quantifies how useful a particular scheme is across other dynamic regimes and prediction timescales, that can be used to experimentally predict what the best-compromise predictive encoding scheme is in cases where a biological system needs to be flexible. We observe that a compromise between representing position and velocity of a single object provides a good, general, predictor for a large set of input behaviors. When adaptation is slower than the timescale over which the environment changes, such a compromise might be beneficial to the biological system. On the other hand, if the biological encoder can adapt,

the optimal predictive encoder for those particular dynamics is the best encoder. We have provided a fully-worked set of examples of what those optimal encoders look like for a variety of parameter choices. The dynamics of natural inputs to biological systems could be mapped onto particular points in these dynamics, providing a hypothesis for what optimal prediction would look like in that system.

We also explored the ability to predict more complex, non-Markovian dynamics. We asked about the usefulness of storing information about the past in the presence of power-law temporal correlations. The optimal information bottleneck solution showed fast diminishing returns as it was allowed to dig deeper and deeper into the past, suggesting that simple encoding schemes with limited temporal span have good predictive power even in complex correlated environments.

Superficially, our framework may seem similar to a Kalman filter (67). There are few major differences in this approach. Kalman filtering algorithms have been used to explain responses to changes in external stimuli in biological system (63). In this framework, the Kalman filters seek to maximize information by minimizing the variance in estimating the true coordinates of an external input. The estimate is, then, a prediction of the next time step, and is iteratively updated. Our information bottleneck approach extracts past information, but explicitly includes another constraint: resource limitations. The tuning of I_{past} is the main difference between our approach and a Kalman filter. Another major difference is that we do not assume the underlying encoder has any explicit representation of the 'physics' of the input. There is no internal model of the input stimulus, apart from our probabilistic mapping from the input to our compressed representation of that input. A biological system could have such an internal model, but that would add significant coding costs that would have to be treated by another term in our framework to draw a precise equivalence between the approaches. We show in the S1 Fig that the Kalman filter approach is not as efficient, in general, as the predictive information bottleneck approach that we present here.

Our results on systems with Wright-Fisher input dynamics reveal that discrete representations that tile input space are optimally predictive encoders. Although we impose discrete internal representations, their non-overlapping character remains even it the limit of a large number of representations. These kinds of solutions are reminiscent of the Laughlin solution for information maximization of input and output in the visual system given a nonlinear noisy channel (80), in which the input space is covered proportionally to the steady state distribution at a given frequency, in chunks given by the size of the noise in the system. Tiling solutions have also been described when optimizing information in gene regulatory networks with nonlinear input-output relations, when one input regulates many gene outputs (168). In this case each gene was expressed in a different region of the input concentration domain. Similarly to our example, where the lifting the degeneracy between multiple representations covering the same frequency range allows for the prediction of more information about the future, lifting the degeneracy between different genes making the same readout, increases the transmitted information between the input concentration and the outputs. More generally, discrete tiling solutions are omnipresent in information optimization problems with boundaries (118; 149).

Biologically, predicting evolutionary dynamics is a different problem than predicting motion. Maybe the accuracy of prediction matters less, while covering the space of potentially very different inputs is important. In our simple example, this is best seen in the strong mutation limit where the mutant allele either fixes or goes extinct with equal probability. In this case, a single Gaussian representation cannot give a large values of predictive information. A discrete representation, which specializes to different regions of input space, is a way to maximize predictive power for very different inputs. It is likely that these kinds of solutions generalize to the case of continuous, multi-dimensional phenotypic spaces, where discrete representations provides a way for the immune system to hedge its bets against pathogens by covering the space of antigen recognition(96). The tiling solution that appears in the

non-Gaussian solution of the problem is also potentially interesting for olfactory systems. The number of odorant molecules is much larger than odor receptors (167; 49), which can be thought of as representation variables that cover the phenotypic input space of odorants. The predictive information bottleneck solution gives us a recipe for covering space, given a dynamical model of evolution of the inputs.

The results in the non-Gaussian problem are different than the Gaussian problem in two important ways: the encoding distributions are not Gaussian (e.g. Fig 3.8D and 3.8E), and the variance of the encoding distributions depends on the the value of $\mathcal{P}(X_t|\tilde{X})$ (Fig 3.11D). These solutions offer more flexibility for internal encoding of external signals.

The information bottleneck approach has received a lot of attention in the machine learning community lately, because it provides a useful framework for creating well-calibrated networks that solve classification problems at human-level performance(3; 29; 4). In these deep networks, variational methods approximate the information quantities in the bottleneck, and have proven their practical utility in many machine learning contexts. These approaches do not always provide intuition about how the networks achieve this performance and what the information bottleneck approach creates in the hidden encoding layers. Here, we have worked through a set of analytically tractable examples, laying the groundwork for building intuition about the structure of information bottleneck solutions and their generalizations in more complex problems.

In summary, the problem of prediction, defined as exploiting correlations about the past dynamics to anticipate the future state comes up in many biological systems from motion prediction to evolution. This problem can be formulated in the same way, although as we have shown, the details of the dynamics matter for how best to encode a predictive representation and maximize the information the system can retain about the future state. Dynamics that results in Gaussian propagators is most informatively predicted using Gaussian representations. However non-Gaussian propagators introduce disjoint non-Gaussian representations that are

nevertheless predictive.

By providing a set of dissected solutions to the predictive information bottleneck problem, we hope to show that not only is the approach feasible for biological encoding questions, it also illuminates connections between seemingly disparate systems (such as visual processing and the immune system). In these systems the overarching goal is the same, but the microscopic implementation might be very different. Commonalities in the optimally predictive solutions as well as the most generalizable ones can provide clues about how to best design experimental probes of this behavior, at both the molecular and cellular level or in networks.

3.6 Computing the optimal representation for jointly Gaussian past-future distributions

We reproduce Chechik et al.(31) to show the analytic construction of the optimally predictive representation variable, $\tilde{X}$, when the input and output variables are jointly Gaussian. The input is $X_t \sim \mathcal{N}(0, \Sigma_{X_t})$ and the output is $X_{t+\Delta t} \sim \mathcal{N}(0, \Sigma_{X_{t+\Delta t}})$. The joint distribution of X_t and $X_{t+\Delta t}$ is Gaussian. To construct the representation, we take a noisy linear transformation of X_t to define $\tilde{X}$

$$\tilde{X} = A_\beta X_t + \xi. \tag{3.20}$$

Here, A_β is a matrix whose elements are a function of β, the tradeoff parameter in the information bottleneck objective function between compressing, in our case, the past while retaining information about the future. ξ is a vector of dimension $\dim(X_t)$. The entries of ξ are Gaussian-distributed random numbers with 0 mean and unit variance. Because the joint distribution of the past and the future is Gaussian, to capture the dependencies of $X_{t+\Delta t}$ on X_t we can use a noisy linear transform of X_t to construct a representation variable that satisfies the information bottleneck objective function(31).

We compute A_β by first computing the left eigenvectors and the eigenvalues of the

regression matrix, $\Sigma_{X_t|X_{t+\Delta t}}\Sigma_{X_t}^{-1}$. Here, $\Sigma_{X_t|X_{t+\Delta t}}$ is the covariance matrix of the probability distribution of $\mathcal{P}(X_t|X_{t+\Delta t})$. These eigenvector–eigenvalue pairs satisfy the following relation

$$v_i^T \Sigma_{X_t|X_{t+\Delta t}} \Sigma_{X_t}^{-1} = \lambda_i v_i^T. \tag{3.21}$$

(We are taking v_i^T to be a row vector, rather than a column vector.)

The matrix, A_β, is then given by

$$A_\beta = \begin{bmatrix} \alpha_1 v_1^T \\ \alpha_2 v_2^T \\ \vdots \end{bmatrix}. \tag{3.22}$$

α_i are scalar values given by

$$\begin{aligned} \alpha_i =& \sqrt{\frac{\beta(1-\lambda_i)-1}{\lambda_i v_i^T \Sigma_{X_t} v_i}} \quad \text{if } \beta > \frac{1}{1-\lambda_i} \\ \alpha_i =& 0 \qquad\qquad\qquad\qquad\quad \text{otherwise.} \end{aligned} \tag{3.23}$$

The α_i define the dimensionality of the most informative representation variable, $\tilde{X}$. The dimension of $\tilde{X}$ is the number of non-zero α_i. The optimal dimension for a given β is, at most, equal to the dimension of $X_{t+\Delta t}$. The set of values, $\{\beta_{c_i}|\beta = 1/(1-\lambda_i)\}$, can be thought of as critical values, as each β_{c_i} triggers the inclusion of the ith left eigenvector into the optimal $\tilde{X}$. The critical values depend strongly on the particular statistics of the input and output variable, so they may be different as the parameters that generate X change.

To compute the information about the past and future contained in $\tilde{X}$, we compute $\mathcal{P}(X_t|\tilde{X})$ and $\mathcal{P}(X_{t+\Delta t}|\tilde{X})$. These distributions are Gaussian. The mean of each distribution corresponds to the encoded value of X_t and $X_{t+\Delta t}$. The variance corresponds to the

uncertainty, or entropy, in this estimate. To compute the variance, we need the variance of $\tilde{X}$

$$\Sigma_{\tilde{X}} = \langle \tilde{X}^T \tilde{X} \rangle = \langle \tilde{X}^T A_\beta^T A_\beta \tilde{X} \rangle + \langle \xi^T \xi \rangle, \tag{3.24}$$

where the excluded terms are zero. Recalling the definition of ξ, we can simplify this expression to yield

$$\Sigma_{\tilde{X}} = A_\beta \Sigma_{X_t} A_\beta^T + I_2. \tag{3.25}$$

Here, I_2 is the identity matrix. To compute the mutual information quantities, we use the following equations,

$$\begin{aligned} I(X_t; \tilde{X}) &= \frac{1}{2} \log_2(|A_\beta \Sigma_{X_t} A_\beta^T + I_2|), \\ I(X_{t+\Delta t}; \tilde{X}) &= I(X_t; \tilde{X}) - \frac{1}{2} \sum_{i=1}^{n(\beta)} \log_2(\beta(1 - \lambda_i)), \end{aligned} \tag{3.26}$$

where $n(\beta)$ corresponds to the number of dimensions included in A_β. We also need the cross covariances between $\tilde{X}$ and X_t and between $\tilde{X}$ and $X_{t+\Delta t}$, which are particularly useful for visualizing the optimal predictive encoding. To obtain these matrices, we use

$$\begin{aligned} \Sigma_{\tilde{X} X_t} &= A_\beta \Sigma_{X_t} \\ \Sigma_{\tilde{X} X_{t+\Delta t}} &= A_\beta \Sigma_{X_{t+\Delta t} X_t}. \end{aligned} \tag{3.27}$$

We can use these results and the Schur complement formula to obtain

$$\begin{aligned} \Sigma_{X_t|\tilde{X}} &= \Sigma_{X_t} - \Sigma_{X_t \tilde{X}} \Sigma_{\tilde{X}}^{-1} \Sigma_{X_t \tilde{X}}^T \\ \Sigma_{X_{t+\Delta t}|\tilde{X}} &= \Sigma_{X_{t+\Delta t}} - \Sigma_{X_{t+\Delta t} \tilde{X}} \Sigma_{\tilde{X}}^{-1} \Sigma_{X_{t+\Delta t} \tilde{X}}^T. \end{aligned} \tag{3.28}$$

3.7 Harmonic Oscillator Model With No Memory

We begin by considering a mass attached to a spring undergoing viscous damping. The mass is being kicked by thermal noise. This mechanical system is largely called the stochastically driven damped harmonic oscillator (SDDHO). A simple model for its position and velocity evolution is given by

$$\begin{aligned} m\frac{dv}{dt} &= -\Gamma v(t) - kx + (2k_B T\Gamma)^{1/2}\xi(t) \\ \frac{dx}{dt} &= v. \end{aligned} \tag{3.29}$$

We use the redefined variables presented in the main text Equations $2-9$ to rewrite the equations as

$$\begin{aligned} \frac{dv}{dt} &= -\frac{x}{4\zeta^2} - v + \frac{\xi(t)}{\sqrt{2}\zeta} \\ \frac{dx}{dt} &= v. \end{aligned} \tag{3.30}$$

There are now two key parameters to explore: ζ and Δt. There are three regimes of motion described by this model. The overdamped regime occurs when $\zeta > 1$. In this regime of motion, the mass, when perturbed from its equilibrium position, relaxes back to its equilibrium position slowly. The underdamped regime occurs when $\zeta < 1$. In this regime of motion, when the mass is perturbed from its equilibrium position, it oscillates about its equilibrium position with an exponentially decaying amplitude. At $\zeta = 1$, we are in the critically damped regime of motion; in this regime, when the mass is perturbed from equilibrium, it returns to equilibrium position as quickly as possible without any oscillatory behavior.

To apply the information bottleneck method to this system, we need to compute the following covariance and cross covariance matrices: Σ_{X_t}, $\Sigma_{X_{t+\Delta t}}$, and $\Sigma_{X_t Y_{t+\Delta t}}$. We note that because the defined motion model is stationary in time, $\Sigma_{X_t} = \Sigma_{X_{t+\Delta t}}$. Using the

procedure given in Flyvbjerg et. al. (119), we can compute the requisite autocorrelations to describe the cross-covariance matrix, $\Sigma_{X_t X_{t+\Delta t}}$.

We begin by using the equipartition theorem that states that

$$\begin{aligned} \langle x_0^2 \rangle &= 1 \\ \langle x_0 v_0 \rangle &= 0 \\ \langle v_0^2 \rangle &= \frac{1}{4\zeta^2}. \end{aligned} \tag{3.31}$$

The covariance matrices are symmetric, so we can use these values to define the elements of Σ_{X_t}. We then obtain expressions for $\Sigma_{X_t X_{t+\Delta t}}$

$$\Sigma_{X_t X_{t+\Delta t}} = \exp\left(-\frac{\Delta t}{2}\right) \begin{bmatrix} \cos(\omega \Delta t) + \frac{\sin(\omega \Delta t)}{2\omega} & -\frac{\sin(\omega \Delta t)}{4\zeta^2 \omega} \\ \frac{\sin(\omega \Delta t)}{4\zeta^2 \omega} & \frac{\cos(\omega \Delta t)}{4\zeta^2} - \frac{\sin(\omega \Delta t)}{8\omega\zeta^2} \end{bmatrix} \tag{3.32}$$

where we have defined $\omega^2 = \frac{1}{4\zeta^2} - \frac{1}{4}$. An alternative approach for the derivation of the above correlation values by methods of Laplace transforms can be found in Sandev et. al. (144).

To construct the optimal representation for prediction, we need the conditional covariance matrices, $\Sigma_{X_t|X_{t+\Delta t}}$ and $\Sigma_{X_{t+\Delta t}|X_t}$. This can be computed using the Schur complement formula to yield

$$\begin{aligned} \Sigma_{X_t|X_{t+\Delta t}} &= \Sigma_{X_t} - \Sigma_{X_t X_{t+\Delta t}} \Sigma_{X_t}^{-1} \Sigma_{X_t X_{t+\Delta t}}^T \\ \Sigma_{X_{t+\Delta t}|X_t} &= \Sigma_{X_t} - \Sigma_{X_t X_{t+\Delta t}}^T \Sigma_{X_t}^{-1} \Sigma_{X_t X_{t+\Delta t}} \end{aligned} \tag{3.33}$$

We provide a graphical representation of these distributions in Fig 2B (main text). These graphical representations correspond to the contour inside which $\sim 68\%$ of observations are observed (i.e. one standard deviation from the mean).

3.7.1 Applying the information bottleneck Solution

To apply the information bottleneck solution, we construct the matrix, $\Sigma_{X_t|X_{t+\Delta t}}\Sigma_{X_t}^{-1}$, and find its eigenvalues and eigenvectors. The left eigenvectors of the matrix will be denoted by the columns of a new matrix, w, given by

$$w = \begin{bmatrix} a+b & a-b \\ 1 & 1 \end{bmatrix}. \tag{3.34}$$

with $a = \omega\cot(\omega\Delta t)$, and $b = \frac{|\csc(\omega\Delta t)|}{2\sqrt{2}\zeta}\sqrt{2-\zeta^2-\zeta^2\cos(2\omega\Delta t)}$. The eigenvalues are then

$$\begin{aligned} \lambda_1 &= 1-\exp(-\Delta t)\left(\frac{1}{4\omega^2\zeta^2} - \frac{\cos(2\omega\Delta t)}{4\omega^2} + \frac{|\sin(\omega\Delta t)|}{2\sqrt{2}\omega^2\zeta}\sqrt{2-\zeta^2-\zeta^2\cos(2\omega\Delta t)}\right) \\ \lambda_2 &= 1-\exp(-\Delta t)\left(\frac{1}{4\omega^2\zeta^2} - \frac{\cos(2\omega\Delta t)}{4\omega^2} - \frac{|\sin(\omega\Delta t)|}{2\sqrt{2}\omega^2\zeta}\sqrt{2-\zeta^2-\zeta^2\cos(2\omega\Delta t)}\right) \end{aligned} \tag{3.35}$$

The transformation matrix, A_β, will now depend on the parameters of the stimulus. Hence, we now refer to this matrix as $A_\beta(\zeta, \Delta t)$, illustrating its functional dependence on those parameters.

Some general intuition can be gained from the form of the above expressions. The eigenvalue gap, $\lambda_1 - \lambda_2$ is proportional to $\frac{\exp(-\Delta t)\|2\sin(\omega\Delta t)\|}{\zeta}$. Intuitively, the eigenvalue gap corresponds the relative importance of the two coding dimensions given by the eigenvectors of the regression matrix. The larger the eigenvalue gap, the more emphasis there is the eigenvector with lower eigenvalue for efficient predictive coding. If the eigenvalue gap is small, there is little benefit for prediction in measuring one dimension over the other. The nature of the dimensions to be measured depends on the direction of the eigenvectors of the regression matrix. This suggests that the eigenvalue gap grows for small Δt, then shrinks for large Δt. Additionally, in the small Δt limit, the eigenvectors align strongly along the position and velocity axes, with the eigenvector corresponding to the smaller eigenvalue being along the

position axis. Hence, for predictions with small Δt, the representation variable must encode a lot of information about the position dimension. For longer timescale predictions, both eigenvectors contribute to large levels of compression, suggesting that the encoding scheme should feature a mix of both position and velocity. This is presented in Fig 4 (main text).

We also compute the total amount of predictive information available in this stimulus. This is given by

$$I(X_t; X_{t+\Delta t}) = \frac{1}{2}\log(|\Sigma_{X_t}|) - \frac{1}{2}\log(|\Sigma_{X_t|X_{t+\Delta t}}|). \tag{3.36}$$

Simplifying this expression yields

$$I(X_t; X_{t+\Delta t}) = \Delta t - \frac{1}{2}\log\Bigg(\exp(2\Delta t) + \cos^4(\omega\Delta t) - \sin^4(\omega\Delta t) \\ -2\exp(\Delta t)\left(\cos^2(\omega\Delta t) + \frac{1+\zeta^2}{1-\zeta^2}\sin^2(\omega\Delta t)\right) + 2\sin^2(\omega\Delta t)\Bigg) \tag{3.37}$$

We can see for very large Δt, this expression becomes

$$I(X_t; X_{t+\Delta t}) \sim \Delta t - \frac{1}{2}\log\left(\exp(2\Delta t) - 2\exp(\Delta t)\right). \tag{3.38}$$

For small Δt, we note there are two conditions: $|\Sigma_{X_t|X_{t+\Delta t}}| < k$ and $|\Sigma_{X_t|X_{t+\Delta t}}| > k$, where k corresponds to width of the distribution. If the width of the Gaussian is below k, we treat this as being effectively deterministic. In this case,

$$I(X_t; X_{t+\Delta t}) \propto \frac{1}{2}\log(|\Sigma_{X_t}|) \tag{3.39}$$

where there are some constants that set the units of the information and the reference point.

For widths larger than k, the expression becomes:

$$I(X_t; X_{t+\Delta t}) \propto \exp(-\Delta t) \tag{3.40}$$

3.7.2 Comparing the information bottleneck Method to Different Encoding Schemes

We compare the encoding scheme discovered by the information bottleneck to alternate encoding schemes. We accomplish this by computing the optimal transformation for a particular parameter set for some value of β, $A_\beta(\zeta, \Delta u)$. We then determine the conditional covariance matrix, $\Sigma_{X_t|\tilde{X}}$. We generate data from this distribution and apply a two-dimensional unitary rotation. We then compute the covariance of the rotated data. This gives us a suboptimal encoding scheme, as represented in Figure 5b in yellow. We note that this representation contains the same amount of mutual information with the past as the optimal representation variable, though the dimensions the suboptimal encoding scheme emphasizes are very different. Evolving the rotated data forward in time and then taking the covariance of the resulting coordinate set gives us $\Sigma_{X_{t+\Delta t}|\tilde{X}}$, as plotted in Fig 5B in purple. We clearly see that encoding the past with the suboptimal representation reduces predictive information, as the predictions of the future are much more uncertain.

3.7.3 Comparing the information bottleneck method to Kalman filters

An alternative approach to predictive coding is Kalman filtering. Kalman filter-based approaches fuses predictions of a system's coordinates at a given time and historical observations of the system's coordinates to achieve increased certainty about the future coordinates of the system(67). However, despite the high-level similarity between Kalman filtering and the information bottleneck method, there are key differences making each technique unique. To show this difference, we present the mathematical structure of a Kalman filter:

$$X^{(\text{naive})}(\Delta t) = \mathcal{H}(\Delta t)X(0) + \xi(\Delta t) \tag{3.41}$$
$$\tilde{X}^{(\text{measured})}(\Delta t) = \mathcal{O}X(\Delta t) + \chi(\Delta t)$$
$$X^{(\text{corrected})} = X^{(\text{naive})}(\Delta t) + K_{\Delta t}(\tilde{X}^{(\text{measured})}(\Delta t) - \mathcal{O}X^{(\text{naive})}(\Delta t)).$$

Here, $\mathcal{O}$ represents a measurement map and $\mathcal{H}(\Delta t)$ represents a dynamical systems model. These features are given to the Kalman filter by the designer. $K_{\Delta t}$ is the filter, and is a function of the measurement map, the dynamical systems model, and the prior uncertainty in the coordinates of the system. The Kalman filter is applied iteratively on each success Δt.

The structure of the Kalman filter reveals two key differences. First, the Kalman filter focuses on the decoding aspect of predictive coding, and is used to improve estimates of a predicted future coordinate via the measurement map and the dynamical systems model. However, the information bottleneck method focuses on the encoding aspect of this problem and generates an optimal encoding scheme for the past. Decoding is not considered explicitly in the information bottleneck. Second, because of the iterative structure of the Kalman Filter, it can use information from an extended time window into the past, while the information bottleneck method can only use one time point of information for predictive coding. This results in Kalman-filtering based approaches using more information about the past than necessary, resulting in inefficient predictive coding. We illustrate this in S1 Fig.

3.7.4 An approach to encoding when the parameters of the stimulus are evolving

We examine prediction in the SDDHO when the underlying parameters governing the trajectory are evolving faster than adaptation timescales. While there are many possible strategies for prediction in this regime, we consider a strategy where the system picks a

representation that provides a maximal amount of information across a large family of stimulus parameters. We chose this strategy because it enables us to analyze the transferability of representations from one parameter set against another. In other words, we can understand how robust representations learned for particular stimulus parameters are.

We first determine the predictive information extracted by an efficient coder for a particular representation level, I_{past} for a particular stimulus with parameters $(\zeta, \Delta t)$, $I^{\text{future}}_{\text{optimal}}((\zeta, \Delta t), I_{\text{past}})$. This predictive mapping is achieved by having a mapping, $\mathcal{P}(\tilde{X}|X_t)$. We apply this mapping to a new stimulus with different parameters $(\zeta, \Delta t)$ to determine the amount of predictive information extracted by this mapping on a different stimulus with parameters $(\zeta', \Delta t')$. We call this predictive information $I^{\text{future}}_{\text{transfer}}((\zeta, \Delta t), I_{\text{past}} \rightarrow (\zeta', \Delta t'))$.

We quantify the quality of these transferred representations in comparison with $I^{\text{future}}_{\text{optimal}}((\zeta', \Delta t'), I_{\text{past}})$ as

$$Q^{\text{transfer}}((\zeta, \Delta t)) = \frac{\int_{\Delta t_{\min}}^{\Delta t_{\max}} \int_{\zeta_{\min}}^{\zeta_{\max}} I^{\text{future}}_{\text{transfer}}((\zeta, \Delta t), I_{\text{past}} \rightarrow (\zeta', \Delta t')) d\zeta' d\Delta t'}{\int_{\Delta t_{\min}}^{\Delta t_{\max}} \int_{\zeta_{\min}}^{\zeta_{\max}} I^{\text{future}}_{\text{optimal}}((\zeta', \Delta t'), I_{\text{past}} d\zeta' d\Delta t'} \tag{3.42}$$

The resulting value is the performance of the mapping against a range of stimuli. In Figure 6, we analyzed the performance of mappings learned on $\frac{1}{3} < \zeta < 3$, $0.1 < t < 10$, on stimuli with parameters $\frac{1}{3} < \zeta' < 3$, $1 < t' < 10$. This choice of range is somewhat arbitrary, but it is large enough to see the asymptotic behavior in Δt, ζ.

3.8 History Dependent Harmonic Oscillators

We extend the results on the Stochastically Driven Damped Harmonic Oscillator to history-dependent stimuli by modifying the original equations of motion to have a history dependent term using the Generalized Langevin Equation

$$\frac{dv}{dt} = -\int_0^t \frac{\gamma v}{|t-t'|^\alpha} dt' - \omega_0^2 x + \xi(t) \tag{3.43}$$
$$\frac{dx}{dt} = v$$

where $-\frac{\gamma}{|t-t'|^\alpha}$ governs how the history impacts the velocity-position evolution. In the main text, we take $\gamma = 1$, $\omega = 1$, and $\alpha = 5/4$. To compute the autocorrelation functions, we compute the Laplace transform of each autocorrelation function and numerically invert the Laplace transform to estimate the value

$$\begin{aligned}
\mathcal{L}[\langle v(t)v(0)\rangle] &= \frac{s}{s^2 + \gamma s^\alpha + \omega^2} \\
\mathcal{L}[\langle v(t)x(0)\rangle] &= -\frac{1}{s^2 + \gamma s^\alpha + \omega^2} \\
\mathcal{L}[\langle x(t)v(0)\rangle] &= -\mathcal{L}[\langle v(t)x(0)\rangle] \\
\mathcal{L}[\langle x(t)x(0)\rangle] &= \frac{1}{\omega^2 s} - \frac{1}{s^2 + \gamma s^\alpha + \omega^2}.
\end{aligned} \tag{3.44}$$

To expand our past and future variables to include multiple time points, we extend the past variable to be observations between $t - t_0$ and t and the future variable to be $t + \Delta t$ to $t + \Delta t + t_0$. The size of the window is set by t_0. We discretize each window with a spacing of $dt = 1$ and compute correlation functions along the discrete points of time, yielding the full covariance matrices. Ideally, we would make the discretization interval arbitrarily small, $dt \to 0$. However, this introduces numerical issues, as the determinant of $\Sigma_{t-t_0:t}$ approaches 0. As such, we make dt as small as possible without causing this numerical issue. We explore a few values of dt in S2 Fig to determine the effect on the information curve. While small dt confers more information, there are diminishing returns and we asymptotically approach the correct values. After this, the recipe is as outlined in Section. 3.6.

3.9 Wright Fisher Dynamics

Wright-Fisher dynamics are used in population genetics to describe the evolution of a population of fixed size over generations. Here, we consider the diffusion approximation to the Wright-Fisher model with continuous time, given by Eq.18. We numerically integrate Eq.18 using a time step of $dt = 0.001$ and use 10000 data points starting from a given initial allele frequencies to estimate the joint distribution, $P(X_{t+\Delta t}, X_t)$. We discretize allele frequency space with $N + 1$ bins. We compute the maximum available predictive information for different values of the parameters (Figure 10) using

$$I(X_t; X_{t+\Delta t}) = -\sum_{X_t} \mathcal{P}(X_t) \log(\mathcal{P}(X_t)) - \sum_{X_{t+\Delta t}} \mathcal{P}(X_{t+\Delta t}) \log(\mathcal{P}(X_{t+\Delta t})) \\ + \sum_{X_t, X_{t+\Delta t}} \mathcal{P}(X_t, X_{t+\Delta t}) \log(\mathcal{P}(X_{t+\Delta t}, X_t)). \tag{3.45}$$

A simple estimate for $I(X_t; \tilde{X})$ can be obtained by considering the case where each individual memory reflects a distinct cluster of allele frequencies. In the optimal encoding case, each memory encodes an equal amount of probability weight on the input variable(80; 151). The upper bound on the information the representation variable has about the past state is $I(X_t; \tilde{X}) = \log(m)$.

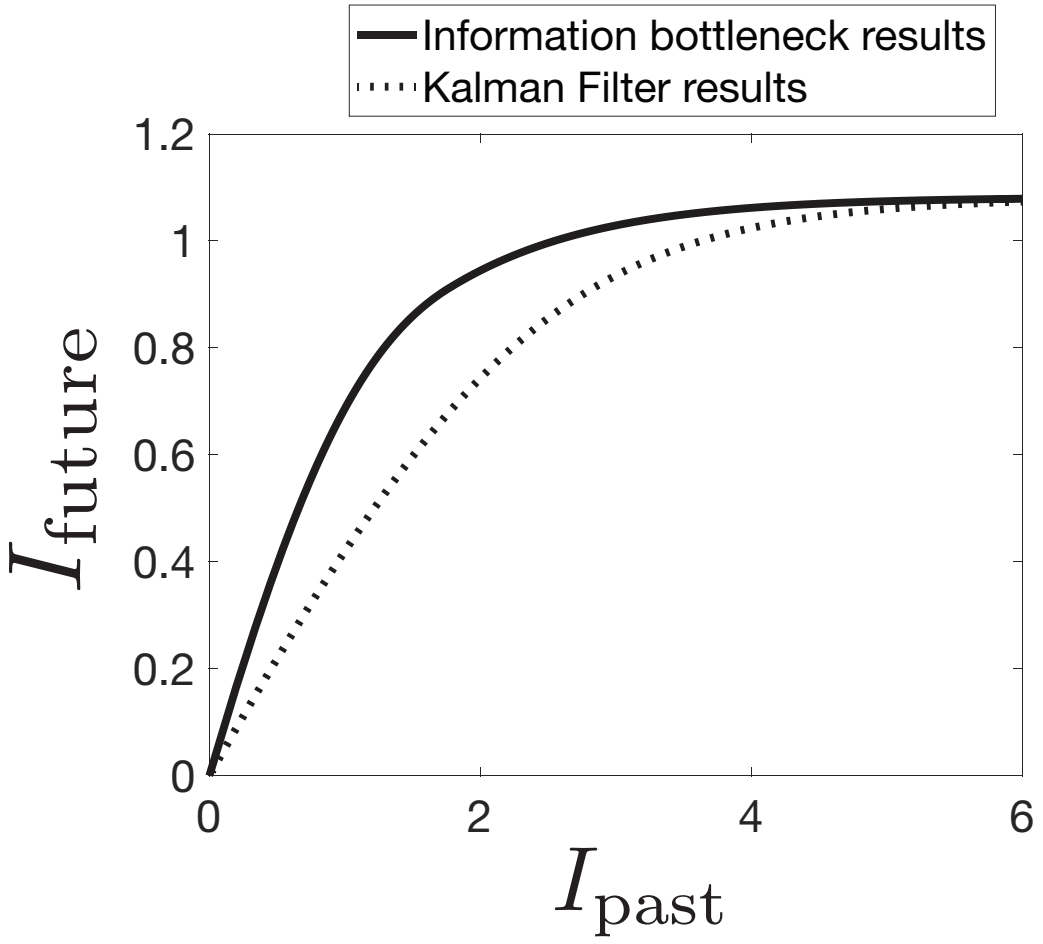

Fig 3.12: Kalman filtering schemes are not efficient coders for a given channel capacity. We compare the amount of information conferred about the future for a given encoding level and find that Kalman Filter-based approaches do not maximize the amount of predictive information conferred, suggesting they are not efficient predictive coding schemes.

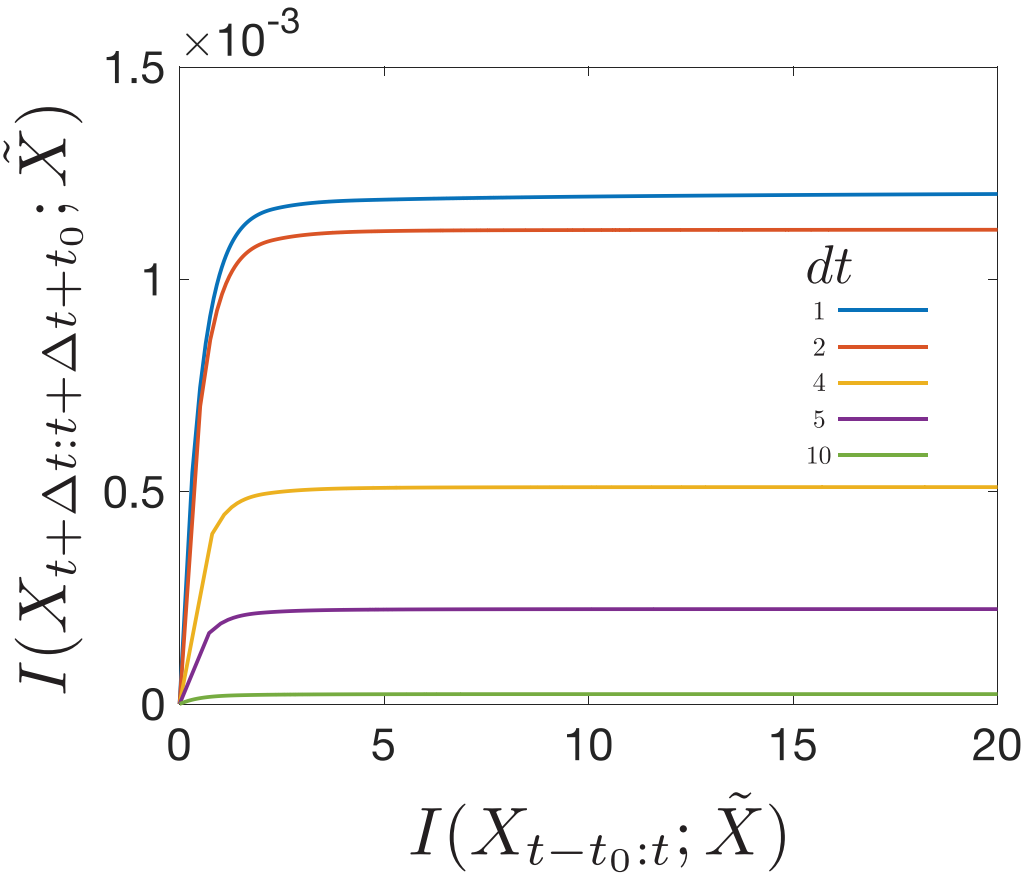

Fig 3.13: We plot the information curve for $\Delta t = 10$, $t_0 = 20$ for different values of dt. We note that there are diminishing returns for increasingly small dt. However, we cannot make dt arbitrarily small, as this introduces numerical errors.

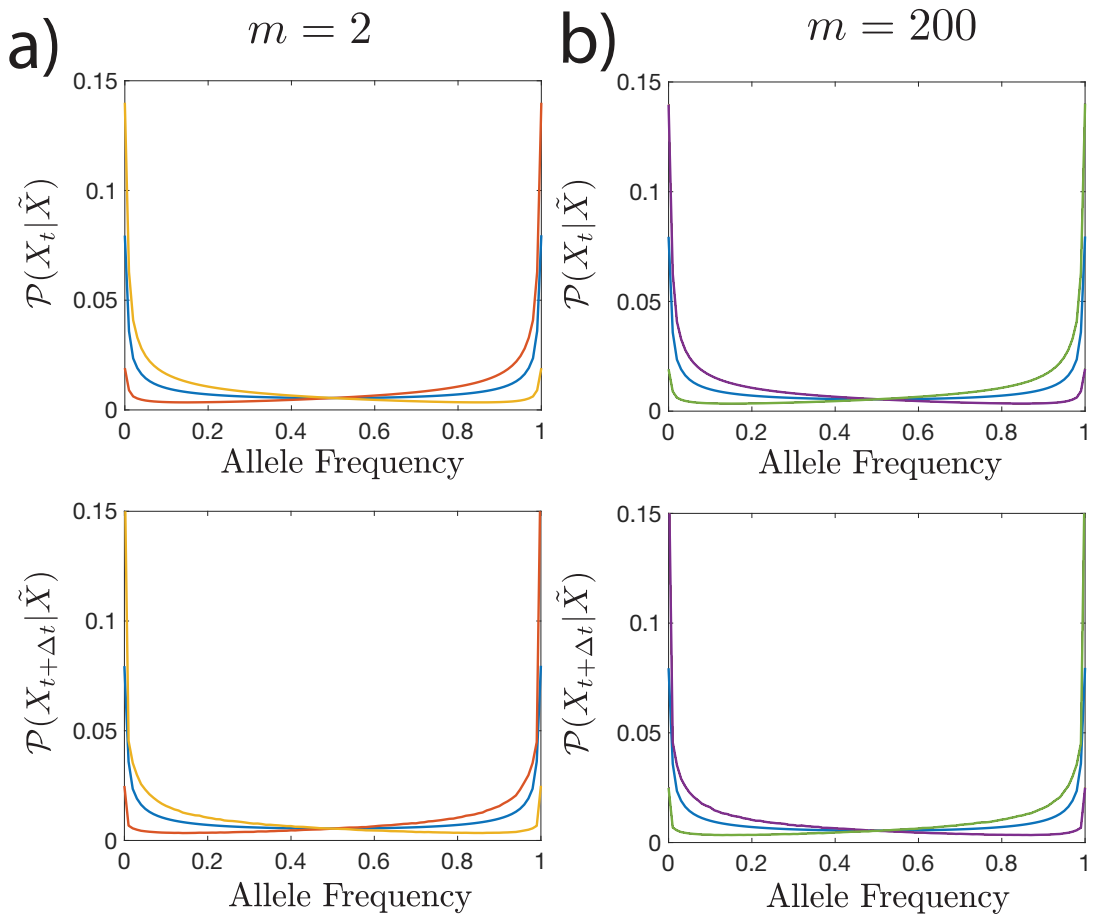

Fig 3.14: The optimal $P(X_t|\tilde{X})$ and $P(X_{t+\Delta t}|\tilde{X})$ for Wright Fisher dynamics with $N = 100$, $N\mu = 0.2$, $Ns = 0.001$, $\Delta t = 1$ with information bottleneck parameters $\beta = 1.01$ ($I(X_t;\tilde{X}) = 0.27$) for $m = 2$. (a) and $m = 200$ (b). Many representations are degenerate in the $m = 200$ in this limit. The encoding schemes for $m = 2$ versus $m = 200$ are nearly identical for this small $I(X_t;\tilde{X})$ limit.

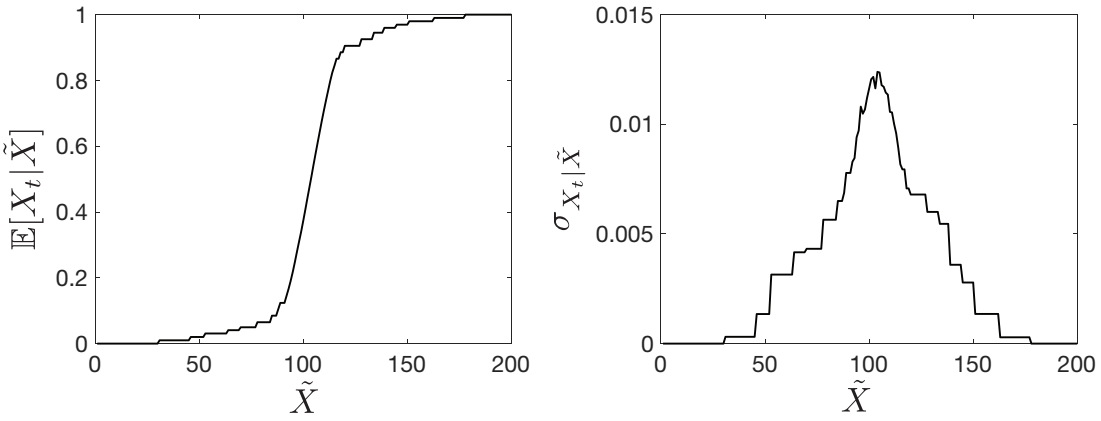

Fig 3.15: Mean (left) and variance (right) of the past allele frequency X_t conditioned on the (categorical) representation variable $\tilde{X}$ (left), for the information bottleneck solution of the Wright-Fisher dynamics with $m = 200$, $N = 100$, $N\mu = 0.2$, $Ns = 0.001$, $\beta = \infty$. The standard deviation is not constant: it is smaller where the prior probability of X_t is large.

CHAPTER 4

INFERRING COUPLINGS ACROSS ORDER-DISORDER PHASE TRANSITIONS

This work was pursued in collaboration with Vuditwat Ngampruetikorn, Hanna Torrence, Jan Humplich, David Schwab, and Stephanie Palmer.

4.1 abstract

Statistical inference is central to many scientific endeavors, yet how it works remains unresolved. Answering this requires a quantitative understanding of the intrinsic interplay between statistical models, inference methods and the structure in the data. To this end, we characterize the efficacy of direct coupling analysis (DCA)—a highly successful method for analyzing amino acid sequence data—in inferring pairwise interactions from samples of ferromagnetic Ising models on random graphs. Our approach allows for physically motivated exploration of qualitatively distinct data regimes separated by phase transitions. We show that inference quality depends strongly on the nature of data-generating distributions: optimal accuracy occurs at an intermediate temperature where the detrimental effects from macroscopic order and thermal noise are minimal. Importantly our results indicate that DCA does not always outperform its local-statistics-based predecessors; while DCA excels at low temperatures, it becomes inferior to simple correlation thresholding at virtually all temperatures when data are limited. Our findings offer new insights into the regime in which DCA operates so successfully and more broadly how inference interacts with the structure in the data.

4.2 Introduction

A quantitative understanding of the limitations and biases of inference methods is critical for developing high performing and trustworthy approaches to data analyzes. While emerging, such an understanding is incomplete, not least because it requires a thorough investigation of the intertwined nature of statistical models, inference methods and the structure in the data (187). Statistical physics models are ideally suited for this investigation for three main reasons. First, they often encompass the statistical models used in practice; take, for example, the Potts model in direct coupling analysis (DCA) (176; 109). Second, they enjoy a number of well-studied inference methods owing to a long history of inverse statistical physics problems (137; 117; 36). Third, they provide a controlled and physically motivated way to alter data-generating distributions across qualitatively distinct regimes. Adopting a statistical physics approach, we characterize the performance of DCA, one of the most oft-used tools in biological sequence analyzes, and highlight the importance of the structure in the data in quantifying the performance of inference methods.

DCA has proved successful as a technique for inferring the physical interactions that underpin the structure of biological molecules from amino acid sequence data (176; 109). This success has led to new insights into the protein folding problem (90) and how RNAs obtain their structures (43; 177; 170). The essence of DCA is to draw a distinction between direct and indirect correlations—those originating from direct physical interactions between two sites in a sequence and those mediated via other sites—by fitting a global statistical model to sequence data. But while DCA supersedes its local-statistics-based predecessors in virtually all applications, relatively little is known about the conditions that underlie its success (74).

The statistical model in DCA, well-known in physics as the Potts model (181), captures a phase transition that results from a competition between disorder-promoting thermal noise and order-promoting interactions. The disordered phase, which prevails at high temperatures,

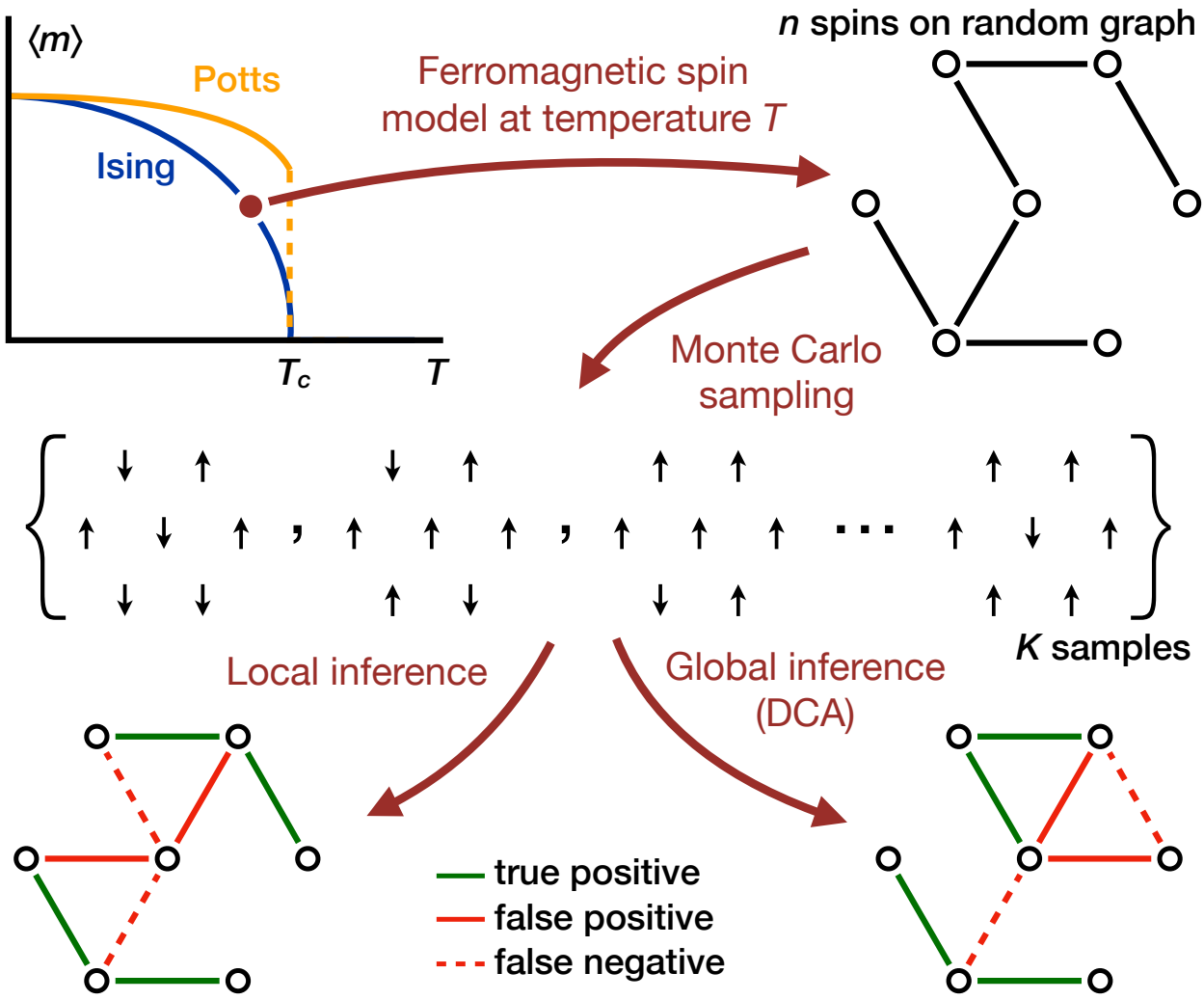

Fig 4.1: Data generation and inference. We generate samples from a ferromagnetic spin model on an Erdős-Rényi random graph and evaluate inference methods on the data at different model temperatures across order-disorder phase transitions. Direct coupling analysis ranks the likelihood of an interaction by leveraging global statistics whereas local inference uses pairwise statistics such as empirical correlations. We obtain predictions by thresholding the likelihood scores. In general, local and global inference method result in different predictions.

describes a system whose constituents (e.g., residues in a sequence) are largely uncorrelated; on the other hand, a macroscopic number of such constituents assume the same state in the low-temperature ordered phase. Both phases make for difficult inference: the data are noisy in the disordered phase and macroscopic ordering leads to strong indirect correlations in the ordered phase (104). A question arises as to the regime in which DCA operates so successfully and more broadly how the nature of data-generating distributions affects inference (see, also Ref (21)).

Recent work suggests that sequence data are drawn from distributions poised at the onset of order (156; 155). This regime sits at the boundary of the two phases, thus minimising the detrimental effects from thermal noise while avoiding precipitation of macroscopic order. In fact signatures of criticality—a defining property of a type of phase transitions—appear ubiquitous across a wide variety of biological systems (105; 17), including antibody diversity (106), genetic regulations (122; 77), neural networks (82; 33; 160; 161; 107; 32; 100), behaviors of individuals (37) and those of groups (20; 10). This apparent ubiquity has inspired a search for the origin of this behavior (146; 2; 110; 12) as well as work that attempts to uncover its function (99). However the structure of data distributions alone cannot capture the complete phenomenology of inference and as such cannot explain the success of DCA relative to local-statistics-based methods.

The use of the Potts model to capture correlations among constituents of a system is neither unique to DCA nor limited to analyzing sequence data. Indeed this approach is applicable to a range of biological systems from neural activity (145; 159) to flocks of birds (19). In addition, the Potts model is closely related to probabilistic graphical models and Markov random fields in probability theory, statistics and machine learning with applications including inferring interactions among genetic transcription factors (51) and computer vision (169). Understanding what affects the performance of DCA and when it outperforms local statistical inference is relevant to a large class of problems beyond the application of DCA in structural

biology.

Here we investigate the efficacy of DCA in inferring pairwise couplings from samples drawn from ferromagnetic spin models on random graphs at different temperatures across order-disorder phase transitions, see Fig 4.1. We demonstrate that the inference quality depends on data-generating distributions; in particular, better inference methods need not be more elaborate nor computationally more expensive. We show that a simple method based on thresholding pairwise correlations can easily outperform DCA at all temperatures in the under-sampled regime—a condition applicable to nearly all amino acid sequence datasets. We find further that more data improve DCA most significantly in the ordered phase where strong indirect correlations limit the performance of local methods. Interestingly we do not observe direct effects of criticality despite its association with diverging Fisher information (26; 65; 38; 95; 130). Instead we attribute the accuracy maximum at an intermediate temperature to the competition between the emergence of macroscopic order at low temperatures and high thermal noise level at high temperatures. Our work underscores the necessity to characterize the role of data-generating distributions when evaluating inference methods and offers a first step towards a deeper understanding of the intertwined nature of inference, models and the structure in the data.

4.3 Data Generation

To highlight the role of a phase transition, we consider the problem of reconstructing the interaction matrix of an Ising model on a random graph. A limiting case of the Potts model, the Ising model is one of the simplest models that captures a phase transition. It describes a system of n spins, $\vec{\sigma}=(\sigma_1,\sigma_2,\ldots,\sigma_n)$, each of which is a binary variable $\sigma_i\in\{\pm 1\}$. The spins interact via the Hamiltonian

$$\mathcal{H}(\vec{\sigma}) = -\sum_{i=1}^{n}\sum_{j=i+1}^{n} J_{ij}\sigma_i\sigma_j - \sum_{i=1}^{n} h_i\sigma_i, \tag{4.1}$$

where J_{ij} denotes the interaction between spins i and j, and h_i the bias field on spin i. The probability distribution of this system is given by

$$P(\vec{\sigma}) = \frac{e^{-\beta\mathcal{H}(\vec{\sigma})}}{\sum_{\vec{\sigma}'} e^{-\beta\mathcal{H}(\vec{\sigma}')}}, \tag{4.2}$$

where $\beta = 1/T$ is the inverse temperature and the summation is over all spin configurations.

Fig 4.1 provides an overview of our work. We generate samples from a uniform-interaction ferromagnetic Ising model on an Erdős-Rényi random graph,

$$\mathcal{H}^{\text{data}}(\vec{\sigma}) = -\sum_{i<j} J_{ij}\sigma_i\sigma_j \quad \text{with} \quad J_{ij} \sim \text{Bern}(\lambda/n) \tag{4.3}$$

for a graph with n vertices and mean degree λ. Each interaction is drawn from a Bernoulli distribution with parameter $p = \lambda/n$, i.e., an interaction is present ($J_{ij} = 1$) with probability p and absent ($J_{ij} = 0$) with probability $1 - p$. In the thermodynamic limit $n \to \infty$, a sharp transition exists between the high-temperature disordered phase and the low-temperature ordered phase. This phase transition is characterized by the order parameter $\Delta \equiv \frac{1}{n}|\langle\sum_i \sigma_i\rangle|$, which vanishes in the disordered phase and grows continuously with decreasing temperature in the ordered phase. A standard mean-field approximation yields the critical temperature $T_c = \lambda$ with the order parameter given by the largest root of the equation $\Delta = \tanh(\lambda\Delta/T)$. As a result, when the mean degree is relatively high, the effect of a change in λ is completely captured by critical temperature rescaling [see, also, Eq (4.34)]. Our results are based on samples generated with exact Monte Carlo sampling (131).

4.4 Mean-field Inversion

While several methods exist for the inverse Ising problem (117), we focus on the so-called naive mean-field inversion which forms the basis for a number of practically relevant algorithms (137; 109; 90; 152). Derived from a mean-field theory and the linear response theorem (68; 154)

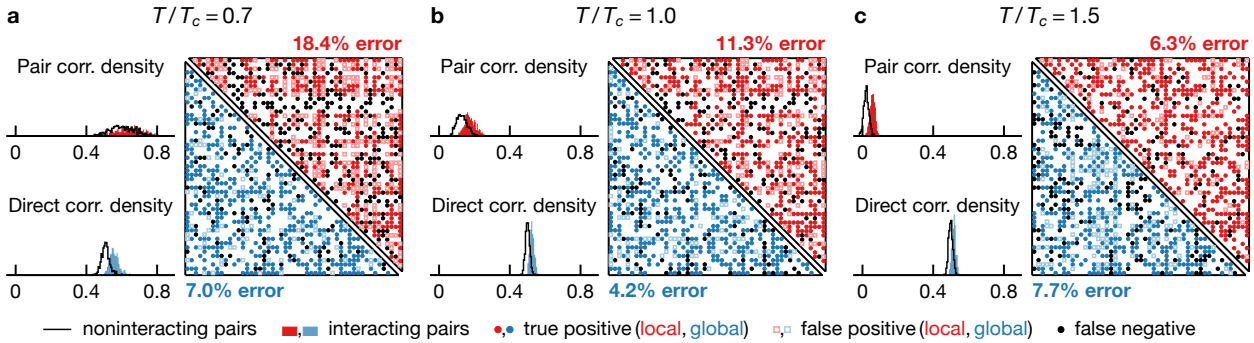

Fig 4.2: Local statistical modeling outperforms mean-field DCA in the disordered phase. We show density histograms of empirical and direct pair correlations—$\langle\sigma_i\sigma_j\rangle_{\text{data}}$ and $\langle\sigma_i\sigma_j\rangle_{\text{dir}}$ [see, Eq (4.5)]—for interacting (filled) and non-interacting (line) pairs of spins at $T/T_c = 0.7, 1.0, 1.5$ (**a-c**, respectively). The predictions of pairwise interactions are depicted in a contact map for local (upper half) and global (lower half) inference. The discrimination threshold is chosen such that the number of positive predictions is equal to the number of real interactions, and false positives and false negatives are equal (see legend). In general both empirical and direct pair correlations are higher among interacting spins and are thus informative of interactions. For local inference, the prediction error decreases with temperature and is smaller than that of global inference at $T/T_c = 1.5$ (**c**). Global inference error exhibits non-monotonic temperature dependence and is minimal at an intermediate temperature $T/T_c = 1.0$ (**b**). Shown results are based on 5×10^3 samples drawn from Ising models on an Erdős-Rényi graph with 50 vertices and mean degree 20.

(see, Appendix 4.8), the naive mean-field inversion expresses interactions J_{ij} in terms of empirically accessible connected correlation matrix C,

$$\beta J_{ij} = -(C^{-1})_{ij} \quad \text{for } i < j, \tag{4.4}$$

where $C_{ij} \equiv \langle\sigma_i\sigma_j\rangle - \langle\sigma_i\rangle\langle\sigma_j\rangle$. In the following, global statistical inference refers to the naive mean-field inversion.

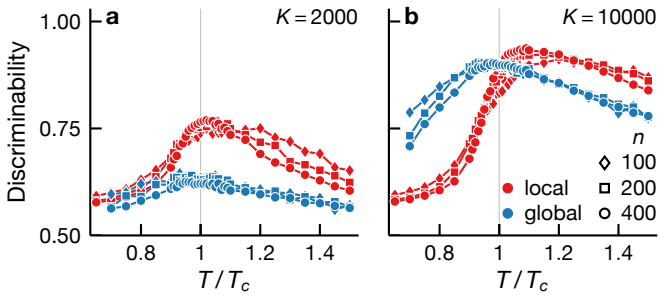

Fig 4.3:
Local inference is more data efficient but more severely affected by macroscopic order.
We depict the local (red) and global (blue) inference discriminability of interactions (area under the ROC curve) for Ising models on Erdős-Rényi graphs with mean degree 40 and different number of vertices n (see legend) for sample sizes $K = 2\times10^3$ and 10^4 (**a** and **b**, respectively). Both local and global inference exhibits discriminability maximum near T_c. Local inference is more discriminating at all temperatures when the data are limited (**a**). But global inference performs better in the ordered phase when more data are available (**b**).

4.5 Results

4.5.1 Discriminability of interactions

One measure of inference quality is the ability to discriminate directly interacting spin pairs from those that interact only via other spins. Fig 4.2 visualizes this discrimination based on local and global statistical inference. For each spin pair, we assign a score that ranks the likelihood of an interaction being present; here, we use empirical correlations $\langle\sigma_i\sigma_j\rangle_{\text{data}}$ and direct correlations $\langle\sigma_i\sigma_j\rangle_{\text{dir}}$ in local and global inference, respectively. The average $\langle\cdots\rangle_{\text{data}}$ is taken with respect to the empirical distribution and $\langle\cdots\rangle_{\text{dir}}$ to the direct pairwise distribution (176),

$$\hat{P}_{ij}^{\text{dir}}(\sigma_i,\sigma_j) \equiv \frac{\exp(\beta\hat{J}_{ij}\sigma_i\sigma_j + \tilde{h}_i\sigma_i + \tilde{h}_j\sigma_j)}{\sum_{\sigma_i',\sigma_j'}\exp(\beta\hat{J}_{ij}\sigma_i'\sigma_j' + \tilde{h}_i\sigma_i' + \tilde{h}_j\sigma_j')} \tag{4.5}$$

where $\hat{J}_{ij}$ denotes the inferred interactions from naive mean-field inversion and the fields $\tilde{h}_i$ and $\tilde{h}_j$ are chosen such that the marginal distributions coincide with empirical single-spin distributions. In Fig 4.2, we see that on average both empirical and direct correlations are

higher among interacting pairs and are thus predictive of true interactions. To turn the likelihood scores into concrete predictions, we need to define a threshold which separates positive and negative predictions. We choose a discrimination threshold that equates the number of positive predictions to the number of true interactions and display inference predictions and errors as a contact map (Fig 4.2a-c). The accuracy of the global approach exhibits non-monotonic temperature dependence with higher error rates at temperatures above and below T_c. In contrast the accuracy of local inference increases with temperature over the range shown in Fig 4.2. (But note that the accuracy must eventually go down at adequately high temperatures, see Fig 4.3.) While the error rate of global inference is less than half of that of local inference at low temperatures (Fig 4.2a-b), a local statistical approach outperforms global inference at high temperature (Fig 4.2c; see also, Fig 4.3).

Although specifying a discrimination threshold allows us to make concrete predictions, its choice is often arbitrary. We now consider a more general measure of discriminability grounded in receiver operating characteristic (ROC) analysis. ROC analysis constructs a curve that traces the true and false positive rates as the discrimination threshold varies. In the following, we identify discriminability with the area under the ROC curve which is equal to the probability that a real positive scores higher than a real negative.

Local and global statistical inference exhibits qualitatively different sample size dependence, see Fig 4.3. At low samples, local inference is more discriminating than naive mean-field inversion at all temperatures (Fig 4.3a). This behavior is a result of the distinct natures of local and global approaches. Global inference requires a good estimate of the full joint distribution whereas local inference relies only on pairwise distributions which are much easier to estimate, especially with limited samples. An increase in samples improves both local and global inference but this improvement diminishes for local inference at low temperatures (Fig 4.3b). This results from the fact that the entropy of the model increases with temperature and thus, given a fixed number of samples, a low-temperature model is better sampled. In

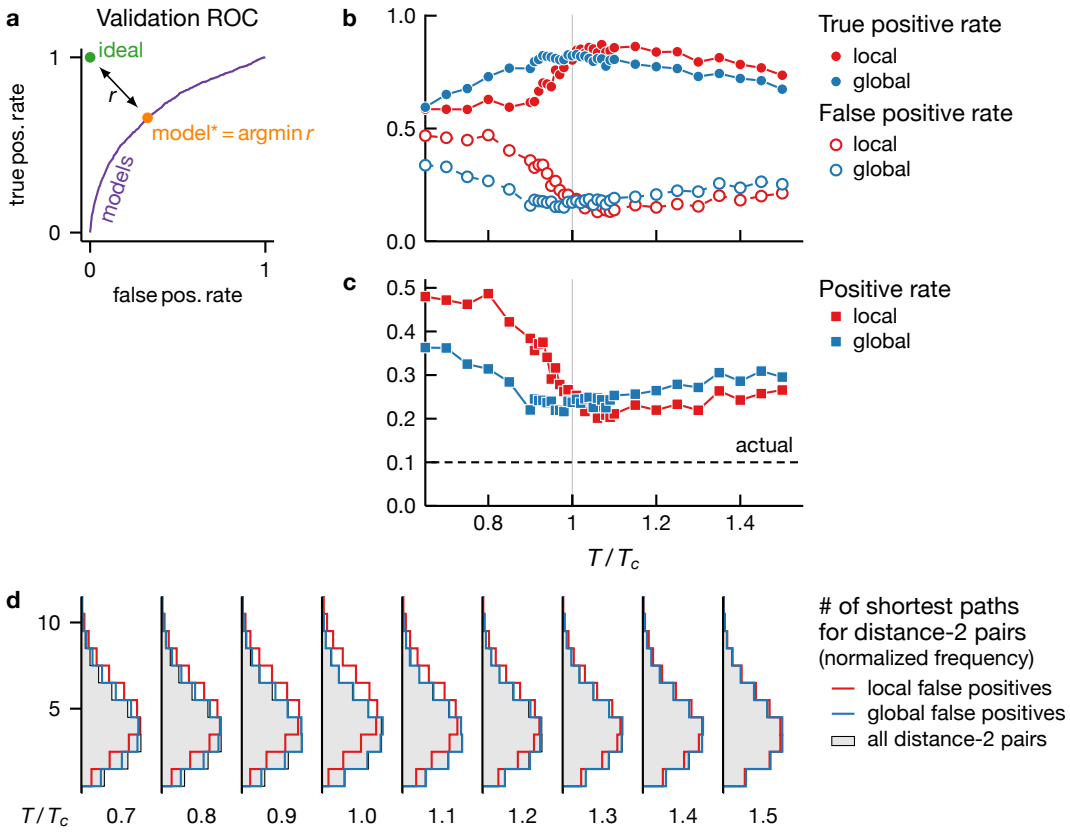

Fig 4.4:
We use 20% of pairs chosen at random (validation set) to compute the discrimination threshold (**a**) and report inference properties on the rest (test set, **b-d**). **a** Typical ROC curve for the validation set. We choose a threshold such that the resulting model is closest to the ideal model, as measured by the Euclidean distance in the ROC space. **b** True and false positive rates *vs* temperature. Both local and global methods are most accurate at a temperature close to T_c but local inference worsens faster at low temperatures. **c** Temperature dependence of the positive rate (the ratio between positive predictions and all pairs). Over-prediction is most acute for local inference at low temperatures. **d** Distribution of the number of shortest paths among false positive pairs with graph distance two at different temperatures. At low temperatures the false positives from local inference contain a larger fraction of highly connected pairs, compared to all pairs with distance two (grey) as well as to the false positives from global inference. Thus non-interacting pairs in denser parts of the graph are likelier to be mis-classified than those in sparser parts. Shown results are based on 10^4 samples from an Ising model on an Erdős-Rényi graph with 400 vertices and mean degree 40.

Fig 4.3a, pairwise distributions are already well-sampled at low temperatures and more samples do not lead to higher accuracy for local inference (Fig 4.3b). However well-sampled pairwise distributions do not imply a good estimate of the full distribution; indeed, more samples improve the discriminability of global inference in the low-temperature regimes, i.e., the blue points in Fig 4.3b are higher than in Fig 4.3a below T_c.

Inference performance depends not only on well-measured probability distributions but also the structure of the distributions. Despite having lower entropy and being better sampled, low-temperature models are more difficult to infer compared to those in the vicinity of the phase transition, see Fig 4.3. This feature is a consequence of macroscopic ordering below T_c. In the ordered phase, two spins are likely to align regardless of the presence of an interaction and therefore pair correlations become less discriminating. While the decrease in discriminability affects both local and global inference, its effect is less severe for global inference (Fig 4.3). The use of global statistics—statistical quantities that require measurements of the entire system such as the inverse connected correlation matrix—helps avoid direct comparisons between spin pairs in dense clusters of the interaction graph and those in sparser parts.

4.5.2 The effects of local interaction networks on inference

Indeed local inference is more likely to mis-classify well-connected non-interacting spin pairs. To illustrate this point, we randomly divide all of the spin pairs into two disjoint sets for validation and testing. We use the validation set to determine a discrimination threshold and report inference quality on the test set. In Fig 4.4 we use 20 percent of pairs in validation and choose the discrimination threshold such that the resulting true and false positive rates are closest to that of ideal classifiers, as measured by the Euclidean distance in the ROC plane (Panel A). Note that while the Euclidean distance is not the only possibility, the concavity of the ROC curve means our results remain qualitatively the same for any metric based on

ℓ_p-norm with $p \geq 1$. Fig 4.4b and c show that the quality of local inference deteriorates faster as temperature decreases below T_c—i.e., decreasing true positive rate, increasing false positive rate and more over-prediction (excess positive predictions compared to ground truth).

We characterize the false positives (mis-classified non-interacting pairs) by the number of shortest paths between spins in each pair (Fig 4.4d). Here we focus only on pairs with a graph distance of two (less than two percent of pairs have distance greater than two for this particular graph). At high temperatures the distribution of the number of shortest paths among false positives is the same as that for non-interacting pairs; that is, any non-interacting pair is equally likely to be mis-classified. As temperature lowers to around T_c, the false positives from local inference contain a disproportionately large fraction of pairs that are connected by more paths. This behavior is a direct consequence of the emergence of order which generates strong correlations, especially among pairs in denser parts of the graph. At very low temperatures, macroscopic order proliferates and pair correlations are strong regardless of the number of paths or physical interactions. While this effect reduces the disproportionate mis-classification among better connected pairs, it increases the discrepancy between the predicted and actual positive rates (Fig 4.4c). In fact the positive rate of $\sim$50 percent results from the fact that any pair leads to a positive prediction with probability $\frac{1}{2}$. We see that in contrast to local inference, mean-field DCA is less likely to confound path multiplicity with interactions, especially close to the onset of order. In addition it suffers less from strong indirect correlations as evidenced by smaller over-prediction rates at low temperatures. In sum, leveraging global statistics helps DCA draw a better distinction between direct and indirect correlations, thus making it more accurate at low temperatures.

4.5.3 Root-mean-square error of inferred couplings

While a useful characterization of discriminability, ROC analysis is agnostic about the magnitude of the inferred interactions. We now show that the root-mean-square (RMS)

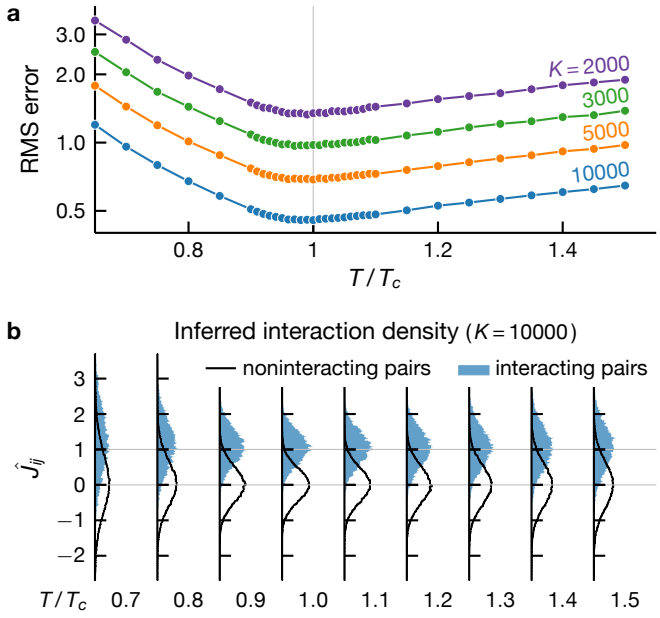

Fig 4.5:
Interactions inferred from mean-field DCA are statistically unbiased with smallest variances around phase transitions.
a Root-mean-square error of inferred interactions as a function of temperature at different sample sizes K (see legend). **b** Density histograms of inferred interactions for non-interacting and interacting pairs whose true interactions are one and zero, respectively. Shown results are for an Ising model on an Erdős-Rényi graph with 400 vertices and mean degree 40.

error of the interactions inferred by naive mean-field inversion exhibits similar temperature dependence to discriminability. In Fig 4.5a, we see that the RMS error is smallest at a temperature slightly below T_c for a range of sample sizes. Fig 4.5b reveals the origin of this temperature dependence. On average mean-field inversion correctly predicts the interactions—$J_{ij} \in \{0, 1\}$ depending on whether an interaction is present—but the prediction variance is minimum around T_c. Above T_c, an increase in temperature leads to a model with higher entropy, thus requiring a larger number of samples to maintain inference accuracy. Below T_c, macroscopic order interferes with inference by generating strong indirect correlations among non-interacting pairs.

4.5.4 The role of data-generating models

Since inference quality is intrinsically a combined property of inference methods and data distributions, it is *a priori* unclear whether the observed non-monotonic temperature dependence (Figs 4.3 and 4.5) originates from the inductive bias in inference methods or the structure in the data. To isolate the role of data-generating models, we consider the response of data distributions to a change in model parameters as a proxy for how informative a data point is about model parameters. We quantify the distributional response by the f-divergence, an information-theoretic distance between two distributions, defined via $\mathrm{D}_f(P_X\|Q_X)\equiv\langle f(P_X/Q_X)\rangle_{X\sim Q_X}$ where $f:[0,\infty)\to(-\infty,\infty)$ is convex and $f(1)=0$. The f-divergence between two zero-field Ising models on different graphs, parametrized by J and J', reads [see, Eqs (4.2) and (4.3)]

$$\mathrm{D}_f(J',J)=\left\langle f\left(\frac{e^{\beta\sum_{i<j}\Delta J_{ij}\sigma_i\sigma_j}}{\left\langle e^{\beta\sum_{i<j}\Delta J_{ij}\sigma'_i\sigma'_j}\right\rangle_{\vec{\sigma}'\sim\mathcal{H}_J}}\right)\right\rangle_{\vec{\sigma}\sim\mathcal{H}_J}, \tag{4.6}$$

where $\Delta J=J'-J$ and the average $\langle\dots\rangle$ is with respect to the model on the graph J.

Before we discuss the numerical results, it is instructive to derive an expression for the f-divergence in a mean-field approximation. Expanding Eq (4.6) around $\beta=0$ and taking $P(\vec{\sigma})=\prod_i\frac{1}{2}(1+\sigma_i\Delta)$ yield

$$\mathrm{D}_f^{\mathrm{mf}}(J',J)=\frac{1}{2}f''(1)\|\Delta J\|_1\frac{1-\Delta(T)^4}{T^2}, \tag{4.7}$$

where $\Delta(T)$ is the mean-field order parameter and the ℓ_1-norm $\|\Delta J\|_1$ counts the number of different edges in J and J'. Note that the elements of J and J' are either zero or one and we set $J_{ij}=0$ for $i\geq j$ as they do not enter the model [see, Eq (4.3)]. In the disorder phase $T>T_c$, high noise level makes models less dependent on the parameters and the f-divergence decays as T^{-2}. The dependence on the order parameter means different parameters also result in

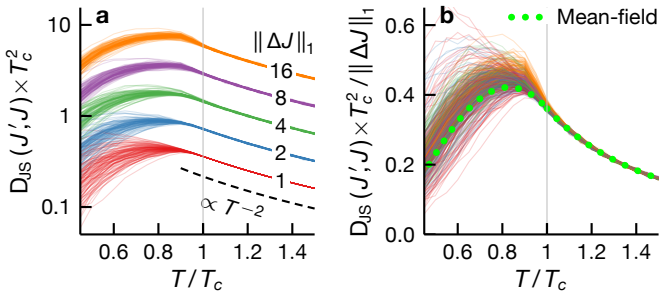

Fig 4.6: Jensen-Shannon (JS) divergence between two Ising models *vs* temperature. **a** JS divergences computed from 10^4 samples using Eq (4.6) for a fixed graph J and many realizations of J' generated by randomly deleting and adding edges to J. The curves are grouped by the number of different edges in J and J' (see legend). **b** empirical JS divergences compared to a mean-field prediction Eq (4.7), showing good agreement for $T > T_c$ (same color code as in **a**). Here J is an Erdős-Rényi graph with 400 vertices and mean degree 40.

more similar models at low temperatures (since $\Delta(T) \to 1$ as $T \to 0$). Indeed the competition between thermal noise and macroscopic order leads to a maximum at $T/T_c \approx 0.83$.

Fig 4.6 illustrates the temperature dependence of the f-divergence between two Ising models. Here we adopt the Jensen-Shannon (JS) divergence which is an f-divergence defined with $f(t) = (t+1)\log_2 \frac{2}{t+1} + t \log_2 t$. We compute the divergence $\mathrm{D}_{\mathrm{JS}}(J', J)$ from data using Eq (4.6) for a fixed Erdős-Rényi graph J and we generate J' by randomly deleting and adding edges in J, allowing J and J' to have different numbers of edges. We see that, as expected from the mean-field analysis, the f-divergence decays as T^{-2} at high temperatures and peaks at a temperature below T_c with its scale controlled by the number of different interactions in J and J' (Fig 4.6a). In Fig 4.6b, we compare the empirical JS-divergence to the mean-field approximation [Eq (4.7)] and find good agreement for $T > T_c$. Below T_c, the mean-field result only captures the qualitative behavior due to large variance in the JS divergence (from different realizations of J'). This is an expected result since the locations where macroscopic order nucleates depend on graph structure and a change to which can yield a range of divergences.

4.5.5 Inference discriminability for Potts models

It is tempting to view the inference quality maximum as a manifestation of critical phenomena, not least because the Fisher information (magnetic susceptibility) diverges at T_c (26; 65; 38; 95; 130). However criticality does not seem to play an important role in inferring the interaction graph. Indeed Fig 4.6 illustrates that the distance between two models on different graph varies smoothly across the critical temperature.

To elaborate this point further, we consider q-state Potts models on an Erdős-Rényi random graph which generalizes the binary spins in Ising models to q states. Unlike the Ising model, a q-state Potts model with $q>2$ exhibits a discontinuous phase transition which does not display critical behaviors and at which the susceptibility remains finite. Fig 4.7 compares the inference discriminability for 3 and 4-state Potts models with that for Ising models ($q=2$). We use the naive mean-field inversion, generalized to Potts models (109) for both Ising and Potts models (see, Appendix 4.8). In Fig 4.7, we see that, in the disordered phase, the discriminability for Potts and Ising models shows similar dependence on sample size and temperature. In the ordered phase, the inference quality decreases with temperature and worsens with increasing q. This q-dependence results from the fact that macroscopic order forms more rapidly for larger q with order parameter discontinuity growing with q, see Appendix 4.9 [Eq (4.36)]. In fact, Fig 4.7b illustrates that the inference discriminability for Potts and Ising cases displays similar dependence on the mean-field order parameter (for a mean-field analysis of the Potts model, see Ref (181) and Appendix 4.7), thus suggesting that macroscopic ordering rather than criticality is an important determinant of inference performance.

4.6 Discussion

Despite being more elaborate and computationally more expensive than local statistical approaches, mean-field DCA does not always lead to better inference quality. Indeed we

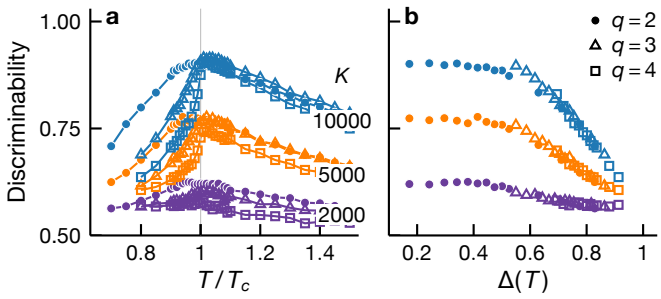

Fig 4.7:
Interaction discriminability for Ising and Potts models.
Discriminability maximum results from the competition between thermal noise and macroscopic ordering but is not a signature of criticality associated with second-order phase transitions. We show DCA discriminability at different sample sizes K (see legend) as a function of temperature (**a**) and mean-field order parameter Δ (**b**). In all cases, discriminability peaks at an intermediate temperature and displays similar temperature dependence above T_c. By plotting discriminability as a function of Δ for $T < T_c$, we see that different temperature dependence for Ising and Potts models at $T < T_c$ originates from the fact that macroscopic order forms more rapidly in Potts models which admit first-order phase transitions. This highlights the detrimental effect of macroscopic order on inference quality. Shown results are based on the same Erdős-Rényi interaction graph with 400 vertices and mean degree 40.

show that local statistical methods can be more accurate when data are limited. More generally, although global statistics encode more information that could potentially improve inference, they are more difficult to estimate in the under-sampled regime. Inference quality depends not only on sample size but also the nature of data distributions. A low-temperature generative model, while better-sampled due to lower entropy, is more difficult to infer, compared to higher-temperature models around the phase transition. This feature highlights how macroscopic ordering, and more broadly data distributions, can interfere with inference. For models exhibiting an order-disorder phase transition, we find that DCA provides the most advantage over local statistical modeling in the ordered phase and when the systems are relatively well-sampled. Our results highlight the fact that inference quality can only be quantified with respect to the structure in the data and illustrate the central role of data-generating distributions in understanding inductive biases of inference methods (41). Finally our work lays a foundation for future investigations seeking to provide a prescription

for inference method selections based on the structure in the data.

While we consider ferromagnetic models on relatively dense interaction networks, our analysis yields qualitative insights applicable to models with sparser interactions. In particular we expect better performance from local inference as each spin pair becomes less connected (see, Sec 4.5.2). In addition, the increased probability of isolate spins means that the connected correlation matrix is more likely to be singular, thus making naive mean-field inversion ill-defined without regularization. A quantitative study of inference for models on sparse networks is an interesting research direction, not least because of the important role of fluctuations in such models.

Although we base our analysis on naive mean-field inversion, a number of methods exist for inferring pairwise interactions (see, e.g., Ref (117)). The general conclusion of our work also applies to these methods; the inference quality must depend on the structure in the data-generating distribution as well as the number of available observations. Revealing the optimal setting for each of these methods is likely to require generative models that capture different types of correlations in the system, and is a promising avenue for future research. For the ferromagnetic model considered here, we expect that our qualitative results hold for other inference methods, not least because the inference performance maximum near the phase transition stems from the property of the generative model (see, Sec 4.5.4).

To isolate the role of a phase transition, we specialize our analysis to uniform-interaction models on Erdős-Rényi random graphs which tend to be less structured than interaction graphs of real systems. For example, the structural organization of proteins leads to a hierarchy of sectors of strongly interacting amino acids (59). Spin models on hierarchical random graphs also capture order-disorder phase transitions (48) and it would be interesting to investigate how such a structure affects inference. Another promising future direction is to extend our analysis beyond ferromagnetic models to systems with richer phase diagrams such as spin-glass models and sparse Hopfield networks.

4.7 Graphical Potts models

Potts models describe a system of q-state spins $\vec{\sigma} = (\sigma_1, \sigma_2, \ldots, \sigma_n)$ with $\sigma_i \in \{1, 2, \ldots, q\}$, interacting via the Hamiltonian,

$$\mathcal{H}(\vec{\sigma}) = -\sum_{i=1}^{n} \sum_{j=i+1}^{n} J_{ij}(\sigma_i, \sigma_j) - \sum_{i=1}^{n} h_i(\sigma_i). \tag{4.8}$$

The probability distribution of this system is given by

$$P(\vec{\sigma}) = \frac{e^{-\beta \mathcal{H}(\vec{\sigma})}}{\sum_{\vec{\sigma}'} e^{-\beta \mathcal{H}(\vec{\sigma}')}}. \tag{4.9}$$

This measure is invariant under the gauge transformation,

$$\begin{aligned} h_i(\mu) &\to h_i(\mu) + \phi_i + \sum\nolimits_{j}^{j \neq i} \Lambda_{ij}(\mu) \\ J_{ij}(\mu, \nu) &\to J_{ij}(\mu, \nu) - \Lambda_{ij}(\mu) - \Lambda_{ji}(\nu) + \psi_{ij} \end{aligned} \tag{4.10}$$

for any ϕ_i, ψ_{ij} and $\Lambda_{ij}(\mu)$. This gauge symmetry means that the Potts measure is characterized by $\binom{n}{2}(q-1)^2 + n(q-1)$ independent parameters, which is the same number of independent parameters in single and two-spin distributions, $P(\sigma_i)$ and $P(\sigma_i, \sigma_j)$ (see, e.g., Ref (109)). Indeed for specified $P(\sigma_i)$ and $P(\sigma_i, \sigma_j)$ the Potts measure is the unique maximum-entropy model (109). Another consequence of the gauge invariance is that a family of model parameters (J, h) can result in the same measure. As a result, inference methods that produce a unique set of parameters must invoke gauge fixing conditions (either explicitly or via implicit regularization).

4.8 Mean-field inversion

For completeness, we reproduce the derivation of the mean-field inversion method for Potts models from Ref (109). We define the free energy

$$\mathcal{F} = \mathcal{F}(J,h) = -\ln \sum\nolimits_{\vec{\sigma}} e^{-\beta \mathcal{H}(\vec{\sigma})} \tag{4.11}$$

It follows that the first and second-order derivatives of this free energy are related to the single-spin and pairwise distributions via

$$\frac{\partial \mathcal{F}}{\partial h_{i\mu}} = -P_{i\mu} \quad \text{and} \quad \frac{\partial^2 \mathcal{F}}{\partial h_{i\mu} \partial h_{j\nu}} = -P_{i\mu,j\nu} + P_{i\mu} P_{j\nu} \tag{4.12}$$

where we introduce the shorthand notations

$$\begin{aligned} h_{i\mu} &= h_i(\sigma_i = \mu), & J_{i\mu,j\nu} &= J_{ij}(\sigma_i = \mu, \sigma_j = \nu), \\ P_{i\mu} &= \sum\nolimits_{\vec{\sigma}} \delta_{\sigma_i,\mu} P(\vec{\sigma}), & P_{i\mu,j\nu} &= \sum\nolimits_{\vec{\sigma}} \delta_{\sigma_i,\mu} \delta_{\sigma_j,\nu} P(\vec{\sigma}). \end{aligned}$$

Eq (4.12) also implies

$$\frac{\partial P_{i\mu}}{\partial h_{j\nu}} = P_{i\mu,j\nu} - P_{i\mu} P_{j\nu} \equiv C_{i\mu,j\nu} \tag{4.13}$$

where $C_{i\mu,j\nu}$ denotes the connected correlation matrix.

4.8.1 Gauge fixing

To infer a unique set of model parameters, we adopt the lattice-gas gauge which explicitly limits the model parameters to those that are independent (see, Eq (4.10) and the text around it). In this gauge each spin has a gauge state, c_i for spin i, for which the pairwise coupling and local field vanish, i.e.,

$$\forall \vec{\sigma}, i, j : J_{ij}(\sigma_i, c_j) = J_{ij}(c_i, \sigma_j) = h_i(c_i) = 0 \tag{4.14}$$

We assume this gauge in the following analysis unless specified otherwise.

4.8.2 Legendre transformation

Since the local field $h_{i\mu}$ is conjugate to the single-spin distributions $P_{i\mu}$ (see, Eq (4.12)), we can define a Legendre transform of the free energy

$$\mathcal{G} = \mathcal{F} + \sum\nolimits_{i\mu} h_{i\mu} P_{i\mu}. \tag{4.15}$$

Note that $\mathcal{G}$ does not depends explicitly on the probability of the gauge state P_{ic_i}; it is left out of the summation by the gauge condition $h_{ic_i}=0$ [Eq (4.14)]. In this ensemble the local fields are given by

$$h_{i\mu} = \frac{\partial \mathcal{G}}{\partial P_{i\mu}}. \tag{4.16}$$

Taking the derivative of the above equation yields

$$\frac{\partial h_{i\mu}}{\partial P_{j\nu}} = \frac{\partial^2 \mathcal{G}}{\partial P_{i\mu} \partial P_{j\nu}} = (C^{-1})_{i\mu,j\nu} \tag{4.17}$$

where the last equality follows from Eq (4.13) and the fact that the first-order derivatives of a function and its Legendre transform are inverse functions of one another. Note that the indices $(i\mu, j\nu)$ in Eqs (4.16) and (4.17) do not include the gauge states.

4.8.3 Small-coupling expansion

To derive the mean-field inversion, we consider a systematic expansion around the non-interacting Hamiltonian, treating the coupling term as a perturbation (57; 186),

$$-\beta \mathcal{H}_\alpha(\vec{\sigma}) = \alpha \sum\nolimits_{i<j} J_{ij}(\sigma_i, \sigma_j) + \sum\nolimits_i h_i(\sigma_i) \tag{4.18}$$

where the parameter α tunes the interaction strength: $\mathcal{H}_0$ corresponds to the non-interacting case and $\mathcal{H}_1$ to the original Hamiltonian. Expanding $\mathcal{G}$ as a power series in α yields

$$\mathcal{G}_\alpha = \mathcal{G}_0 + \mathcal{G}_0'\alpha + \tfrac{1}{2}\mathcal{G}_0''\alpha^2 + \mathcal{O}(\alpha^3) \tag{4.19}$$

where $\mathcal{G}_\alpha' = d\mathcal{G}_\alpha/d\alpha$ and $\mathcal{G}_\alpha'' = d^2\mathcal{G}_\alpha/d\alpha^2$. Substituting the above expression in Eqs (4.16) and (4.17) gives

$$\begin{aligned} h_{i\mu} &= \frac{\partial \mathcal{G}_0}{\partial P_{i\mu}} + \frac{\partial \mathcal{G}_0'}{\partial P_{i\mu}}\alpha + \mathcal{O}(\alpha^2) \\ (C^{-1})_{i\mu,j\nu} &= \frac{\partial \mathcal{G}_0}{\partial P_{i\mu}\partial P_{j\nu}} + \frac{\partial \mathcal{G}_0'}{\partial P_{i\mu}\partial P_{j\nu}}\alpha + \mathcal{O}(\alpha^2) \end{aligned}, \tag{4.20}$$

for $i\mu \neq ic_i$ and $j\nu \neq jc_j$.

4.8.4 Zeroth order

When $\alpha = 0$, the spins decouple and the free energy reads

$$\mathcal{F}_0 = -\sum\nolimits_i \ln \sum\nolimits_\nu e^{h_{i\nu}} \tag{4.21}$$

From Eq (4.12), we have $P_{i\mu} = e^{h_{i\mu}}/\sum_\nu e^{h_{i\nu}}$ and

$$\mathcal{G}_0 = \sum_{i\mu \neq ic_i} P_{i\mu} \ln P_{i\mu} + \sum_i \left(1 - \sum_{\nu \neq c_i} P_{i\nu}\right) \ln \left(1 - \sum_{\nu \neq c_i} P_{i\nu}\right). \tag{4.22}$$

Taking the derivatives, we have

$$\frac{\partial \mathcal{G}_0}{\partial P_{i\mu}} = \ln \frac{P_{i\mu}}{P_{ic_i}} \quad \text{and} \quad \frac{\partial^2 \mathcal{G}_0}{\partial P_{i\mu}\partial P_{j\nu}} = \delta_{ij}\left(\frac{\delta_{\mu\nu}}{P_{i\mu}} + \frac{1}{P_{ic_i}}\right), \tag{4.23}$$

where $P_{ic_i} = 1 - \sum_{\mu \neq c_i} P_{i\mu}$. We note that the pairwise coupling does not appear in the zeroth-order expansion.

4.8.5 First order

Differentiating the thermodynamic potential $\mathcal{G}_\alpha$ with respect to α gives

$$\mathcal{G}'_\alpha = -\sum_{\vec{\sigma}} \frac{e^{-\beta\mathcal{H}_\alpha(\vec{\sigma})}}{\sum_{\vec{\sigma}'} e^{-\beta\mathcal{H}_\alpha(\vec{\sigma}')}} \sum_{i<j} J_{ij}(\sigma_i, \sigma_j). \tag{4.24}$$

Note that the expression for $\mathcal{G}_\alpha$ can be obtained from Eqs (4.11) and (4.15) for the small-coupling Hamiltonian in Eq (4.18). In the limit $\alpha \to 0$, the Boltzmann weight becomes that of the non-interacting system and the above equation reduces to

$$\mathcal{G}'_0 = -\sum\nolimits_{i<j} \sum\nolimits_{\mu\nu} P_{i\mu} P_{j\nu} J_{i\mu,j\nu} \tag{4.25}$$

Therefore we have

$$\frac{\partial \mathcal{G}'_0}{\partial P_{i\mu}} = -\sum\nolimits_{j\nu}^{j\neq i} P_{j\nu} J_{i\mu,j\nu}. \tag{4.26}$$

Here the gauge condition on J ensures that the single-spin probability of the gauge state does not appear on the *r.h.s.* Note that $J_{i\mu,j\nu}$ for $j < i$ does not enter the model and we let $J_{i\mu,j\nu} = J_{j\nu,i\mu}$ for convenience. Taking the derivative of Eq (4.26), we obtain

$$\frac{\partial^2 \mathcal{G}'_0}{\partial P_{i\mu} \partial P_{j\nu}} = -(1-\delta_{ij}) J_{i\mu,j\nu} \tag{4.27}$$

Substituting Eq (4.23) and the above equation in Eq (4.20) gives

$$(C^{-1})_{i\mu,j\nu} \approx \begin{cases} \frac{\delta_{\mu\nu}}{P_{i\mu}} + \frac{1}{P_{ic_i}} & \text{if } i = j \\ -\alpha J_{i\mu,j\nu} & \text{if } j \neq i \end{cases} \tag{4.28}$$

Finally we combine Eqs (4.20),(4.23) and (4.26) to obtain the self-consistent condition for the local fields

$$h_{i\mu} = \ln \frac{P_{i\mu}}{P_{ic_i}} - \alpha \sum\nolimits_{j\nu}^{j\neq i} P_{j\nu} J_{i\mu,j\nu} + \mathcal{O}(\alpha^2) \tag{4.29}$$

The naive mean-field inversion method is based on Eqs (4.28) and (4.29) which relate the model parameters to the empirically accessible connected correlation matrix.

4.9 Phase transitions in Potts models on homogeneous random graphs

Here we reproduce the mean-field analysis of Potts models (see, e.g., Ref (181, Sec. IC)). Consider a uniform-interaction ferromagnetic q-state Potts model on a graph,

$$\mathcal{H}(\vec{\sigma}) = -\sum_{(ij)\in\mathcal{E}} \delta_{\sigma_i,\sigma_j}, \tag{4.30}$$

where $\delta_{\sigma_i,\sigma_j}$ denotes the Kronecker delta and the summation is over the graph's edges $\mathcal{E}$. In the mean-field approximation, all spins are identical and the internal energy and entropy of the system read

$$U = -|\mathcal{E}| \sum_{\mu=1}^{q} p_\mu^2 \quad \text{and} \quad S = -n \sum_{\mu=1}^{q} p_\mu \ln p_\mu \tag{4.31}$$

where p_μ is the fraction of spins in state μ, n the number of spins and $|\mathcal{E}|$ the numbers of edges (interactions). To analyze the ferromagnetic transition, we consider the ansatz

$$p_\mu = \frac{1}{q}(1-\Delta) + \delta_{\mu,q}\Delta \tag{4.32}$$

where Δ is the order parameter and we chose the state q as the spin state of the ferromagnetic phase. This ansatz yields the free energy per spin

$$\begin{aligned} \beta(f(\Delta) - f(0)) = {} & \frac{1+(q-1)\Delta}{q} \ln(1+(q-1)\Delta) \\ & + \frac{q-1}{q}(1-\Delta)\ln(1-\Delta) - \frac{q-1}{2q}\frac{\lambda}{T}\Delta^2 \end{aligned} \tag{4.33}$$

where $\lambda = 2|\mathcal{E}|/n$ is the mean coordination number. In the thermodynamic limit $n \to \infty$, a phase transition exists at the critical temperature

$$\frac{1}{T_c} = \frac{1}{\lambda} \times \begin{cases} q & \text{if } q \leq 2 \\ 2\frac{q-1}{q-2}\ln(q-1) & \text{if } q > 2 \end{cases} \tag{4.34}$$

The free energy is minimized by $\Delta = 0$ for $T > T_c$ and by the largest root of the equation

$$e^{-\lambda\Delta/T} = \frac{1-\Delta}{1+(q-1)\Delta} \tag{4.35}$$

for $T < T_c$. This phase transition is continuous for $q \leq 2$ and discontinuous for $q > 2$ in which the order parameter and internal energy per spin are discontinuous across the transition,

$$\begin{aligned} \Delta(T_c^-) - \Delta(T_c^+) &= \frac{q-2}{q-1} \\ u(T_c^-) - u(T_c^+) &= -\lambda\frac{(q-2)^2}{2q(q-1)}. \end{aligned} \tag{4.36}$$

Finally we note that the above analysis is exact for complete graphs in which all spins in the system are truly (as opposed to statistically) identical.

CHAPTER 5

ORGANIZATION OF MEMORY IN INFORMATION BOTTLENECK ENHANCED KALMAN FILTERS

This work was done in collaboration with Thierry Mora, Aleksandra Walczak, and Stephanie Palmer.

5.1 Abstract

In order to respond efficiently to changes in the external world, living systems evolve sensory encoding schemes that make effective measurements of the relevant external features in their input stimuli, subject to resource constraints. A typical method for quantifying that trade-off is the information bottleneck (IB) approach. Some sensory encoding schemes have been shown to be optimized for encoding stimulus features that are relevant for prediction in this IB sense. However, typical models do not consider how the organism might make use of a previous stimulus estimate. The Kalman filter (KF) approach makes estimation memory explicit. KF also posits a known model of the input's dynamical system and the measurement model is fixed and is not optimized for efficiency. Connecting these approaches reveals the structure of an optimal encoding that has both a notion of efficient measurement and a memory of the previous stimulus estimate. We derive the encoding that maximally reduces uncertainty in the stimulus estimate, subject to a fixed model 'size', analogous to a power constraint. We demonstrate that by optimizing a sensory model of this type, organisms can achieve larger reductions in uncertainty about their inputs than is possible without memory of their past estimates.

5.2 Introduction

A core challenge for biological systems is estimating the state of the external environment and projecting that forward in time. Examples include predicting sunset in systems that have internal circadian clocks(179; 62), representing the position and velocity of an incoming visual object to drive escape from a predator(125; 15; 174; 143), and sculpting a antibody repertoire in anticipation of future pathogenic attacks (112; 5). All of these important functions involve prediction of future inputs(18; 11). These estimates come at a cost to the organism, as they require resources to develop, maintain, and operate a predictive sensory encoding scheme. Biological systems must make trade-offs between the cost of a measurement and the benefit of knowing the future external state(153).

The classic information bottleneck approach to quantifying this trade-off does not usually include a memory of past estimates(158; 94). The well-know Kalman Filter approach to the same type of prediction problem relies heavily on these past estimates, but does not optimize the encoding to consider just these predictive features(67). Optimizing the encoding subject to that resource constraint can make the measurement model most efficient. Can the combination of these two approaches yield a better, efficient encoder? How useful is estimation memory when measurements are constrained? We set out to answer these questions by merging IB with KF for a few toy problems.

We have previously solved predictive IB problems for these toy models(29; 142), but now include an internal estimate in our IB framework. By incorporating internal estimates into the optimization of the sensory encoder, different features of the stimulus are encoded. This difference emerges from the fact that measurement and internal prediction reduce uncertainty along different input dimensions. The allows for the combination to perform better stimulus estimation than the typical IB framework alone.

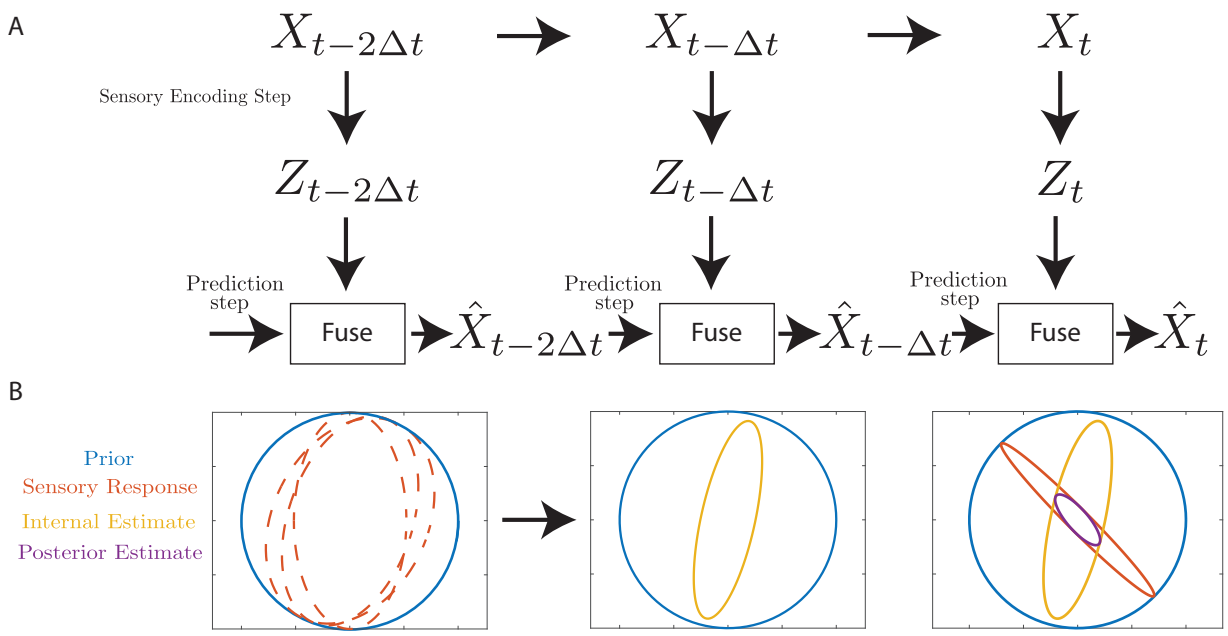

Fig 5.1: Internal estimates based on past sensory responses can be used to improve the state estimates when combined with current sensory responses.(A) Internal estimates can be constructed through prediction based on previous sensory responses and then combining the information provided by each prediction. (B) Leveraging both the current sensory response with the internal estimate enables significant reduction in uncertainty of the posterior estimate of the environment.

5.3 Results

We consider a dynamical system with coordinates, $X_{t-\Delta t}$. A biological system uses an encoder to build a sensory representation, $Z_{t-\Delta t}$. This sensory representation can be used to construct an internal estimate of the external world, $\hat{X}_{t-\Delta t}$. The biological system can then infer further states of the environment either through a prediction step or an instantaneous measurement via the sensory encoder (Fig. 5.1A). Together, these two distinct estimates can be fused to provide a more precise estimate of the external environment(Fig. 5.1B).

For a fixed sensory encoder, a biological system can minimize its uncertainty of the external environment by varying its internal estimate construction based on previous internal estimates and sensory responses. This is accomplished by optimizing the objective function:

$$\max_{p(x_t|\{\hat{x}_{t-\Delta t},z_t\})} I(X_t;\hat{X}_t) \tag{5.1}$$

This is an information theoretic generalization of the Kalman Filter objective function(162; 67; 87).

Biological systems are capable, however, of evolving their sensory encoders to make them more efficient. This is to say that a sensory encoder may be evolved to provide preferentially provide information about one dimension of the input over another for a fixed overall level of power. As such, we propose the optimization of a new objective function:

$$\mathcal{L} = \max_{p(\hat{x}_t|\{\hat{x}_{t-\Delta t},z_t\}),p(z_t|x_t)} I(X_t;\hat{X}_t) - \beta I(X_t;Z_t). \tag{5.2}$$

The optimal sensory encoder will maximize the amount of mutual information between the internal estimate and the external environment for a given cost of encoding, controlled by β.

The optimization of the sensory encoder is reminiscent of the optimization of the representation variable in the IB method. The method proposed here differs, however, in that the internal representation is constructed as a function of previous internal representations and the sensory response, while in the IB method, no additional information is incorporated.

5.3.1 One-dimensional stimulus

We present the results for the objective function for a one-dimensional stimulus. Our stimulus is a Brownian motion process with viscosity parameter γ. In this model, there are two parameters which impact the shape of the sensory encoding: the autocorrelation of the stimulus, $\exp(-\gamma\Delta t)$, and the cost of encoding, β.

We fix the autocorrelation of the stimulus to be $\exp(-\gamma\Delta t) = 0.9$ and the cost to be $\beta = 3.7$ and compute the uncertainty reduction achieved by the sensory response, the

prediction based on prior internal estimates, and the combination of the two (Fig. 5.2A). We see that the sensory representation, at each time step, provides only a small reduction in the uncertainty of the state estimate. However, when the sensory encoding estimate is fused with the forecast estimate, a significant reduction in uncertainty of state estimate can be achieved. This reduction in uncertainty emerges from forecasting based on previous estimates. Forecasting based on previous estimates effectively allows the biological organism to take advantage of the repeated measurement of the system and internal prediction mechanisms.

We vary the encoding cost and identify the optimal encoding strategy for each cost(Fig. 5.2B). Initially, there is a rapid increase in the mutual information between the state estimate and the sensory encoding scheme. However, for low encoding costs, the information about the external world is coming primarily from the instantaneous measurement, as $I(X_t; Z_t) \approx I(X_t; \hat{X}_t)$. Thus, there appears to be encoding cost at which the memory no longer becomes useful.

We then vary both the autocorrelation parameter and the encoding cost simultaneously, and observe three distinct regimes (Fig. 5.2C). In the upper left regime (blue), we have a trivial regime where encoding costs are too high to maintain an encoding scheme. In the bottom regime (yellow), encoding costs are small relative to the autocorrelation of the stimulus. Consequently, memory is unnecessary for estimating the external world. This corresponds to the regime where the internal estimate uncertainty is primarily dependent on the sensory response. Finally, in between these two regimes, there exists some pair of autocorrelations and encoding costs that demonstrate that memory can provide significant enhancements in the precision of the external environment estimate.

The boundaries between each regime can be determined analytically for a one-dimensional stimulus. The derivation is given in the SI. Using this framework, we have identified sensory encoding costs for which maintaining previous internal estimates can enable reduced uncertainty in external environments.

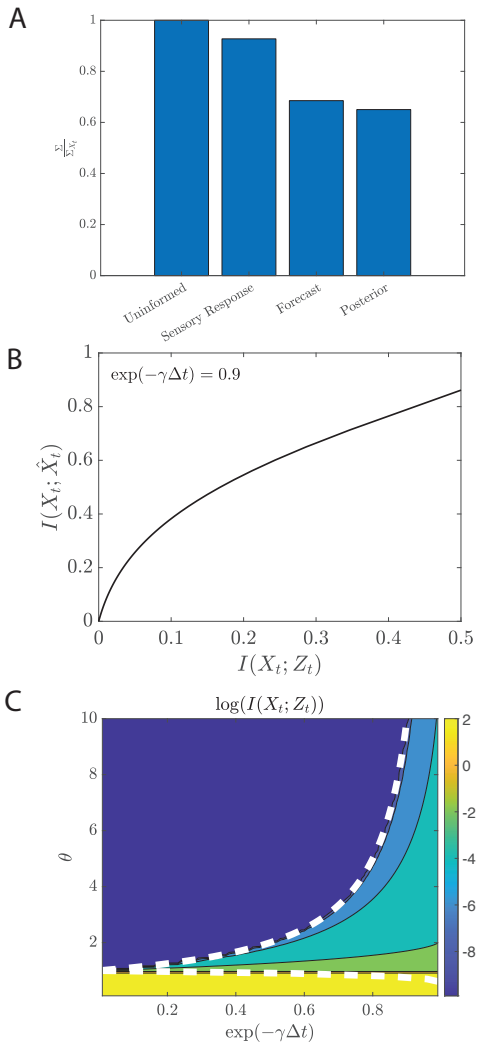

Fig 5.2: Combining internal estimates with current sensory responses provides significant improvements in state estimate uncertainty under certain conditions on encoding costs and environmental correlations. (A) The uninformed uncertainty corresponds to the level of uncertainty in an estimate of the state one could achieve given only an unordered list of velocities. The sensory response uncertainty corresponds to the uncertainty about the state when inferring based only on the sensory response. The forecast uncertainty corresponds to the uncertainty about the state one would have using only previous internal estimates of the state. The posterior, or fused, uncertainty, corresponds to the uncertainty arising when optimally fusing both the sensory response and the forecast estimate. (B) The maximal amount of information the posterior estimate could provide about the state of the system at a given level of sensory response. (C) The phase diagram of sensory response models, visualizing it by plotting $\log(I(X_t; Z_t))$. We observe three phases: a low cost high correlation phase, where $I(X_t; Z_t) \to \infty$; a high cost low correlation phase, where $I(X_t; Z_t) \to 0$; and an intermediate phase, where the sensory response is non trivial. The phase boundaries, drawn as dashed lines, are analytically determined.

5.3.2 Two-dimensional stimulus

We extend our results on one-dimensional stimuli to two-dimensional stimuli by extending our brownian motion model to include a harmonic oscillator force. This results in a first-order Markov process of two input variables. There is now one additional defining parameter, ζ, corresponding to oscillatory motion about an equilibrium $\zeta < 1$ and relaxation to an equilibrium $\zeta \geq 1$(119). Sensory encoding schemes are now optimized by both identifying a dimension along which to reduce uncertainty and the magnitude by which uncertainty should be reduced. This is represented through the use of confidence regions, where the confidence region outlined in blue corresponds to the prior and the confidence region corresponding to the sensory encoder is in orange (Fig. 5.3A). Due to the numerical intractability of the optimization of the objective function, the optimal encoding scheme is computed by scanning over the dimension along which uncertainty is reduced and the magnitude by which the uncertainty is reduced (Fig. 5.3B).

We visualize the optimal sensory encoder (Fig. 5.3C(left)) and compare to a suboptimal sensory encoder (Fig. 5.3C(right)). The uncertainty of the sensory encoder is given in the orange ellipse, the uncertainty of the forecase estimate is given in the yellow ellipse, and the posterior estimate is given as the purple ellipse. Though the optimal sensory encoder and suboptimal sensory encoder presented have equal areas, the difference in the dimension along which they reduce uncertainty impacts the forecast based on previous internal estimates. In particular, the forecast based on previous internal estimates for the optimal sensory encoder provides uncertainty reduction along a dimension distinct from the sensory encoder. Meanwhile, in the suboptimal case, the internal estimate is redundant with the sensory encoder, and thus, provides no additional information about the future state of the environment. We formalize this by calculating the synergy between the sensory encoder and the prediction based on the internal estimate (Fig. 5.3D). We find that optimal dimension for the sensory encoder is the one that maximizes this synergy.

We now explore the structure of the optimal sensory encoder for different motion statistics and values of the encoding cost. The differing motion statistics are given by varying ζ and Δt (Fig. 5.4A). We see that the underlying motion statistics do not impact the types of encoders as much as the timescale and sensory encoding costs. We also consider the impact of reducing uncertainty along two-dimensions in the sensory encoder (Fig. 5.4B). We find that encoding both dimensions of the stimulus offers no benefit in reducing the uncertainty of the estimate of the external environment. This is due to the fact that forecast based on previous internal estimates already provides information about a dimension distinct from the sensory encoder.

Finally, we present the results of optimizing the objective function using a two-dimensional and a one-dimensional encoder for different values of β(Fig. 5.4C). We find that though the two-dimensional encoder can reduce uncertainty along both dimensions, it is degenerate with the one-dimensional encoder up until a memorization phase, where the previous internal estimates do not provide additional information about the external environment.

5.4 Discussion

In our work, we have connected the Kalman filter to models of sensory information processing. We demonstrate the impact of a neural Kalman filter on the structure of an efficient sensory encoder, extending on previous works that have demonstrated how a neural system could implement a Kalman filter(102). It also provides predictions on the structure of efficient representations neural networks may discover(93).

demonstrated the impact of Kalman filtering on the . This extends on explored a novel framework that connects Kalman filter approaches to integrating data over a long timescale to improve state estimates with information theoretic measure-based constraints to identify the relevant dimensions of the data to measure. This presents novel insight over previous work, in which observations from previous time points' impact to the optimization of the sensory model were not considered. This may be relevant for behaviors whose timescales are

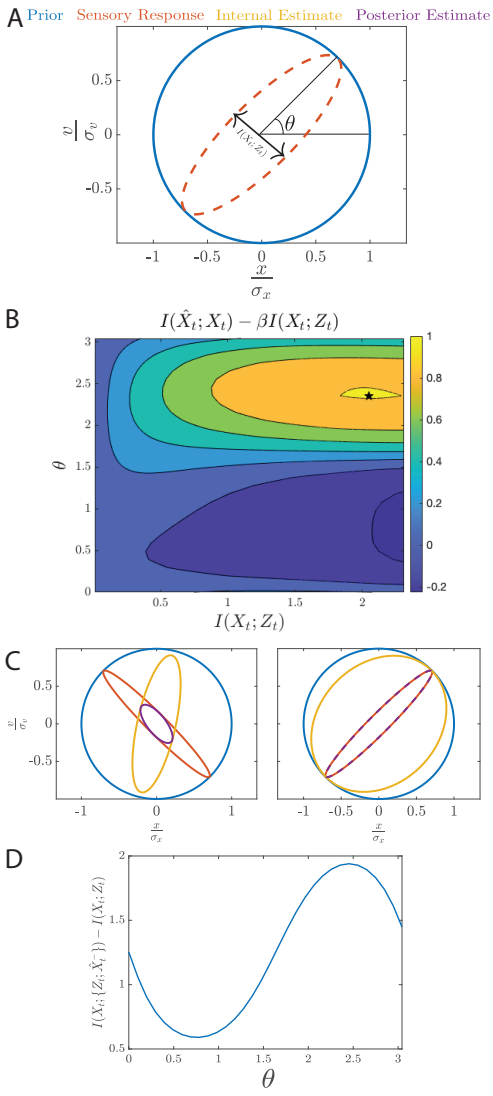

Fig 5.3: The results for the optimal sensory encoding scheme for a two-dimesional stimuli show that the preferred encoding dimension is the dimension which provides the largest marginal benefit over using just the previous internal estimate.(A) To determine the optimal sensory response organ, we sweep through a range of dimensions along which we reduce uncertainty and vary the amount by which we reduce the uncertainty. We will analyze this for $\zeta = 1$, $\beta = 1.1$, $\Delta t = 1$. (B) By sweeping through this range of parameters, we are capable of identifying which sensory organ optimizes the objective function. We present one example, and denote the maximal point with a star. (C) A comparison of the impact of the sensory encoding angle. The sensory encoding scheme with the optimal angle (yellow) is to the left, while the suboptimal encoding scheme is plotted to the right. The comparison suggests that the optimal sensory encoding scheme will be minimally redundant with the internal estimate uncertainty (orange). The suboptimal encoding encodes $\pi/2$ radians off the correct angle. The suboptimal encoding encodes in a (D) A cross section of the objective function along the θ axis. The objective function maximizes where $I(X_t; \{Z_t; \hat{X}_{t-\Delta t}) - I(X_t; Z_t)$ is maximized.

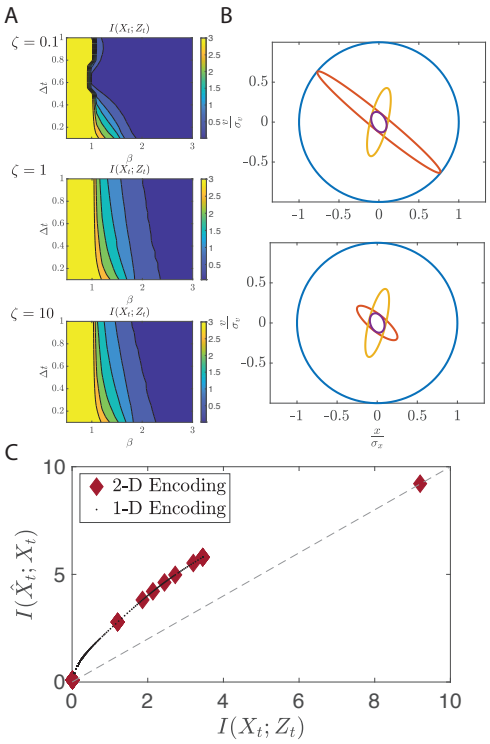

Fig 5.4: Encoding two feature dimensions provides no additional benefit in the posterior estimate uncertainty.(A) The magnitude of the optimal sensory encoding scheme, $I(X_t; Z_t)$ for different values of the tradeoff parameter for different underlying motion models. (B) A comparison of one-dimensional sensory encoding schemes (top) to two-dimensional encoding schemes (bottom). Here, $\zeta = 0.1$, $\beta = 1.1$, $\Delta t = 0.1$. Two-dimensional encoding schemes do not provide significant improvements in the poster estimate uncertainty. (C) Two-dimensional encodings can, at best, match one-dimensional encodings until the sensory estimate is a memorization of the input statistic.

faster than that of neural computations, such as reflexive reaction to a predator hunting prey, while it may not be relevant for longer timescale behavior. For example, updating internal estimates from one sensory modality with information from another sensory modality has been shown to improve the efficiency of olfactory searches in computational models(136).

We demonstrate the solution for this problem for both one and two-dimensional Gaussian stimuli, and note two key principles that emerge. First, there emerges a regime where memories can be used to improve state estimates. This regime depends on the encoding cost, relative to the cost of keeping an internal estimate based on previous sensory responses. Second, the sensory encoding scheme discovered differs from the encoding scheme predicted by the information bottleneck method. This is because by leveraging previous internal estimates, the organism is capable of computing features that might otherwise need to be measured. For example, the organism no longer precisely needs to measure velocity because the organism is capable of discovering velocity through its previous estimates of position.

In general, optimizing the objective function for this problem is analytically intractable due to the update structure of the prior and posterior estimate. We make a key simplification by considering the model to be in the stationary regime; that is, the regime where the stimulus had been observed long enough that the posterior and prior estimate reach steady state. However, this may not be the case for stimuli whose statistics may be changing. Novel behavior may emerge if the underlying statistics themselves change. There may regimes of changes where only the gain needs to be changed to make significant improvements in state inference, and others where the observation model itself needs to be changed.

We have not considered cases where the external dynamical model also needs to be learned. However, learning the dynamical systems model is also a critical element of this

state estimation framework. While we have not explored this in this work, learning the dynamical systems model could be included through the implementation of an Ensemble Kalman filer(50). By imposing constraints on the learned dynamical systems model, this extension could be used to make predictions on how precisely an external dynamical systems model needs to be learned for "good enough" state inference. This could be relevant for circadian rhythms, where intrinsic noise is always present, even though the dynamical model governing the external system is fully deterministic. Such a framework could also be used to explore the condition in which a learned model can be transferred to dynamical systems with varying underlying statistics.

CHAPTER 6

DISCUSSION

In this work, we demonstrated that biological systems adapt to changing environments by exploiting the temporal correlation structure of the environment, enabling them to predict and avoid future threats. We demonstrated this in both molecular and neural settings. In molecular settings, we are able to demonstrate that depending upon the rate of environmental variation, biological systems can be driven towards prediction, even when prediction is not the best strategy at any given time. In neural settings, we demonstrate how information about the external environment should be encoded in order to enable prediction.

Excitingly, these short timescale prediction problems are solved by the long timescale evolutionary process. On top of that, the long timescale evolutionary process is responsible for evolving the nature of evolution itself. This suggests that by evolving an evolutionary process, short timescale prediction becomes an emergent phenomenon. Future work will explore how mutation rates themselves evolve, and how prediction emerges from the extant variation generated by those mutations.

The evolution of predictive strategies depends on a biological system's ability to navigate tradeoffs. In molecular settings, these tradeoffs emerge from the entropic cost of discovering a predictive strategy and the benefit of being well fit to future environments with little adaptation. This tradeoff has molecular motivations, as in binding the conserved residues of HIV obligates binding an embedded residue on the HIV envelop protein. This is inherently less energetically favorable than binding an exposed residue.

In the neural setting, the tradeoff emerges between the cost of encoding information about the external environment and the need to be predictive. The particular optimization achieved by a predictive strategy then depends on the type of prediction problem. We observe that for visual scenes, which can vary rapidly, the optimal encoder is very informative of future states and continuous representations are preferable. However, when attempting to build predictive

strategies for dynamics with discrete outcomes, rather than attempt to be predictive, it becomes preferable to discretely tile the possible outcomes.

Studying adaptation to changing environments offers key insights into the way real biological systems may have evolved, as biological systems must respond to changes in the external environment in order to survive. While in this work, we considered adaptive strategies that enable explicit responses to external environments, such as sensing the environment or generalizing against families of environments, there are many strategies that provide implicit responses. Such implicit responses include rapidly changing phenotype in response to a new environmental threat. Many molecular mechanisms can give rise to such rapidly adaptable phenotypes, such as chaperone proteins that buffer mutations(**?**), evolving mutation rates(**?**), and sensory encoding schemes.

None of these rapidly adaptable phenotypes confer immediate fitness benefit, and some may even come at a cost, but can offer fitness benefits if the environment shifts. Hence, these phenotypes should be described in terms of their second-order impact. Second-order impact means how a given phenotype impacts the reproductive viability of the lineage of the individuals carrying the phenotype, as opposed to first-order impact, which reflects how a phenotype impacts the reproductive viability of the individual themselves. Such an impact cannot be straightforwardly explained by the Wright-Fisher equation, as $\frac{df}{dt}$, the change in frequency of an allele over time, is not impacted by the future of the allele. Instead it could be explained by the structure of the second derivative of the change in frequency of allele in time. By changing $\frac{d^2f}{dt^2}$, however, the rate of evolution itself can be changed. By exploring how the fundamental forces of evolution impact $\frac{d^2f}{dt^2}$, one can identify what kinds of selective and mutational forces give rise to phenotypes which are rapidly adaptable.

The impact of mutation rates on adaptation can be explored experimentally. For example, by competing populations with distinct mutation rates in families of selective pressures, one can identify under what conditions higher mutation rates are preferable. Such experiments

are exciting because they offer empirical evidence for the strength and sign of the epistasis between first- and second-order selective impacts. Depending on the structure of epistasis, it may be evolvable traits have no clear notion of fitness. Instead, their viability depends on the background in which they are evolved. Such a dramatic outcome is not unique to mutation rates. There are many other traits, such as recombination rate or cryptic variation, which may give rise to the same effect.

The impact of extinction rates on the rate of adaptation can be explored mathematically. The rate of evolution can be given by identifying the rate at which a new mutant can fix in a population. While the probability of fixation of a mutation has been given by Kimura et. al.(73), studies on what parameters govern the rate of fixation of a mutation are less prevalent. By identifying how the rate of mutation is impacted by extinction rates, one can understand why, for example, B-cell evolution is poised near extinction. A simple theoretical argument would suggest that extinction rates need to be higher to promote evolvability, as being nearer to extinction offers larger opportunities for mutants to fix.

Finally, exploring the landscape of viable mutations themselves may yield insight into how rapidly adaptation can occur. Biological systems are faced with a serious constraint during evolution - that is, they can never fix a phenotype that results in extinction. Consequently, evolving systems may sometimes traverse longer mutational pathways than necessary, in order to avoid strongly deleterious pathways. As such, continued work on characterizing the topology of the mutational landscape of a given trait could explain under what conditions being evolvable is even possible.

Altogether, the work presented here presents some preliminary steps in studying evolution in a changing environment. This allows for the relaxation of strong assumptions made on previous studies of evolution, and. clear identification for when such assumptions are valid. Violating these assumptions yields novel behaviors, some of which are explored here.

REFERENCES

[1] Rhys M Adams, Justin B Kinney, Aleksandra M Walczak, and Thierry Mora. Epistasis in a fitness landscape defined by Antibody-Antigen binding free energy. *Cell Syst*, 8(1): 86–93.e3, January 2019. ISSN 2405-4712. doi: 10.1016/j.cels.2018.12.004.

[2] Laurence Aitchison, Nicola Corradi, and Peter E. Latham. Zipf's law arises naturally when there are underlying, unobserved variables. *PLOS Comput. Biol.*, 12(12):e1005110, 12 2016. doi: 10.1371/journal.pcbi.1005110.

[3] Alexander A. Alemi. Variational predictive information bottleneck, 2019.

[4] Alexander A. Alemi, Ian Fischer, Joshua V. Dillon, and Kevin Murphy. Deep variational information bottleneck, 2016.

[5] Christopher D.C. Allen, Takaharu Okada, and Jason G. Cyster. Germinal-center organization and cellular dynamics. *Immunity*, 27(2):190–202, 2007. ISSN 1074-7613. doi: https://doi.org/10.1016/j.immuni.2007.07.009. URL `https://www.sciencedirect.com/science/article/pii/S1074761307003706`.

[6] Daniel Amit, Hanoch Gutfreund, and H Sompolinsky. Storing infinite numbers of patterns in a spin-glass model of neural networks. *Phys. Rev. Lett.*, 55(14):1530–1533, September 1985. ISSN 0031-9007.

[7] S. Arimoto. An algorithm for computing the capacity of arbitrary discrete memoryless channels. *IEEE Transactions on Information Theory*, 18(1):14–20, January 1972. doi: 10.1109/TIT.1972.1054753.

[8] Peter F Arndt and Terence Hwa. Regional and time-resolved mutation patterns of the human genome. *Bioinformatics*, 20(10):1482–1485, July 2004. ISSN 1367-4803. doi: 10.1093/bioinformatics/bth105.

[9] Joseph J. Atick and A. Norman Redlich. Towards a Theory of Early Visual Processing. *Neural Computation*, 2(3):308–320, 09 1990. ISSN 0899-7667. doi: 10.1162/neco.1990.2.3.308. URL `https://doi.org/10.1162/neco.1990.2.3.308`.

[10] Alessandro Attanasi, Andrea Cavagna, Lorenzo Del Castello, Irene Giardina, Stefania Melillo, Leonardo Parisi, Oliver Pohl, Bruno Rossaro, Edward Shen, Edmondo Silvestri, and Massimiliano Viale. Finite-size scaling as a way to probe near-criticality in natural swarms. *Phys. Rev. Lett.*, 113:238102, Dec 2014. doi: 10.1103/PhysRevLett.113.238102.

[11] Horace B. Barlow. Possible principles underlying the transformation of sensory messages. In *Sensory communication*. MIT Press, 2012.

[12] Pierre Barrat-Charlaix, Anna Paola Muntoni, Kai Shimagaki, Martin Weigt, and Francesco Zamponi. Sparse generative modeling via parameter reduction of boltzmann machines: Application to protein-sequence families. *Phys. Rev. E*, 104:024407, Aug 2021. doi: 10.1103/PhysRevE.104.024407.

[13] Normand J. Beaudry and Renato Renner. An intuitive proof of the data processing inequality, 2011.

[14] H.C. Berg and E.M. Purcell. Physics of chemoreception. *Biophysical Journal*, 20(2): 193 – 219, 1977. ISSN 0006-3495. doi: https://doi.org/10.1016/S0006-3495(77)85544-6. URL http://www.sciencedirect.com/science/article/pii/S0006349577855446.

[15] Michael J Berry and Gregory Schwartz. The retina as embodying predictions about the visual world. *Predictions in the brain: Using our past to generate a future*, page 295, 2011.

[16] W. Bialek. *Biophysics: Searching for Principles*. Princeton University Press, 2012. ISBN 9780691138916. URL https://books.google.com/books?id=5In_FKA2rmUC.

[17] William Bialek. Perspectives on theory at the interface of physics and biology. *Rep. Prog. Phys.*, 81(1):012601, dec 2017. doi: 10.1088/1361-6633/aa995b.

[18] William Bialek, Ilya Nemenman, and Naftali Tishby. Predictability, complexity, and learning. *Neural Computation*, 13(11):2409–2463, 2001. doi: 10.1162/089976601753195969. URL https://doi.org/10.1162/089976601753195969.

[19] William Bialek, Andrea Cavagna, Irene Giardina, Thierry Mora, Edmondo Silvestri, Massimiliano Viale, and Aleksandra M. Walczak. Statistical mechanics for natural flocks of birds. *Proc. Natl. Acad. Sci. U.S.A.*, 109(13):4786–4791, 2012. doi: 10.1073/pnas.1118633109.

[20] William Bialek, Andrea Cavagna, Irene Giardina, Thierry Mora, Oliver Pohl, Edmondo Silvestri, Massimiliano Viale, and Aleksandra M. Walczak. Social interactions dominate speed control in poising natural flocks near criticality. *Proc. Natl. Acad. Sci. U.S.A.*, 111(20):7212–7217, 2014. doi: 10.1073/pnas.1324045111.

[21] William Bialek, Stephanie E. Palmer, and David J. Schwab. What makes it possible to learn probability distributions in the natural world?, 2020. URL https://arxiv.org/abs/2008.12279.

[22] Vincent A Billock, Gonzalo C de Guzman, and J.A Scott Kelso. Fractal time and 1/f spectra in dynamic images and human vision. *Physica D: Nonlinear Phenomena*, 148(1): 136 – 146, 2001. ISSN 0167-2789. doi: https://doi.org/10.1016/S0167-2789(00)00174-3. URL http://www.sciencedirect.com/science/article/pii/S0167278900001743.

[23] Richard E. Blahut. Computation of channel capacity and rate-distortion functions. *IEEE Trans. Inform. Theory*, 18:460–473, 1972.

[24] Celia Blanco, Evan Janzen, Abe Pressman, Ranajay Saha, and Irene A Chen. Molecular fitness landscapes from High-Coverage sequence profiling. *Annu. Rev. Biophys.*, 48:1–18, May 2019. ISSN 1936-122X, 1936-1238. doi: 10.1146/annurev-biophys-052118-115333.

[25] Mattia Bonsignori, Tongqing Zhou, Zizhang Sheng, Lei Chen, Feng Gao, M Gordon Joyce, Gabriel Ozorowski, Gwo-Yu Chuang, Chaim A Schramm, Kevin Wiehe, S Munir Alam, Todd Bradley, Morgan A Gladden, Kwan-Ki Hwang, Sheelah Iyengar, Amit Kumar, Xiaozhi Lu, Kan Luo, Michael C Mangiapani, Robert J Parks, Hongshuo Song, Priyamvada Acharya, Robert T Bailer, Allen Cao, Aliaksandr Druz, Ivelin S Georgiev, Young D Kwon, Mark K Louder, Baoshan Zhang, Anqi Zheng, Brenna J Hill, Rui Kong, Cinque Soto, NISC Comparative Sequencing Program, James C Mullikin, Daniel C Douek, David C Montefiori, Michael A Moody, George M Shaw, Beatrice H Hahn, Garnett Kelsoe, Peter T Hraber, Bette T Korber, Scott D Boyd, Andrew Z Fire, Thomas B Kepler, Lawrence Shapiro, Andrew B Ward, John R Mascola, Hua-Xin Liao, Peter D Kwong, and Barton F Haynes. Maturation pathway from germline to broad HIV-1 neutralizer of a CD4-Mimic antibody. *Cell*, 165(2):449–463, April 2016. ISSN 0092-8674, 1097-4172. doi: 10.1016/j.cell.2016.02.022.

[26] Dorje Brody and Nicolas Rivier. Geometrical aspects of statistical mechanics. *Phys. Rev. E*, 51:1006–1011, Feb 1995. doi: 10.1103/PhysRevE.51.1006.

[27] Dennis R Burton, Ronald C Desrosiers, Robert W Doms, Wayne C Koff, Peter D Kwong, John P Moore, Gary J Nabel, Joseph Sodroski, Ian A Wilson, and Richard T Wyatt. HIV vaccine design and the neutralizing antibody problem. *Nat. Immunol.*, 5 (3):233–236, March 2004. ISSN 1529-2908. doi: 10.1038/ni0304-233.

[28] Dennis R Burton, Pascal Poignard, Robyn L Stanfield, and Ian A Wilson. Broadly neutralizing antibodies present new prospects to counter highly antigenically diverse viruses. *Science*, 337(6091):183–186, July 2012. ISSN 0036-8075, 1095-9203. doi: 10.1126/science.1225416.

[29] Matthew Chalk, Olivier Marre, and Gašper Tkačik. Toward a unified theory of efficient, predictive, and sparse coding. *Proceedings of the National Academy of Sciences*, 115(1):186–191, 2018. ISSN 0027-8424. doi: 10.1073/pnas.1711114115. URL `https://www.pnas.org/content/115/1/186`.

[30] Sidhartha Chaudhury, Jaques Reifman, and Anders Wallqvist. Simulation of B cell affinity maturation explains enhanced antibody cross-reactivity induced by the polyvalent malaria vaccine AMA1. *J. Immunol.*, 193(5):2073–2086, September 2014. ISSN 0022-1767, 1550-6606. doi: 10.4049/jimmunol.1401054.

[31] Gal Chechik, Amir Globerson, Naftali Tishby, and Yair Weiss. Information bottleneck for gaussian variables. In S. Thrun, L. K. Saul, and B. Schölkopf, editors, *Advances in Neural Information Processing Systems 16*, pages 1213–1220. MIT Press, 2004. URL `http://papers.nips.cc/paper/2457-information-bottleneck-for-gaussian-variables.pdf`.

[32] Xiaowen Chen, Francesco Randi, Andrew M. Leifer, and William Bialek. Searching for collective behavior in a small brain. *Phys. Rev. E*, 99:052418, May 2019. doi: 10.1103/PhysRevE.99.052418.

[33] Dante R. Chialvo. Emergent complex neural dynamics. *Nat. Phys.*, 6(10):744–750, 2010. doi: 10.1038/nphys1803.

[34] Childs Lauren M., Baskerville Edward B., and Cobey Sarah. Trade-offs in antibody repertoires to complex antigens. *Philos. Trans. R. Soc. Lond. B Biol. Sci.*, 370(1676): 20140245, September 2015. ISSN 0962-8436. doi: 10.1098/rstb.2014.0245.

[35] Cobey Sarah, Wilson Patrick, and Matsen Frederick A. The evolution within us. *Philos. Trans. R. Soc. Lond. B Biol. Sci.*, 370(1676):20140235, September 2015. ISSN 0962-8436. doi: 10.1098/rstb.2014.0235.

[36] Simona Cocco, Christoph Feinauer, Matteo Figliuzzi, Rémi Monasson, and Martin Weigt. Inverse statistical physics of protein sequences: a key issues review. *Rep. Prog. Phys.*, 81(3):032601, jan 2018. doi: 10.1088/1361-6633/aa9965.

[37] Antonio C. Costa, Tosif Ahamed, and Greg J. Stephens. Adaptive, locally linear models of complex dynamics. *Proc. Natl. Acad. Sci. U.S.A.*, 116(5):1501–1510, 2019. doi: 10.1073/pnas.1813476116.

[38] Gavin E. Crooks. Measuring thermodynamic length. *Phys. Rev. Lett.*, 99:100602, Sep 2007. doi: 10.1103/PhysRevLett.99.100602.

[39] Ivana Cvijović, Benjamin H Good, Elizabeth R Jerison, and Michael M Desai. Fate of a mutation in a fluctuating environment. *Proc. Natl. Acad. Sci. U. S. A.*, 112(36): E5021–8, September 2015. ISSN 0027-8424, 1091-6490. doi: 10.1073/pnas.1505406112.

[40] Ivana Cvijovic, Benjamin H. Good, Elizabeth R. Jerison, and Michael M. Desai. Fate of a mutation in a fluctuating environment. *Proceedings of the National Academy of Sciences*, 112(36):E5021–E5028, 2015. doi: 10.1073/pnas.1505406112. URL `https://www.pnas.org/doi/abs/10.1073/pnas.1505406112`.

[41] A Das and I R Fiete. Systematic errors in connectivity inferred from activity in strongly recurrent networks. *Nat. Neurosci.*, 23:1286–1296, 2020. doi: 10.1038/s41593-020-0699-2.

[42] Maxwell G De Jong and Kevin B Wood. Tuning spatial profiles of selection pressure to modulate the evolution of drug resistance. *Phys. Rev. Lett.*, 120(23):238102, June 2018. ISSN 0031-9007, 1079-7114. doi: 10.1103/PhysRevLett.120.238102.

[43] Eleonora De Leonardis, Benjamin Lutz, Sebastian Ratz, Simona Cocco, Rémi Monasson, Alexander Schug, and Martin Weigt. Direct-coupling analysis of nucleotide coevolution facilitates RNA secondary and tertiary structure prediction. *Nucleic Acids Res.*, 43 (21):10444–10455, 09 2015. doi: 10.1093/nar/gkv932.

[44] R. R. de Ruyter van Steveninck and S. B. Laughlin. The rate of information transfer at graded-potential synapses. *Nature*, 379(6566):642–645, 1996. doi: 10.1038/379642a0. URL `https://doi.org/10.1038/379642a0`.

[45] Michael W Deem and Ha Youn Lee. Sequence space localization in the immune system response to vaccination and disease. *Phys. Rev. Lett.*, 91(6):068101, August 2003. ISSN 0031-9007. doi: 10.1103/PhysRevLett.91.068101.

[46] Jonathan Desponds, Thierry Mora, and Aleksandra M Walczak. Fluctuating fitness shapes the clone-size distribution of immune repertoires. *Proc. Natl. Acad. Sci. U. S. A.*, 113(2):274–279, January 2016. ISSN 0027-8424, 1091-6490. doi: 10.1073/pnas.1512977112.

[47] Patrick T. Dolan, Zachary J. Whitfield, and Raul Andino. Mapping the evolutionary potential of rna viruses. *Cell Host & Microbe*, 23(4):435 – 446, 2018. ISSN 1931-3128. doi: https://doi.org/10.1016/j.chom.2018.03.012. URL `http://www.sciencedirect.com/science/article/pii/S1931312818301410`.

[48] S. N. Dorogovtsev, A. V. Goltsev, and J. F. F. Mendes. Ising model on networks with an arbitrary distribution of connections. *Phys. Rev. E*, 66:016104, Jul 2002. doi: 10.1103/PhysRevE.66.016104.

[49] Mathias Dunkel, Ulrike Schmidt, Swantje Struck, Lena Berger, Bjoern Gruening, Julia Hossbach, Ines S. Jaeger, Uta Effmert, Birgit Piechulla, Roger Eriksson, Jette Knudsen, and Robert Preissner. SuperScent?a database of flavors and scents. *Nucleic Acids Research*, 37(suppl_1):D291–D294, 10 2008. ISSN 0305-1048. doi: 10.1093/nar/gkn695. URL `https://doi.org/10.1093/nar/gkn695`.

[50] Geir Evensen. The ensemble kalman filter: Theoretical formulation and practical implementation. *Ocean dynamics*, 53(4):343–367, 2003.

[51] Nir Friedman. Inferring cellular networks using probabilistic graphical models. *Science*, 303(5659):799–805, 2004. doi: 10.1126/science.1094068.

[52] Nir Friedman, Ori Mosenzon, Noam Slonim, and Naftali Tishby. Multivariate information bottleneck, 2013. URL `https://arxiv.org/abs/1301.2270`.

[53] Feng Gao, Mattia Bonsignori, Hua-Xin Liao, Amit Kumar, Shi-Mao Xia, Xiaozhi Lu, Fangping Cai, Kwan-Ki Hwang, Hongshuo Song, Tongqing Zhou, Rebecca M Lynch, S Munir Alam, M Anthony Moody, Guido Ferrari, Mark Berrong, Garnett Kelsoe, George M Shaw, Beatrice H Hahn, David C Montefiori, Gift Kamanga, Myron S Cohen, Peter Hraber, Peter D Kwong, Bette T Korber, John R Mascola, Thomas B Kepler, and Barton F Haynes. Cooperation of B cell lineages in induction of HIV-1-broadly neutralizing antibodies. *Cell*, 158(3):481–491, July 2014. ISSN 0092-8674, 1097-4172. doi: 10.1016/j.cell.2014.06.022.

[54] C. W. Gardiner. *Handbook of stochastic methods for physics, chemistry and the natural sciences*, volume 13 of *Springer Series in Synergetics*. Springer-Verlag, Berlin, third edition, 2004. ISBN 3-540-20882-8.

[55] Elizabeth Gardner. Maximum storage capacity in neural networks. *EPL*, 4(4):481, 1987.

[56] Robert A Gatenby, Ariosto S Silva, Robert J Gillies, and B Roy Frieden. Adaptive therapy. *Cancer Res.*, 69(11):4894–4903, June 2009. ISSN 0008-5472, 1538-7445. doi: 10.1158/0008-5472.CAN-08-3658.

[57] A Georges and J S Yedidia. How to expand around mean-field theory using high-temperature expansions. *J. Phys. A*, 24(9):2173–2192, 1991. doi: 10.1088/0305-4470/24/9/024.

[58] Nigel Goldenfeld and Carl Woese. Life is physics: Evolution as a collective phenomenon far from equilibrium. *Annu. Rev. Condens. Matter Phys.*, February 2011. doi: 10.1146/annurev-conmatphys-062910-140509.

[59] Najeeb Halabi, Olivier Rivoire, Stanislas Leibler, and Rama Ranganathan. Protein sectors: Evolutionary units of three-dimensional structure. *Cell*, 138(4):774–786, 2009. doi: 10.1016/j.cell.2009.07.038.

[60] Mathieu Hemery and Olivier Rivoire. Evolution of sparsity and modularity in a model of protein allostery. *Phys. Rev. E Stat. Nonlin. Soft Matter Phys.*, 91(4):042704, April 2015. ISSN 1539-3755, 1550-2376. doi: 10.1103/PhysRevE.91.042704.

[61] J J Hopfield. Neural networks and physical systems with emergent collective computational abilities. In *Proceedings of the International Association for Shell and Spatial Structures (IASS) Symposium 2009*, January 1982.

[62] Kabir Husain, Weerapat Pittayakanchit, Gopal Pattanayak, Michael J. Rust, and Arvind Murugan. Kalman-like self-tuned sensitivity in biophysical sensing. *Cell Systems*, 9(5):459–465.e6, 2019. ISSN 2405-4712. doi: https://doi.org/10.1016/j.cels.2019.08.008. URL `https://www.sciencedirect.com/science/article/pii/S2405471219303060`.

[63] Kabir Husain, Weerapat Pittayakanchit, Gopal Pattanayak, Michael J. Rust, and Arvind Murugan. Kalman-like self-tuned sensitivity in biophysical sensing. *Cell Systems*, 9(5):459 – 465.e6, 2019. ISSN 2405-4712. doi: https://doi.org/10.1016/j.cels.2019.08.008. URL `http://www.sciencedirect.com/science/article/pii/S2405471219303060`.

[64] Kavita Jain and Joachim Krug. Deterministic and stochastic regimes of asexual evolution on rugged fitness landscapes. *Genetics*, 175(3):1275–1288, March 2007. ISSN 0016-6731. doi: 10.1534/genetics.106.067165.

[65] W. Janke, D.A. Johnston, and R. Kenna. Information geometry and phase transitions. *Physica A*, 336(1):181–186, 2004. doi: 10.1016/j.physa.2004.01.023. Proceedings of the XVIII Max Born Symposium "Statistical Physics outside Physics".

[66] Jae-Hyung Jeon and Ralf Metzler. Fractional brownian motion and motion governed by the fractional langevin equation in confined geometries. *Phys. Rev. E*, 81:021103,

Feb 2010. doi: 10.1103/PhysRevE.81.021103. URL https://link.aps.org/doi/10.1103/PhysRevE.81.021103.

[67] Rudolph Emil Kalman. A new approach to linear filtering and prediction problems. *Transactions of the ASME–Journal of Basic Engineering*, 82(Series D):35–45, 1960.

[68] H. J. Kappen and F. B. Rodríguez. Efficient learning in Boltzmann machines using linear response theory. *Neural Comput.*, 10(5):1137–1156, 1998. doi: 10.1162/089976698300017386.

[69] Nadav Kashtan and Uri Alon. Spontaneous evolution of modularity and network motifs. *Proc. Natl. Acad. Sci. U. S. A.*, 102(39):13773–13778, September 2005. ISSN 0027-8424. doi: 10.1073/pnas.0503610102.

[70] R Kassen. The experimental evolution of specialists, generalists, and the maintenance of diversity: Experimental evolution in variable environments. *J. Evol. Biol.*, 15(2):173–190, March 2002. ISSN 1010-061X, 1420-9101. doi: 10.1046/j.1420-9101.2002.00377.x.

[71] Allen A Katouli and Natalia L Komarova. The worst drug rule revisited: mathematical modeling of cyclic cancer treatments. *Bull. Math. Biol.*, 73(3):549–584, March 2011. ISSN 0092-8240, 1522-9602. doi: 10.1007/s11538-010-9539-y.

[72] Motoo Kimura. Diffusion models in population genetics. *Journal of Applied Probability*, 1(2):177–232, 1964. ISSN 00219002. URL http://www.jstor.org/stable/3211856.

[73] Motoo Kimura. Theoretical foundation of population genetics at the molecular level. *Theoretical Population Biology*, 2(2):174–208, 1971. ISSN 0040-5809. doi: https://doi.org/10.1016/0040-5809(71)90014-1. URL https://www.sciencedirect.com/science/article/pii/0040580971900141.

[74] Yaakov Kleeorin, William P. Russ, Olivier Rivoire, and Rama Ranganathan. Undersampling and the inference of coevolution in proteins, 2021. URL https://www.biorxiv.org/content/10.1101/2021.04.22.441025v1.

[75] Tetsuya J. Kobayashi and Yuki Sughiyama. Stochastic and information-thermodynamic structures of population dynamics in a fluctuating environment. *Phys. Rev. E*, 96:012402, Jul 2017. doi: 10.1103/PhysRevE.96.012402. URL https://link.aps.org/doi/10.1103/PhysRevE.96.012402.

[76] Artemy Kolchinsky, Brendan D Tracey, and David H Wolpert. Nonlinear information bottleneck. *Entropy*, 21(12):1181, 2019.

[77] Dmitry Krotov, Julien O. Dubuis, Thomas Gregor, and William Bialek. Morphogenesis at criticality. *Proc. Natl. Acad. Sci. U.S.A.*, 111(10):3683–3688, 2014. doi: 10.1073/pnas.1324186111.

[78] E Kussell and M Vucelja. Non-equilibrium physics and evolution–adaptation, extinction, and ecology: a key issues review. *Rep. Prog. Phys.*, 77(10):102602, October 2014. ISSN 1361-6633, 0034-4885. doi: 10.1088/0034-4885/77/10/102602.

[79] Edo Kussell, Stanislas Leibler, and Alexander Grosberg. Polymer-population mapping and localization in the space of phenotypes. *Phys. Rev. Lett.*, 97(6):068101, August 2006. ISSN 0031-9007. doi: 10.1103/PhysRevLett.97.068101.

[80] Simon B. Laughlin. A simple coding procedure enhances a neuron's information capacity. *Zeitschrift für Naturforschung C*, 36:910 – 912, 1981.

[81] Tai Sing Lee and David Mumford. Hierarchical bayesian inference in the visual cortex. *J. Opt. Soc. Am. A*, 20(7):1434–1448, Jul 2003. doi: 10.1364/JOSAA.20.001434. URL `http://josaa.osa.org/abstract.cfm?URI=josaa-20-7-1434`.

[82] A. Levina, J. M. Herrmann, and T. Geisel. Dynamical synapses causing self-organized criticality in neural networks. *Nat. Phys.*, 3(12):857–860, 2007. doi: 10.1038/nphys758.

[83] Richard Levins. *Evolution in Changing Environments: Some Theoretical Explorations*. Princeton University Press, August 1968. ISBN 9780691080628.

[84] Hua-Xin Liao, Rebecca Lynch, Tongqing Zhou, Feng Gao, S Munir Alam, Scott D Boyd, Andrew Z Fire, Krishna M Roskin, Chaim A Schramm, Zhenhai Zhang, Jiang Zhu, Lawrence Shapiro, NISC Comparative Sequencing Program, James C Mullikin, S Gnanakaran, Peter Hraber, Kevin Wiehe, Garnett Kelsoe, Guang Yang, Shi-Mao Xia, David C Montefiori, Robert Parks, Krissey E Lloyd, Richard M Scearce, Kelly A Soderberg, Myron Cohen, Gift Kamanga, Mark K Louder, Lillian M Tran, Yue Chen, Fangping Cai, Sheri Chen, Stephanie Moquin, Xiulian Du, M Gordon Joyce, Sanjay Srivatsan, Baoshan Zhang, Anqi Zheng, George M Shaw, Beatrice H Hahn, Thomas B Kepler, Bette T M Korber, Peter D Kwong, John R Mascola, and Barton F Haynes. Co-evolution of a broadly neutralizing HIV-1 antibody and founder virus. *Nature*, 496 (7446):469–476, April 2013. ISSN 0028-0836, 1476-4687. doi: 10.1038/nature12053.

[85] Hod Lipson, Jordan B Pollack, and Nam P Suh. On the origin of modular variation. *Evolution*, 56(8):1549–1556, August 2002. ISSN 0014-3820.

[86] Marta Łuksza and Michael Lässig. A predictive fitness model for influenza. *Nature*, 507(7490):57–61, 2014. doi: 10.1038/nature13087. URL `https://doi.org/10.1038/nature13087`.

[87] Yarong Luo, Jianlang Hu, and Chi Guo. Kalman filter from the mutual information perspective. *CoRR*, abs/2101.00757, 2021. URL `https://arxiv.org/abs/2101.00757`.

[88] Francesco Mainardi and Paolo Pironi. The fractional langevin equation: Brownian motion revisited, 2008.

[89] Delphine C Malherbe, Nicole A Doria-Rose, Lynda Misher, Travis Beckett, Wendy Blay Puryear, Jason T Schuman, Zane Kraft, Jean O'Malley, Motomi Mori, Indresh Srivastava, Susan Barnett, Leonidas Stamatatos, and Nancy L Haigwood. Sequential immunization with a subtype B HIV-1 envelope quasispecies partially mimics the in vivo development of neutralizing antibodies. *J. Virol.*, 85(11):5262–5274, June 2011. ISSN 0022-538X, 1098-5514. doi: 10.1128/JVI.02419-10.

[90] Debora S. Marks, Lucy J. Colwell, Robert Sheridan, Thomas A. Hopf, Andrea Pagnani, Riccardo Zecchina, and Chris Sander. Protein 3D structure computed from evolutionary sequence variation. *PLOS ONE*, 6(12):e28766, 12 2011. doi: 10.1371/journal.pone.0028766.

[91] Loïc Marrec and Anne-Florence Bitbol. Quantifying the impact of a periodic presence of antimicrobial on resistance evolution in a homogeneous microbial population of fixed size. *J. Theor. Biol.*, 457:190–198, November 2018. ISSN 0022-5193, 1095-8541. doi: 10.1016/j.jtbi.2018.08.040.

[92] Yosef E. Maruvka, David A. Kessler, and Nadav M. Shnerb. The birth-death-mutation process: A new paradigm for fat tailed distributions. *PLOS ONE*, 6(11):1–7, 11 2011. doi: 10.1371/journal.pone.0026480. URL `https://doi.org/10.1371/journal.pone.0026480`.

[93] Sarah E. Marzen and James P. Crutchfield. Probabilistic deterministic finite automata and recurrent networks, revisited. *Entropy*, 24(1), 2022. ISSN 1099-4300. doi: 10.3390/e24010090. URL `https://www.mdpi.com/1099-4300/24/1/90`.

[94] Sarah E. Marzen and Simon DeDeo. The evolution of lossy compression. *Journal of The Royal Society Interface*, 14(130):20170166, 2017. doi: 10.1098/rsif.2017.0166. URL `https://royalsocietypublishing.org/doi/abs/10.1098/rsif.2017.0166`.

[95] Iacopo Mastromatteo and Matteo Marsili. On the criticality of inferred models. *J. Stat. Mech.: Theory Exp.*, 2011(10):P10012, oct 2011. doi: 10.1088/1742-5468/2011/10/p10012.

[96] Andreas Mayer, Vijay Balasubramanian, Thierry Mora, and Aleksandra M. Walczak. How a well-adapted immune system is organized. *Proceedings of the National Academy of Sciences*, 112(19):5950–5955, 2015. ISSN 0027-8424. doi: 10.1073/pnas.1421827112. URL `https://www.pnas.org/content/112/19/5950`.

[97] Andreas Mayer, Thierry Mora, Olivier Rivoire, and Aleksandra M Walczak. Transitions in optimal adaptive strategies for populations in fluctuating environments. *Phys Rev E*, 96(3-1):032412, September 2017. ISSN 2470-0053, 2470-0045. doi: 10.1103/PhysRevE.96.032412.

[98] Andreas Mayer, Vijay Balasubramanian, Aleksandra M. Walczak, and Thierry Mora. How a well-adapting immune system remembers. *Proceedings of the National Academy*

of Sciences, 116(18):8815–8823, 2019. ISSN 0027-8424. doi: 10.1073/pnas.1812810116. URL https://www.pnas.org/content/116/18/8815.

[99] Matthijs Meijers, Sosuke Ito, and Pieter Rein ten Wolde. Behavior of information flow near criticality. *Phys. Rev. E*, 103:L010102, Jan 2021. doi: 10.1103/PhysRevE.103.L010102.

[100] Leenoy Meshulam, Jeffrey L. Gauthier, Carlos D. Brody, David W. Tank, and William Bialek. Coarse graining, fixed points, and scaling in a large population of neurons. *Phys. Rev. Lett.*, 123:178103, Oct 2019. doi: 10.1103/PhysRevLett.123.178103.

[101] Nicholas Metropolis, Arianna W. Rosenbluth, Marshall N. Rosenbluth, Augusta H. Teller, and Edward Teller. Equation of state calculations by fast computing machines. *The Journal of Chemical Physics*, 21(6):1087–1092, 1953. doi: 10.1063/1.1699114. URL https://doi.org/10.1063/1.1699114.

[102] Beren Millidge, Alexander Tschantz, Anil Seth, and Christopher Buckley. Neural kalman filtering, 2021. URL https://arxiv.org/abs/2102.10021.

[103] Victor Minden, Cengiz Pehlevan, and Dmitri B. Chklovskii. Biologically plausible online principal component analysis without recurrent neural dynamics, 2018. URL https://arxiv.org/abs/1810.06966.

[104] Andrea Montanari and José Pereira. Which graphical models are difficult to learn? In Y. Bengio, D. Schuurmans, J. Lafferty, C. Williams, and A. Culotta, editors, *Advances in Neural Information Processing Systems*, volume 22, pages 1303–1311. Curran Associates, Inc., 2009. URL https://proceedings.neurips.cc/paper/2009/file/22fb0cee7e1f3bde58293de743871417-Paper.pdf.

[105] Thierry Mora and William Bialek. Are biological systems poised at criticality? *J. Stat. Phys.*, 144(2):268–302, 2011. doi: 10.1007/s10955-011-0229-4.

[106] Thierry Mora, Aleksandra M. Walczak, William Bialek, and Curtis G. Callan. Maximum entropy models for antibody diversity. *Proc. Natl. Acad. Sci. U.S.A.*, 107(12):5405–5410, 2010. doi: 10.1073/pnas.1001705107.

[107] Thierry Mora, Stéphane Deny, and Olivier Marre. Dynamical criticality in the collective activity of a population of retinal neurons. *Phys. Rev. Lett.*, 114:078105, Feb 2015. doi: 10.1103/PhysRevLett.114.078105.

[108] P. A. P. Moran. Random processes in genetics. *Mathematical Proceedings of the Cambridge Philosophical Society*, 54(1):60?71, 1958. doi: 10.1017/S0305004100033193.

[109] Faruck Morcos, Andrea Pagnani, Bryan Lunt, Arianna Bertolino, Debora S. Marks, Chris Sander, Riccardo Zecchina, José N. Onuchic, Terence Hwa, and Martin Weigt. Direct-coupling analysis of residue coevolution captures native contacts across many protein families. *Proc. Natl. Acad. Sci. U.S.A.*, 108(49):E1293–E1301, 2011. doi: 10.1073/pnas.1111471108.

[110] Mia C. Morrell, Audrey J. Sederberg, and Ilya Nemenman. Latent dynamical variables produce signatures of spatiotemporal criticality in large biological systems. *Phys. Rev. Lett.*, 126:118302, 2021. doi: 10.1103/PhysRevLett.126.118302.

[111] Enrique T Muñoz and Michael W Deem. Epitope analysis for influenza vaccine design. *Vaccine*, 23(9):1144–1148, January 2005. ISSN 0264-410X. doi: 10.1016/j.vaccine.2004.08.028.

[112] K. Murphy and C. Weaver. *Janeway's Immunobiology*. CRC Press, 2016. ISBN 9781315533247. URL `https://books.google.com/books?id=GmPLCwAAQBAJ`.

[113] Ville Mustonen and Michael Lässig. Molecular evolution under fitness fluctuations. *Phys. Rev. Lett.*, 100(10):108101, March 2008. ISSN 0031-9007. doi: 10.1103/PhysRevLett.100.108101.

[114] Ville Mustonen and Michael Lässig. From fitness landscapes to seascapes: non-equilibrium dynamics of selection and adaptation. *Trends Genet.*, 25(3):111–119, March 2009. ISSN 0168-9525. doi: 10.1016/j.tig.2009.01.002.

[115] Ville Mustonen and Michael Lässig. Fitness flux and ubiquity of adaptive evolution. *Proc. Natl. Acad. Sci. U. S. A.*, 107(9):4248–4253, March 2010. ISSN 0027-8424, 1091-6490. doi: 10.1073/pnas.0907953107.

[116] Vudtiwat Ngampruetikorn, Vedant Sachdeva, Johanna Torrence, Jan Humplik, David J. Schwab, and Stephanie E. Palmer. Inferring couplings in networks across order-disorder phase transitions, 2021. URL `https://arxiv.org/abs/2106.02349`.

[117] H. Chau Nguyen, Riccardo Zecchina, and Johannes Berg. Inverse statistical problems: from the inverse Ising problem to data science. *Adv. Phys.*, 66(3):197–261, 2017. doi: 10.1080/00018732.2017.1341604.

[118] Alexander P. Nikitin, Nigel G. Stocks, Robert P. Morse, and Mark D. McDonnell. Neural population coding is optimized by discrete tuning curves. *Phys. Rev. Lett.*, 103:138101, Sep 2009. doi: 10.1103/PhysRevLett.103.138101. URL `https://link.aps.org/doi/10.1103/PhysRevLett.103.138101`.

[119] Simon F. Nørrelykke and Henrik Flyvbjerg. Harmonic oscillator in heat bath: Exact simulation of time-lapse-recorded data and exact analytical benchmark statistics. *Phys. Rev. E*, 83:041103, Apr 2011. doi: 10.1103/PhysRevE.83.041103. URL `https://link.aps.org/doi/10.1103/PhysRevE.83.041103`.

[120] Armita Nourmohammad and Ceyhun Eksin. Optimal evolutionary control for artificial selection on molecular phenotypes, 2019.

[121] Armita Nourmohammad, Stephan Schiffels, and Michael Lassig. Evolution of molecular phenotypes under stabilizing selection. *Journal of Statistical Mechanics: Theory and Experiment*, 2013(01):P01012, jan 2013. doi: 10.1088/1742-5468/2013/01/p01012. URL `https://doi.org/10.1088%2F1742-5468%2F2013%2F01%2Fp01012`.

[122] Matti Nykter, Nathan D. Price, Maximino Aldana, Stephen A. Ramsey, Stuart A. Kauffman, Leroy E. Hood, Olli Yli-Harja, and Ilya Shmulevich. Gene expression dynamics in the macrophage exhibit criticality. *Proc. Natl. Acad. Sci. U.S.A.*, 105(6): 1897–1900, 2008. doi: 10.1073/pnas.0711525105.

[123] Bruno A. Olshausen and David J. Field. Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature*, 381(6583):607—609, 1996. URL `https://doi.org/10.1038/381607a0`.

[124] Amichai Painsky and Naftali Tishby. Gaussian lower bound for the information bottleneck limit. *J. Mach. Learn. Res.*, 18:213–1, 2017.

[125] Stephanie E. Palmer, Olivier Marre, Michael J. Berry, and William Bialek. Predictive information in a sensory population. *Proceedings of the National Academy of Sciences*, 112(22):6908–6913, 2015. ISSN 0027-8424. doi: 10.1073/pnas.1506855112. URL `https://www.pnas.org/content/112/22/6908`.

[126] Alan S Perelson. Immune network theory. *Immunol. Rev*, 110(5), 1989.

[127] Alan S. Perelson and George F. Oster. Theoretical studies of clonal selection: Minimal antibody repertoire size and reliability of self-non-self discrimination. *Journal of Theoretical Biology*, 81(4):645 – 670, 1979. ISSN 0022-5193. doi: https://doi.org/10.1016/0022-5193(79)90275-3. URL `http://www.sciencedirect.com/science/article/pii/0022519379902753`.

[128] Franco Pissani, Delphine C Malherbe, Harlan Robins, Victor R DeFilippis, Byung Park, George Sellhorn, Leonidas Stamatatos, Julie Overbaugh, and Nancy L Haigwood. Motif-optimized subtype a HIV envelope-based DNA vaccines rapidly elicit neutralizing antibodies when delivered sequentially. *Vaccine*, 30(37):5519–5526, August 2012. ISSN 0264-410X, 1873-2518. doi: 10.1016/j.vaccine.2012.06.042.

[129] Abe D Pressman, Ziwei Liu, Evan Janzen, Celia Blanco, Ulrich F Müller, Gerald F Joyce, Robert Pascal, and Irene A Chen. Mapping a systematic ribozyme fitness landscape reveals a frustrated evolutionary network for Self-Aminoacylating RNA. *J. Am. Chem. Soc.*, 141(15):6213–6223, April 2019. ISSN 0002-7863, 1520-5126. doi: 10.1021/jacs.8b13298.

[130] Mikhail Prokopenko, Joseph T. Lizier, Oliver Obst, and X. Rosalind Wang. Relating Fisher information to order parameters. *Phys. Rev. E*, 84:041116, Oct 2011. doi: 10.1103/PhysRevE.84.041116.

[131] James Gary Propp and David Bruce Wilson. Exact sampling with coupled Markov chains and applications to statistical mechanics. *Random Struct. Alg.*, 9:223–252, 1996. doi: 10.1002/(SICI)1098-2418(199608/09)9:1/2⟨223::AID-RSA14⟩3.0.CO;2-O.

[132] Klaus Rajewsky. Clonal selection and learning in the antibody system. *Nature*, 381(6585):751–758, 1996. doi: 10.1038/381751a0. URL `https://doi.org/10.1038/381751a0`.

[133] Arjun S Raman, K Ian White, and Rama Ranganathan. Origins of allostery and evolvability in proteins: A case study. *Cell*, 166(2):468–480, July 2016. ISSN 0092-8674, 1097-4172. doi: 10.1016/j.cell.2016.05.047.

[134] Rajesh P. N. Rao and Dana H. Ballard. Dynamic model of visual recognition predicts neural response properties in the visual cortex. *Neural Computation*, 9(4):721–763, 1997. doi: 10.1162/neco.1997.9.4.721. URL https://doi.org/10.1162/neco.1997.9.4.721.

[135] Rajesh P. N. Rao and Dana H. Ballard. Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects. *Nature Neuroscience*, 2(1):79–87, 1999. doi: 10.1038/4580. URL https://doi.org/10.1038/4580.

[136] Nicola Rigolli, Nicodemo Magnoli, Lorenzo Rosasco, and Agnese Seminara. Learning to predict target location with turbulent odor plumes, 2021. URL https://arxiv.org/abs/2106.08988.

[137] Yasser Roudi, Erik Aurell, and John Hertz. Statistical physics of pairwise probability models. *Front. Comput. Neurosci.*, 3:22, 2009. doi: 10.3389/neuro.10.022.2009.

[138] Elsa Rousseau, Benoit Moury, Ludovic Mailleret, Rachid Senoussi, Alain Palloix, Vincent Simon, Sophie Valiere, Frederic Grognard, and Frederic Fabre. Estimating virus effective population size and selection without neutral markers. *PLOS Pathogens*, 13(11):1–25, 11 2017. doi: 10.1371/journal.ppat.1006702. URL https://doi.org/10.1371/journal.ppat.1006702.

[139] Daniel Ruderman and William Bialek. Statistics of natural images: Scaling in the woods. *Advances in neural information processing systems*, 6, 1993.

[140] Daniel L. Ruderman. Origins of scaling in natural images. *Vision Research*, 37(23):3385 – 3398, 1997. ISSN 0042-6989. doi: https://doi.org/10.1016/S0042-6989(97)00008-4. URL http://www.sciencedirect.com/science/article/pii/S0042698997000084.

[141] Vedant Sachdeva, Kabir Husain, Jiming Sheng, Shenshen Wang, and Arvind Murugan. Tuning environmental timescales to evolve and maintain generalists, 2019.

[142] Vedant Sachdeva, Thierry Mora, Aleksandra M. Walczak, and Stephanie E. Palmer. Optimal prediction with resource constraints using the information bottleneck. *PLOS Computational Biology*, 17(3):1–27, 03 2021. doi: 10.1371/journal.pcbi.1008743. URL https://doi.org/10.1371/journal.pcbi.1008743.

[143] Jared M. Salisbury and Stephanie E. Palmer. Optimal prediction in the retina and natural motion statistics. *Journal of Statistical Physics*, 162(5):1309–1323, Mar 2016. ISSN 1572-9613. doi: 10.1007/s10955-015-1439-y. URL https://doi.org/10.1007/s10955-015-1439-y.

[144] Trifce Sandev, Ralf Metzler, and Živorad Tomovski. Correlation functions for the fractional generalized langevin equation in the presence of internal and external noise. *Journal of Mathematical Physics*, 55(2):023301, 2014. doi: 10.1063/1.4863478. URL https://doi.org/10.1063/1.4863478.

[145] Elad Schneidman, Michael J. Berry, Ronen Segev, and William Bialek. Weak pairwise correlations imply strongly correlated network states in a neural population. *Nature*, 440(7087):1007–1012, 2006. doi: 10.1038/nature04701.

[146] David J. Schwab, Ilya Nemenman, and Pankaj Mehta. Zipf's law and criticality in multivariate data without fine-tuning. *Phys. Rev. Lett.*, 113:068102, Aug 2014. doi: 10.1103/PhysRevLett.113.068102.

[147] Audrey J. Sederberg, Jason N. MacLean, and Stephanie E. Palmer. Learning to make external sensory stimulus predictions using internal correlations in populations of neurons. *Proceedings of the National Academy of Sciences*, 115(5):1105–1110, 2018. ISSN 0027-8424. doi: 10.1073/pnas.1710779115. URL https://www.pnas.org/content/115/5/1105.

[148] Noam Slonim. The information bottleneck: Theory and applications, 2002.

[149] Joel G. Smith. The information capacity of amplitude- and variance-constrained scalar gaussian channels. *Information and Control*, 18(3):203 – 219, 1971. ISSN 0019-9958. doi: https://doi.org/10.1016/S0019-9958(71)90346-9. URL http://www.sciencedirect.com/science/article/pii/S0019995871903469.

[150] Kayla G. Sprenger, Joy E. Louveau, and Arup K. Chakraborty. Optimizing immunization protocols to elicit broadly neutralizing antibodies. *bioRxiv*, 2020. doi: 10.1101/2020.01.04.894857. URL https://www.biorxiv.org/content/early/2020/01/06/2020.01.04.894857.

[151] Mandyam Veerambudi Srinivasan, Simon Barry Laughlin, A. Dubs, and George Adrian Horridge. Predictive coding: a fresh view of inhibition in the retina. *Proceedings of the Royal Society of London. Series B. Biological Sciences*, 216(1205):427–459, 1982. doi: 10.1098/rspb.1982.0085. URL https://royalsocietypublishing.org/doi/abs/10.1098/rspb.1982.0085.

[152] Richard R. Stein, Debora S. Marks, and Chris Sander. Inferring pairwise interactions from biological data using maximum-entropy probability models. *PLOS Comput. Biol.*, 11(7):e1004182, 07 2015. doi: 10.1371/journal.pcbi.1004182.

[153] Martin Stevens. *Sensory ecology, behaviour, and evolution*. Oxford University Press, 2013.

[154] Toshiyuki Tanaka. Mean-field theory of Boltzmann machine learning. *Phys. Rev. E*, 58:2302–2310, Aug 1998. doi: 10.1103/PhysRevE.58.2302.

[155] Qian-Yuan Tang and Kunihiko Kaneko. Long-range correlation in protein dynamics: Confirmation by structural data and normal mode analysis. *PLOS Comput. Biol.*, 16 (2):e1007670, 02 2020. doi: 10.1371/journal.pcbi.1007670.

[156] Qian-Yuan Tang, Yang-Yang Zhang, Jun Wang, Wei Wang, and Dante R. Chialvo. Critical fluctuations in the native state of proteins. *Phys. Rev. Lett.*, 118:088102, Feb 2017. doi: 10.1103/PhysRevLett.118.088102.

[157] Paula Tataru, Maria Simonsen, Thomas Bataillon, and Asger Hobolth. Statistical inference in the wright-fisher model using allele frequency data. *Systematic biology*, 66 (1):e30–e46, 01 2017. doi: 10.1093/sysbio/syw056. URL `https://pubmed.ncbi.nlm.nih.gov/28173553`.

[158] Naftali Tishby, Fernando C. Pereira, and William Bialek. The information bottleneck method. pages 368–377, 1999.

[159] Gašper Tkačik, Elad Schneidman, Michael J Berry II, and William Bialek. Ising models for networks of real neurons, 2006. URL `https://arxiv.org/abs/q-bio/0611072`.

[160] Gašper Tkačik, Olivier Marre, Thierry Mora, Dario Amodei, Michael J Berry II, and William Bialek. The simplest maximum entropy model for collective behavior in a neural network. *J. Stat. Mech.: Theory Exp.*, 2013(03):P03011, mar 2013. doi: 10.1088/1742-5468/2013/03/p03011.

[161] Gašper Tkačik, Thierry Mora, Olivier Marre, Dario Amodei, Stephanie E. Palmer, Michael J. Berry II, and William Bialek. Thermodynamics and signatures of criticality in a network of neurons. *Proc. Natl. Acad. Sci. U.S.A.*, 112(37):11508–11513, 2015. doi: 10.1073/pnas.1514188112.

[162] Yutaka Tomita, S. Ohmatsu, and Takashi Soeda. An application of the information theory to estimation problems. *Inf. Control.*, 32:101–111, 1976.

[163] Erdal Toprak, Adrian Veres, Jean-Baptiste Michel, Remy Chait, Daniel L Hartl, and Roy Kishony. Evolutionary paths to antibiotic resistance under dynamically sustained drug selection. *Nat. Genet.*, 44(1):101–105, December 2011. ISSN 1061-4036, 1546-1718. doi: 10.1038/ng.1034.

[164] Hildegard Uecker and Joachim Hermisson. On the fixation process of a beneficial mutation in a variable environment. *Genetics*, 188(4):915–930, August 2011. ISSN 0016-6731, 1943-2631. doi: 10.1534/genetics.110.124297.

[165] N.G. Van Kampen. *Stochastic Processes in Physics and Chemistry*. North-Holland Personal Library. Elsevier Science, 1992. ISBN 9780080571386. URL `https://books.google.com/books?id=3e7XbMoJzmoC`.

[166] E van Nimwegen and J P Crutchfield. Metastable evolutionary dynamics: crossing fitness barriers or escaping via neutral paths? *Bull. Math. Biol.*, 62(5):799–848, September 2000. ISSN 0092-8240. doi: 10.1006/bulm.2000.0180.

[167] Christophe Verbeurgt, Françoise Wilkin, Maxime Tarabichi, Françoise Gregoire, Jacques E Dumont, and Pierre Chatelain. Profiling of olfactory receptor gene expression in whole human olfactory mucosa. *PloS one*, 9(5):e96333–e96333, 05 2014. doi: 10.1371/journal.pone.0096333. URL https://pubmed.ncbi.nlm.nih.gov/24800820.

[168] Aleksandra M. Walczak, Ga šper Tkačik, and William Bialek. Optimizing information flow in small genetic networks. ii. feed-forward interactions. *Phys. Rev. E*, 81:041905, Apr 2010. doi: 10.1103/PhysRevE.81.041905. URL https://link.aps.org/doi/10.1103/PhysRevE.81.041905.

[169] Chaohui Wang, Nikos Komodakis, and Nikos Paragios. Markov random field modeling, inference & learning in computer vision & image understanding: A survey. *Comput. Vis. Image. Underst.*, 117(11):1610–1627, 2013. doi: 10.1016/j.cviu.2013.07.004.

[170] Jian Wang, Kangkun Mao, Yunjie Zhao, Chen Zeng, Jianjin Xiang, Yi Zhang, and Yi Xiao. Optimization of RNA 3D structure prediction using evolutionary restraints of nucleotide–nucleotide interactions from direct coupling analysis. *Nucleic Acids Res.*, 45 (11):6299–6309, 05 2017. doi: 10.1093/nar/gkx386.

[171] Shenshen Wang. Optimal sequential immunization can focus antibody responses against diversity loss and distraction. *PLoS Comput. Biol.*, 13(1):e1005336, January 2017. ISSN 1553-734X, 1553-7358. doi: 10.1371/journal.pcbi.1005336.

[172] Shenshen Wang and Lei Dai. Evolving generalists in switching rugged landscapes. *PLOS Computational Biology*, 15(10):1–21, 10 2019. doi: 10.1371/journal.pcbi.1007320. URL https://doi.org/10.1371/journal.pcbi.1007320.

[173] Shenshen Wang, Jordi Mata-Fink, Barry Kriegsman, Melissa Hanson, Darrell J Irvine, Herman N Eisen, Dennis R Burton, K Dane Wittrup, Mehran Kardar, and Arup K Chakraborty. Manipulating the selection forces during affinity maturation to generate cross-reactive HIV antibodies. *Cell*, 160(4):785–797, February 2015. ISSN 0092-8674, 1097-4172. doi: 10.1016/j.cell.2015.01.027.

[174] Siwei Wang, Idan Segev, Alexander Borst, and Stephanie Palmer. Maximally efficient prediction in the early fly visual system may support evasive flight maneuvers. *PLOS Computational Biology*, 17(5):1–27, 05 2021. doi: 10.1371/journal.pcbi.1008965. URL https://doi.org/10.1371/journal.pcbi.1008965.

[175] Yihang Wang, João Marcelo Lamim Ribeiro, and Pratyush Tiwary. Past–future information bottleneck for sampling molecular reaction coordinate simultaneously with thermodynamics and kinetics. *Nature Communications*, 10(1):3573, 2019. doi: 10.1038/s41467-019-11405-4. URL https://doi.org/10.1038/s41467-019-11405-4.

[176] Martin Weigt, Robert A. White, Hendrik Szurmant, James A. Hoch, and Terence Hwa. Identification of direct residue contacts in protein–protein interaction by message passing. *Proc. Natl. Acad. Sci. U.S.A.*, 106(1):67–72, 2009. doi: 10.1073/pnas.0805923106.

[177] Caleb Weinreb, Adam J. Riesselman, John B. Ingraham, Torsten Gross, Chris Sander, and Debora S. Marks. 3D RNA and functional interactions from evolutionary couplings. *Cell*, 165(4):963–975, 2016. doi: 10.1016/j.cell.2016.03.030.

[178] Daniel B Weissman, Michael M Desai, Daniel S Fisher, and Marcus W Feldman. The rate at which asexual populations cross fitness valleys. *Theor. Popul. Biol.*, 75(4): 286–300, June 2009. ISSN 0040-5809, 1096-0325. doi: 10.1016/j.tpb.2009.02.006.

[179] Arthur Winfree. The geometry of biological time. 1980.

[180] S Wright. The differential equation of the distribution of gene frequencies. *Proceedings of the National Academy of Sciences of the United States of America*, 31(12):382–389, 12 1945. doi: 10.1073/pnas.31.12.382. URL `https://pubmed.ncbi.nlm.nih.gov/16588707`.

[181] F. Y. Wu. The Potts model. *Rev. Mod. Phys.*, 54:235–268, Jan 1982. doi: 10.1103/RevModPhys.54.235.

[182] Tailin Wu and Ian Fischer. Phase transitions for the information bottleneck in representation learning. In *International Conference on Learning Representations*, 2020. URL `https://openreview.net/forum?id=HJloElBYvB`.

[183] Tailin Wu, Ian Fischer, Isaac L. Chuang, and Max Tegmark. Learnability for the information bottleneck. *Entropy*, 21(10):924, Sep 2019. ISSN 1099-4300. doi: 10.3390/e21100924. URL `http://dx.doi.org/10.3390/e21100924`.

[184] Xueling Wu, Zhi-Yong Yang, Yuxing Li, Carl-Magnus Hogerkorp, William R Schief, Michael S Seaman, Tongqing Zhou, Stephen D Schmidt, Lan Wu, Ling Xu, Nancy S Longo, Krisha McKee, Sijy O'Dell, Mark K Louder, Diane L Wycuff, Yu Feng, Martha Nason, Nicole Doria-Rose, Mark Connors, Peter D Kwong, Mario Roederer, Richard T Wyatt, Gary J Nabel, and John R Mascola. Rational design of envelope identifies broadly neutralizing human monoclonal antibodies to HIV-1. *Science*, 329(5993): 856–861, August 2010. ISSN 0036-8075, 1095-9203. doi: 10.1126/science.1187659.

[185] Bingkan Xue and Stanislas Leibler. Evolutionary learning of adaptation to varying environments through a transgenerational feedback. *Proceedings of the National Academy of Sciences*, 113(40):11266–11271, October 2016. doi: 10.1073/pnas.1608756113.

[186] Jonathan S Yedidia. An idiosyncratic journey beyond mean field theory. In Manfred Opper and David Saad, editors, *Advanced mean field methods: theory and practice*, Neural Information Processing, chapter 3, pages 21–36. MIT Press, Cambridge MA, 2001. ISBN 0262150549.

[187] Lenka Zdeborová. Understanding deep learning is also a job for physicists. *Nat. Phys.*, 16(6):602–604, 2020. doi: 10.1038/s41567-020-0929-2.

[188] Christina Zelano, Aprajita Mohanty, and JayÂ A. Gottfried. Olfactory predictive codes and stimulus templates in piriform cortex. *Neuron*, 72(1):178 – 187, 2011. ISSN 0896-6273. doi: https://doi.org/10.1016/j.neuron.2011.08.010. URL `http://www.sciencedirect.com/science/article/pii/S0896627311007318`.