Optimal prediction with resource constraints using the information bottleneck

Vedant Sachdeva[1], Thierry Mora[2], Aleksandra M. Walczak[3], Stephanie Palmer[4]
[1] *Graduate Progam in Biophysical Sciences, University of Chicago, Chicago, IL,* [2] *Laboratoire de physique statistique, CNRS, Sorbonne Université, Université Paris-Diderot, and École Normale Supérieure (PSL University), 24, rue Lhomond, 75005 Paris, France,* [3] *Laboratoire de physique théorique, CNRS, Sorbonne Université, and École Normale Supérieure (PSL University), 24, rue Lhomond, 75005 Paris, France,* [4] *Deparment of Organismal Biology and Anatomy, University of Chicago, Chicago, IL**
(Dated: April 29, 2020)

Responding to stimuli requires that organisms encode information about the external world. Not all parts of the signal 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 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 coordinates prove more predictive. We identify the properties of global, transferrable strategies for time-varying stimuli. For non-Markovian dynamics we explore the role of long-term memory of the internal representation. Lastly, we show that prediction in evolutionary population dynamics is linked to clustering allele frequencies into non-overlapping memories, revealing a very different prediction strategy from motion prediction.

I. 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 capacity constraints, has been successful in explaining some early sensory representations in neuroscience. Barlow suggested sensory circuits may reduce redundancy in the neural code and minimize metabolic costs for signal transmission [1–4]. 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 [5–7], 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 [8–10]. Molecular networks have also been shown to be predictive of future states, suggesting prediction may be one of the underlying principles of biological computation [11, 12]. 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 [12–14]. 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.

The information bottleneck framework allows us to define a 'relevance' variable in the encoded sensory stream, which we take 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, prediction of 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 stochastic prediction by calculating the future form of the probability distribution of the variable X_t or its moments [15, 16]. The higher-order moments yield an estimate of X_t and the uncertainty in the our estimate. However, biological systems cannot precisely know X_t due to inherently limited readout precision [17, 18] and limited availability of resources tasked with remembering the measured statistics.

Constructing internal representations of sensory stimuli illustrates a tension between predicting the future, for which the past must be known with higher certainty, and compression of knowledge of the past, due to finite resources. We explore this intrinsic trade-off using the in-

*Correspondence should be addressed to sepalmer@uchicago.edu

formation bottleneck (IB) approach proposed by Tishby et. al. [13]. This method assumes that the input variable, in our case the past signal X_t, can be used to make inferences about the relevance variable, in our case the future signal $X_{t+\Delta t}$. By introducing a representation variable, $\tilde{X}$, we can construct the conditional distribution of the representation variable on the input variable $\mathcal{P}(\tilde{X}|X_t)$ to be maximally informative of the output variable (Fig. 1).

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

$$\mathcal{L} = \min_{\mathcal{P}(\tilde{X}|X_t)} I(X_t;\tilde{X}) - \beta I(\tilde{X};X_{t+\Delta t}). \tag{1}$$

Each term is the mutual information between two variables: the first between the past input and estimate of the past 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, the representation variable 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 the past and the future 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 the future cannot exceed the amount encoded about the past [19]. Additionally, the information about the past that the representation can extract is bounded by the amount of information the uncompressed past, itself, contains about the future, $I(\tilde{X};X_{t+\Delta t}) \leq I(X_t;X_{t+\Delta t})$.

We use this framework to study prediction 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 simultaneously 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 with different noise profiles to study the effect of noise correlations on prediction. Our exploration of the SDDHO system has a two-fold motivation: it is the simplest possible continuous stochastic system whose full dynamics can be solved exactly. Additionally, a visual stimulus in the form of a moving bar that was driven by an SDDHO process was used in retinal response studies [9, 20, 21]. The Wright-Fisher model [22] is a canonical model of evolution [23] for which has been used to consider how the adaptive immune system predicts the future state of the pathogenic environment [11, 24].

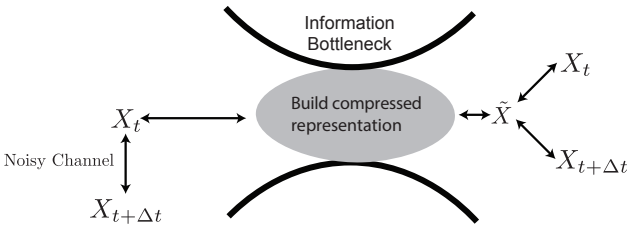

FIG. 1: A schematic representation our predictive information bottleneck. On the left hand side, we have coordinates X_t evolving in time, subject to noise to give $X_{t+\Delta t}$. We construct a representation, $\tilde{X}$, that compresses the past input (minimizes $I(X_t;\tilde{X})$) while retaining as much information about the future (maximizes $I(\tilde{X};X_{t+\Delta t})$) up to the weighting of the prediction compared to the compression set by β.

II. RESULTS

A. 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 [9] 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 [9]. 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. 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 Figure 2a. 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

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

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 $D = \frac{k_B T}{\Gamma}$,

$$\frac{d\bar{v}}{d\bar{t}} = -\frac{\bar{x}}{4\tau^2\zeta^2} - \frac{\bar{v}}{\tau} + \frac{\sqrt{2D}}{\tau}\xi(\bar{t}), \qquad (3)$$
$$\frac{d\bar{x}}{d\bar{t}} = \bar{v}.$$

We introduce a dimensionless parameter, the damping coefficient, $\zeta = 1/(2\omega_0\tau)$. When $\zeta < 1$, the motion of the mass will be oscillatory. When $\zeta \geq 1$, the motion will be non-oscillatory. Additionally, we note that the equipartition theorem tells us that $\langle \bar{x}(\bar{t})^2 \rangle \equiv x_0^2 = k_B T/k = D/(\tau\omega_0^2)$

Expressing the equations of motion in terms of ζ, τ, and x_0, we obtain

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

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

$$\frac{dv}{dt} = -\frac{x}{4\zeta^2} - v + \frac{\xi(t)}{\sqrt{2}\zeta}. \qquad (5)$$

There are now 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. 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 Appendix A) [25],

$$\tilde{X} = A_\beta X_t + \xi. \qquad (6)$$

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 Appendix B.2). The past and future variables in the SDDHO bottleneck problem are jointly Gaussian, which means that the optimal compression can be summarized by its second-order statistics. We generalize analytically the results that were numerically obtained in Ref. [9] 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 the past conditioned on the compression, $\mathcal{P}(X_t|\tilde{X})$, the distribution of the future conditioned the compressed variable $\mathcal{P}(X_{t+\Delta t}|\tilde{X})$, and the distribution of the future conditioned on exact knowledge of the past $\mathcal{P}(X_{t+\Delta t}|X_t)$. We represent the uncertainty reduction 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 $\mathcal{P}(X_t)$.

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 λ_2 of the matrix $\Sigma_{X_t|X_{t+\Delta t}}\Sigma_{X_t}^{-1}$ as $\beta_c = 1/(1-\lambda_2)$ (see Appendix B.2). $\Sigma_{X_t|X_{t+\Delta t}}$ 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 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. 2b). However, if the encoding contains information about the past, 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. 2b), 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.

From Figure 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. 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 a full information bottleneck curve for this damping coefficient and desired prediction timescale, as shown in Figure 3. Velocity uncertainty is only reduced as we allow for less compression, as shown in Fig. 3a (right). For both of the cases represented in Fig. 3a, the illustrated encoding strategy yields a maximal amount of mutual information

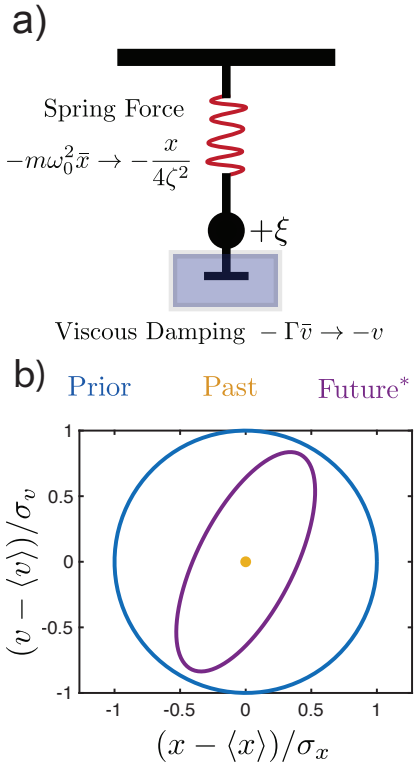

FIG. 2: Schematic of the stochastically driven damped harmonic oscillator (SDDHO). (a) The SDDHO consists of a mass attached to a spring undergoing viscous damping and experiencing Gaussian thermal noise of magnitude. There are two parameters to be explored in this model: $\zeta = \frac{1}{2\omega_0\tau}$ and $\Delta t = \frac{\Delta t}{\tau}$. (b) We can represent the statistics of the stimulus through error ellipses. $\zeta = \frac{1}{2}$, and $\Delta t = 1$, we plot two-dimensional confidence intervals under various conditions. In blue, we plot the two-dimensional confidence interval of the prior. In yellow, we plot the certainty with which we measure the position and velocity at time t. Here, it is measured with infinite precision, meaning $I_{\text{past}} \to \infty$. In purple, we plot the two-dimensional confidence interval of the future conditioned on the measurement given in yellow, for this particular choice of parameters. 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.

between the compressed representation, $\tilde{X}$, and the future for the given level of compression, as indicated by the red dots in Fig. 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 Appendix A. 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

$$w_1 = \begin{bmatrix} \omega\cot(\omega\Delta t) + \frac{|\csc(\omega\Delta t)|}{2\sqrt{2}\zeta}\sqrt{2-\zeta^2-\zeta^2\cos(2\omega\Delta t)} \\ 1 \end{bmatrix}. \tag{7}$$

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

$$\phi = \arctan\left(\left(\omega\cot(\omega\Delta t) + \frac{|\csc(\omega\Delta t)|}{2\sqrt{2}\zeta}\sqrt{2-\zeta^2-\zeta^2\cos(2\omega\Delta t)}\right)^{-1}\right). \tag{8}$$

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

$$\phi = \arctan\left(\frac{\Delta t}{1+\omega^2}\right). \tag{9}$$

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

$$\phi = \arctan\left(\frac{2\sinh(\frac{\Delta t}{2})}{\cosh(\frac{\Delta t}{2}) + \sqrt{\frac{1+\cosh(\Delta t)}{2}}}\right). \tag{10}$$

In the large Δt limit, $\phi \to \frac{\pi}{4}$. In the small Δt limit, $\phi \to \arctan(\Delta t)$. Past 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

$$\phi = \arctan\left(\frac{2\zeta\sin(\frac{\Delta t}{2\zeta})}{\cos(\frac{\Delta t}{2\zeta}) + \sqrt{2-\zeta^2-\zeta^2\cos(\frac{\Delta t}{\zeta})}}\right). \tag{11}$$

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 the past, 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

$$1-\lambda_1 = \exp(-\Delta t)\left(\frac{1}{4\omega^2\zeta^2}+\frac{\cos(2\omega\Delta t)}{4\omega^2}+\frac{|\sin(\omega\Delta t)|}{2\sqrt{2}\omega^2\zeta}\sqrt{2-\zeta^2-\zeta^2\cos(2\omega\Delta t)}\right). \quad (12)$$

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

$$I(X_t; X_{t+\Delta t}) = \frac{1}{2}\log(|\Sigma_{X_t}|) - \frac{1}{2}\log(|\Sigma_{X_t|X_{t+\Delta t}}|). \quad (13)$$

For large Δt, this expression becomes

$$I(X_t; X_{t+\Delta t}) \propto \exp(-\Delta t). \quad (14)$$

For small Δt,

$$I(X_t; X_{t+\Delta t}) \propto \Delta t - \frac{1}{2}\log(\Delta t). \quad (15)$$

The constants emerge from the physical parameters of the input dynamics.

1. *Optimal representations in all parameter regimes for fixed past information*

We sweep over all possible parameter regimes of the SDDHO keeping $I(X_t;\tilde{X})$ fixed to 5 bits and find the optimal representation for a variety of timescales (Fig. 4), keeping a fixed amount of information encoded about the past for each realization of the stimulus and prediction. More information can be transmitted for shorter delays (Fig. 4a,d,g) between the past and future signal than for longer delays (Fig. 4c,f,i). In addition, at shorter prediction timescales more information about the past 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. 4a), the evolution of the equations of motion are well approximated by integrating Eq. 3 with the left hand side set to zero, and the optimal representation encodes mostly positional information. This can be observed by noting that the encoding ellipse remains on-axis and mostly compressed in the position dimension. For the underdamped case, in short time predictions (Fig. 4g), a similar strategy is effective. However, for longer predictions (Fig. 4h,i), 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. 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. 4f, the optimal encoding is non-predictive.

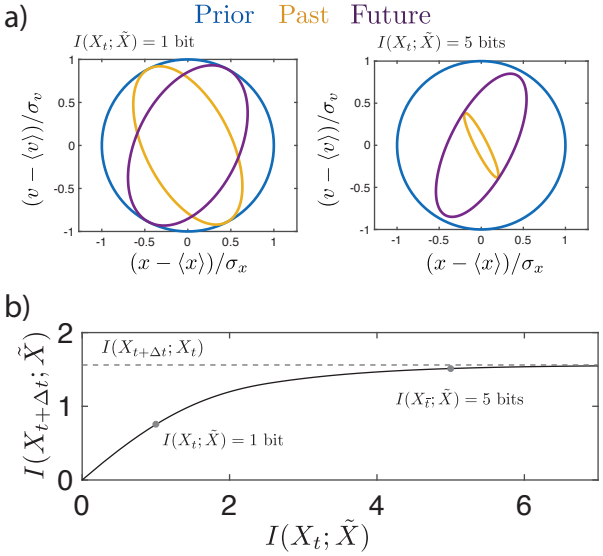

FIG. 3: We consider the task of predicting the path of an SDDHO with $\zeta = \frac{1}{2}$ and $\Delta t = 1$. (a) (left) We encode the history of the stimulus, X_t, with a representation generated by the information bottleneck, $\tilde{X}$, that can store 1 bit of information. Knowledge of the coordinates 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 the future scales with the information in the past, highlighting the points represented in panel (a).

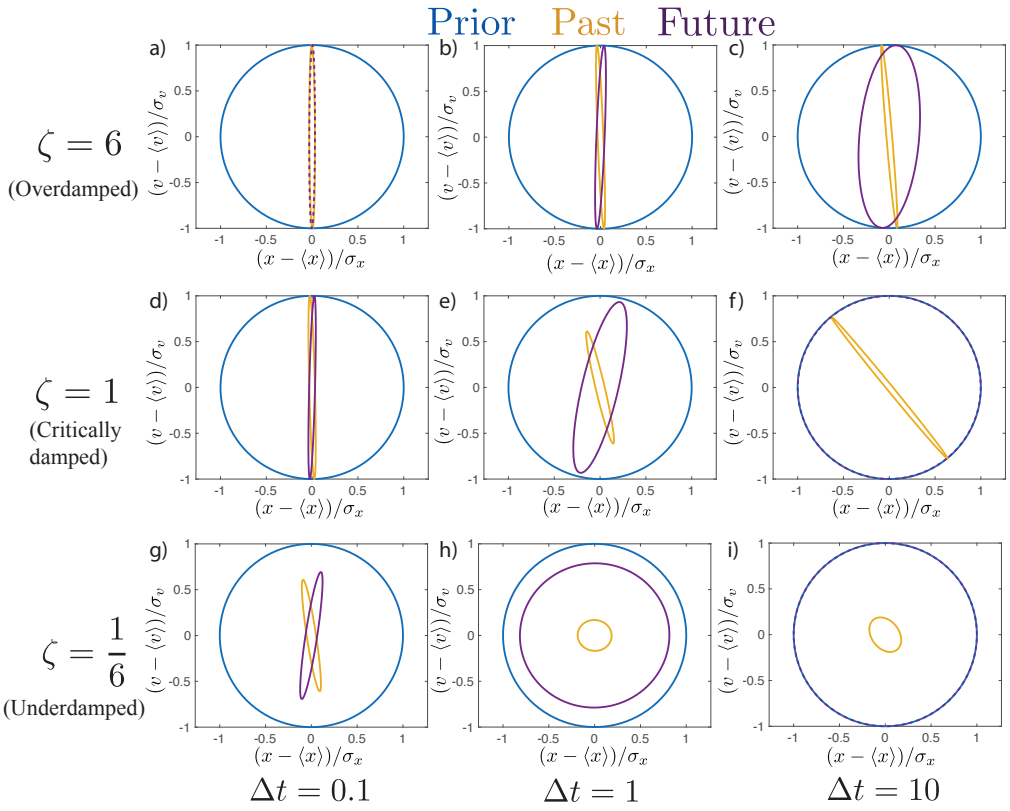

FIG. 4: Possible behaviors associated for the SDDHO for a variety of timescales with a fixed $I(X_t; \tilde{X})$ of 5 bits. For an overdamped SDDHO, panel a-c, the optimal representation continues to encode mostly position information, as velocity is hard to predict. For the underdamped case, panels 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.

2. *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 the past, $\mathcal{P}(X_t|\tilde{X}_{\text{suboptimal}})$, and the future, $\mathcal{P}(X_{t+\Delta t}|\tilde{X}_{\text{suboptimal}})$. For a fixed amount of information about the past, $I(X_t; \tilde{X}_{\text{optimal}}) = I(X_t, \tilde{X}_{\text{suboptimal}})$, we compare the predictive information in the optimal (Fig. 5a) and the suboptimal representations (Fig. 5b). In this example, we are exploring the impact of encoding velocity with high certainty as compared to encoding position with high certainty. We observe that encoding velocity provides very little predictive power, indicating that encoding velocity and position is not equally important, even for equal compression levels. In addition, it shows that encoding schemes discovered by IB are optimal for predictive purposes.

3. *Kalman filters versus information bottleneck*

We can also compare our information bottleneck solutions to what one would obtain using Kalman filters [26]. We note that Kalman filters are not designed to be *efficient* strategies for extracting predictive information, as shown in the Appendix, Figure B.1. This is because the Kalman filter approach does not constrain the representation entropy (i.e. it does not have a resource-limit constraint). A Kalman filter also always explicitly makes a model of the dynamics that generate updates to the input variable, an explicit model of the 'physics of the external world'. The information bottleneck framework enables exploration of representations without explicitly developing an internal model of the dynamics and also includes resource constraints. Thus, for a given amount of compression, the information bottleneck solution to the prediction problem is as predictive as possible, whereas

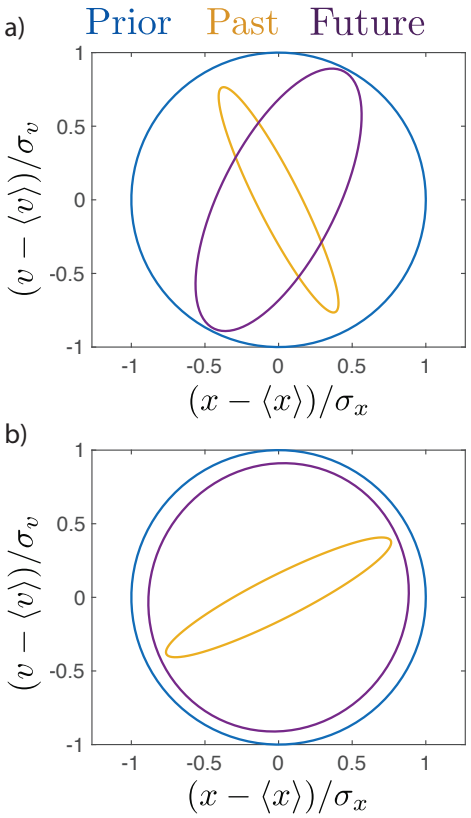

FIG. 5: Example of a sub-optimal compression. An optimal predictive, compressed representation, in panel (a) compared to a suboptimal representation, in panel (b) for a prediction of $\Delta t = 1$ away in the underdamped regime ($\zeta = 1/2$). We fix the mutual information between the representations and the past ($I(X_t;\tilde{X}) = 3$ bits), but find that, as expected, the sub-optimal representation contains significantly less information about the future.

a Kalman filter may miss important predictive features of the input while representing noisy, unpredictable features. In that sense, the Kalman filter approach is agnostic about what input bits matter for prediction, and is a less efficient coding scheme of predictive information for a given channel capacity.

4. *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. 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. Figure 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 Figure 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 Figure 6c, (Figure 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 Figure 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.

B. History-dependent Gaussian Stimuli

In the above analysis, we considered stimuli with correlations that fall off exponentially. However, natural

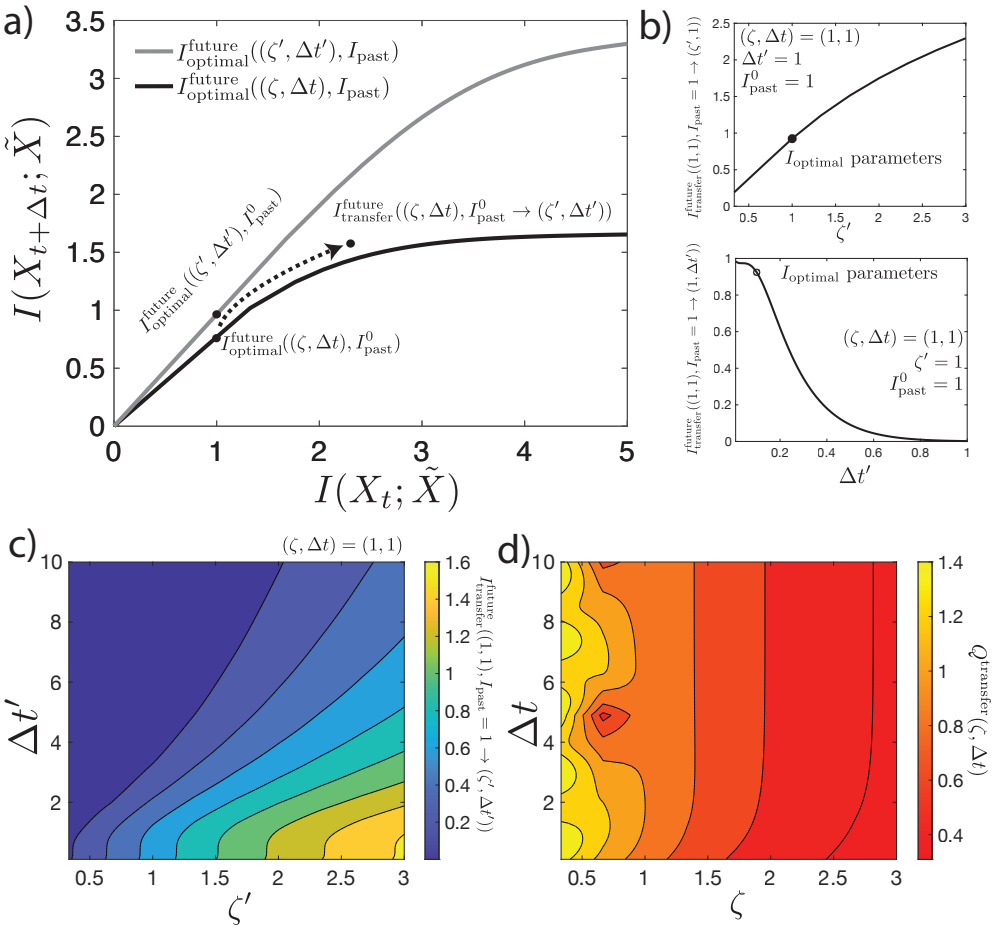

FIG. 6: Representations learned on underdamped systems can be transferred to other types of motion, while representations learned on overdamped systems cannot be easily transferred. (a) Here, we consider the information bottleneck bound curve (black) for a stimulus with underlying parameters, $(\zeta, \Delta t)$. 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.

scenes, such as leaves blowing in the wind or bees moving in their hives, are shown to have heavy-tailed statistics [21, 27, 28], and we extend our results to models of motion stimuli with heavy-tailed temporal correlation. Despite long-ranged temporal order, prediction is still possible. We show this through the use of the Generalized Langevin equation [29–31]:

$$\frac{dv}{dt} = -\int_0^t \frac{\gamma v}{|t-t'|^\alpha} dt - \omega_0^2 x + \xi(t) \tag{16}$$

$$\frac{dx}{dt} = v \tag{17}$$

Here, we have returned to unscaled definitions of v, and t. The damping force here is 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 SDDHO, 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. 7a). Using the approach described in Appendix A, 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,

$$\mathcal{L} = \min_{\mathcal{P}(\tilde{X}|X_{t-t_0:t})} I(X_{t-t_0:t};\tilde{X}) - \beta I(X_{t+\Delta t:t+\Delta t+t_0};\tilde{X}). \tag{18}$$

The information bottleneck curves show more predictive information as the prediction process uses more past information (larger t_0 in Fig. 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. 7c). These results indicate that efficient coding strategies can enable prediction even in complex temporally correlated environments.

C. 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 up an immunity. Exploiting these correlations can be done at a population level, in terms of vaccine design [32–35], and has been postulated as a means for the immune system to adapt to future threats [11, 36]. Here, we present efficient predictive coding strategies for the Wright-Fisher model, which is commonly used to describe viral evolution [37]. In contrast to the two models studied so far, Wright-Fisher dynamics are not Gaussian. We use this model to explore how the results obtained in the previous sections generalize to non-Gaussian statistics of the past and future distributions.

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. 8a. For large enough N, the update rule of the allele frequencies is given through the diffusion approximation interpreted with the Ito convention [38]:

$$\frac{dX_t}{dt} = sX_t(1-X_t) + \mu(1-2X_t) + \sqrt{X_t(1-X_t)/N}\eta(t), \tag{19}$$

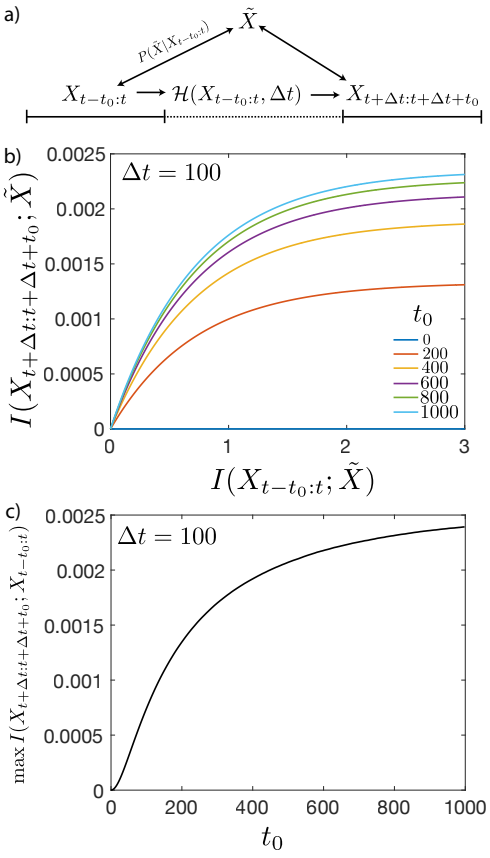

FIG. 7: The ability of the information bottleneck Method to predict history-dependent stimuli. (a) The prediction problem, using an extended history and a future. This problem is largely similar to the one set up for the SDDHO but the past and the future are larger composites of observations within a window of time $t-t_0 : t$ for the past and $t+\Delta t : t+\Delta t+t_0$ 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.

where $\langle\eta(t)\rangle = 0$, $\langle\eta(t)\eta(t')\rangle = \delta(t-t')$.

For this model, defining the representation $\tilde{X}$ as a noisy linear transformation of the past frequency X_t as we did for the Gaussian case in Eq. 21 does not capture all of the dependencies of the future on the past due to the non-Gaussian character of the joint distribution of $X_{t+\Delta t}$ and X_t stemming from the non-linear form of Eq. 19. Instead, we determine the mapping of X_t to $\tilde{X}$ numerically using the Blahut-Arimoto algorithm [39, 40]. For ease of computation, we will take the representation variable $\tilde{X}$ to be discrete (Fig. 8b) and later, approximate continuous $\tilde{X}$ by driving the cardinality of $\tilde{X}$, denoted by m, to be high. The assumption that $\tilde{X}$ is discrete results in each realization of the representation tiling a distinct part of frequency space. This encoding scheme can be thought of as lymphocyte antigen-receptors in the adaptive immune system corresponding to different regions of phenotypic space [41].

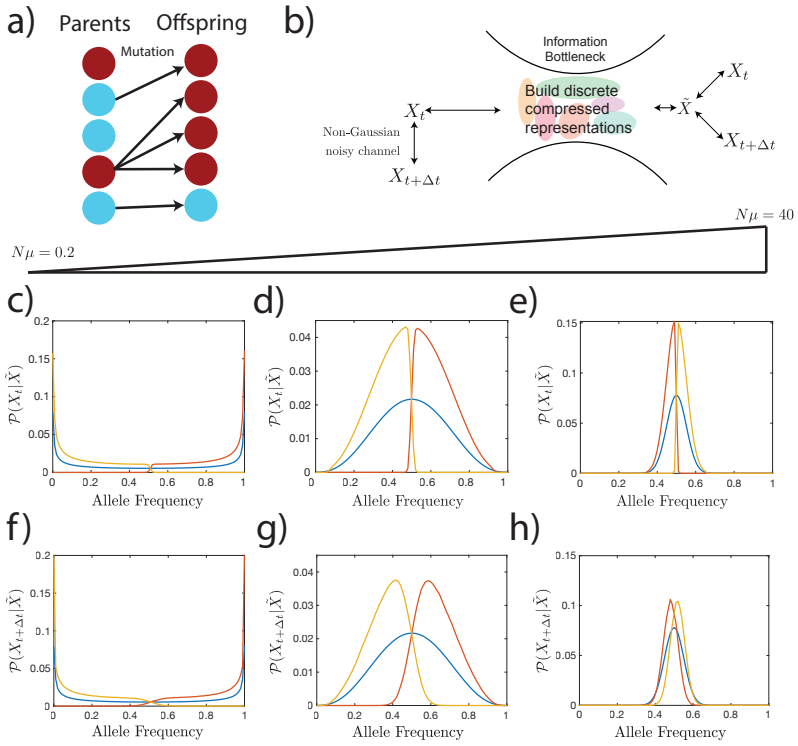

FIG. 8: The information bottleneck solution for a Wright Fisher process. (a) The Wright-Fisher model of evolution can be visualized as a population of N parents giving rise to a population of N children. Genotypes of the children are selected as a function of the parents' generation genotypes subject to mutation rates, μ, and selective pressures s. (b) Information bottleneck schematic with a discrete (rather than continuous) representation variable, $\tilde{X}$. (c-h) 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). (c-e) 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. (f-h) 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.

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,

$$P_s(X) \propto [X(1-X)]^{N\mu-1} e^{NsX} \tag{20}$$

(blue line in Fig. 8c) 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. 8c, 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. 8f. The predictive information conferred by the representation variable is limited by the information it has about the past, as in the Gaussian case (Fig. 10a.)

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. 8d and e). 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. 8g and h), 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 past distribution biases the representation variable towards frequency space regions with larger steady state probability mass.

In Fig. 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

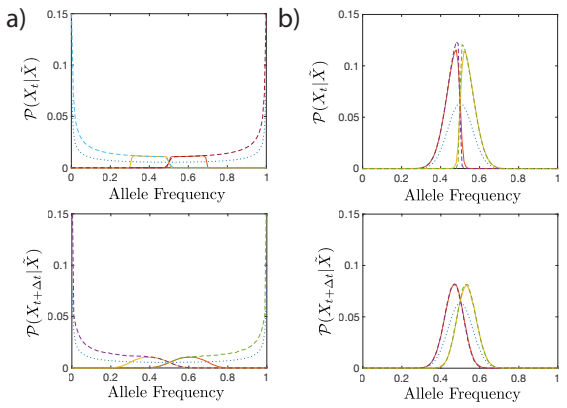

FIG. 9: Transferability of prediction schemes in Wright-Fisher dynamics. We transfer a mapping, $\mathcal{P}(\tilde{X}|X_t)$, trained on one set of parameters and apply it to another. We consider transfers between two choices of mutability, $N\mu_1 = 0.2$ (low) and $N\mu_2 = 20$ (high), with $N = 100$, $Ns = 0.001$, $\Delta t = 1$. 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: $I^{\text{past}}_{\text{transfer}}((N\mu_2), I_{\text{past}} = 0.92 \rightarrow (N\mu_1)) = 0.14$ and $I^{\text{future}}_{\text{transfer}}((N\mu_2), I_{\text{past}} = 0.92 \rightarrow (N\mu_1)) = 0.05$. 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: $I^{\text{past}}_{\text{transfer}}((N\mu_1), I_{\text{past}} = 0.98 \rightarrow (N\mu_2)) = 0.79$ and $I^{\text{future}}_{\text{transfer}}((N\mu_1), I_{\text{past}} = 0.98 \rightarrow (N\mu_2)) = 0.27$. Transfer in this direction yields good predictive informations.

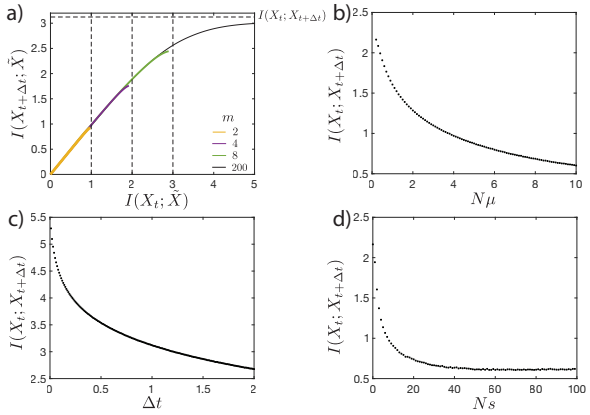

FIG. 10: Amount of predictive information in the Wright Fisher dynamics as a function of the model parameters. (a) Predictive information as a function of compression level. Predictive information increases with the cardinality, m, of the representation variable. The amount of predictive information is limited by $\log(m)$ (vertical dashed lines) for small m, and the mutual information between the future and the past, $I(X_{t+\Delta t}; X_t)$ (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: $N = 100$, $N\mu = 0.2$, $N\mu = 0.2$, $Ns = 0.001$, $\Delta t = 1$. (b-d), Value of the asymptote of the information bottleneck curve, $I(X_t; X_{t+\Delta t})$ with: (b) $N = 100$, $Ns = 0.001$, $\Delta t = 1$ as a function of μ; (c) $N = 100$, $N\mu = 0.2$, $Ns = 0.001$ as a function of Δt; and (d) $N = 100$, $N\mu = 0.2$, and $\Delta t = 1$ as a function of s.

information, as seen in Fig. 10b. 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. 8 to see how the predictive information depends on the strength of the selection and mutation rates (Fig. 10b and d). 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. 10c).

So far we have discussed the results for $m = 2$ representations. As we increase the tradeoff parameter, β in Eq. 1, the amount of predictive information increases, since we retain more information about the past. 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. 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. 10a)). Comparing the $m = 2$ and $m = 3$ representations for maximum values of β for each of them (Fig. 11a and b), shows that larger numbers of representations tile allele frequency space more finely, allowing for more precise encodings of the past and fu-

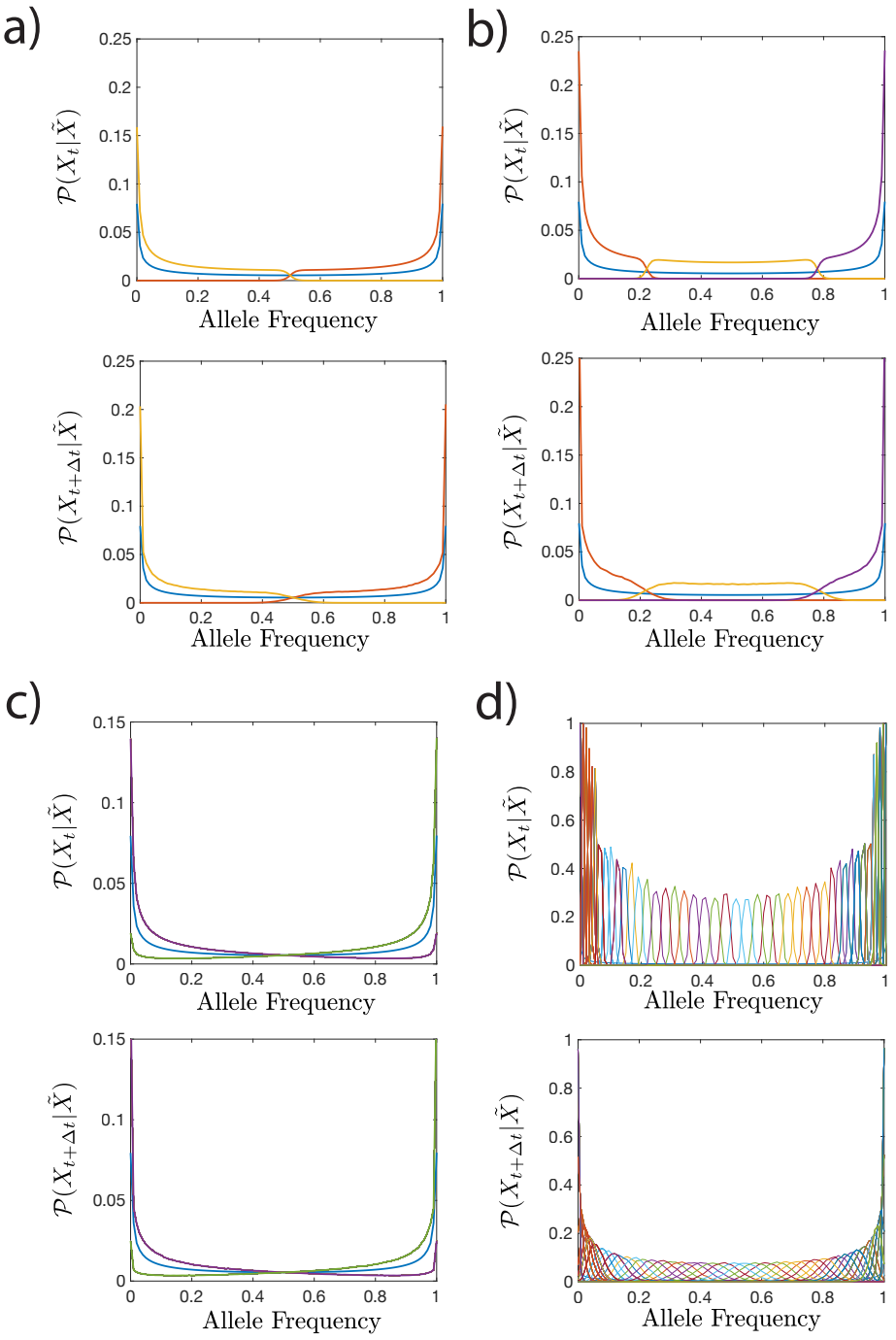

FIG. 11: Encoding schemes with $m > 2$ representation variables. The representations which carry maximum predictive information for $m = 2$ at $I(X_t;\tilde{X}) \approx \log(m) = 1$(a) and $m = 3$ at $I(X_t;\tilde{X}) \approx \log(m) \approx 1.5$. (b). The optimal representations at large m tile space 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 Appendix Fig. C.2. At high $I(X_t;\tilde{X})$, the degeneracy is lifted. All computations done at $N = 100$, $N\mu = 0.2$, $Ns = 0.001$, $\Delta t = 1$.

ture distributions. The maximum amount of information about the past goes as $\log(m)$ (Fig. 10a). The predictive information curves for different m values are the same, until the branching point $\lesssim \log(m)$ for each m (Fig. 10a).

We analyze the nature of this branching by taking $m \gg 1$, $m = 200$ (Fig. 11c and d). 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. 11c), with additional degenerate representations (Fig. C.2). 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 [42, 43].

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. 11d top panel, Fig. C.3). 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 [2], but with the added constraint of choosing the tiling to enable the most informative predictions.

III. 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.

We hypothesize that biological systems that need to operate flexibly across a wide range of different input statistics may use a best-compromise predictive encoding of their inputs. We have used a transferability metric, Q, to quantify just how useful a particular scheme is across other dynamic regimes and prediction timescales. What we have shown is 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 [26]. 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 [44]. In this framework, the Kalman filters seek to maximize information by minimizing the variance of the true coordinates of an external input and the estimate of those coordinates. 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 Appendix that the Kalman filter approach is not as efficient, in general, as the predictive information bottleneck approach that we present here.

The evolutionary context shows another set of solutions to predictive information in terms of discrete representations that tile input space. Although we impose discrete representations, their non-overlapping character remains even it the limit of many 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 [2], since input space is covered proportionally to the steady state distribution at a given frequency. 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 [45]. 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 [46, 47].

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[24]. 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 [48, 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. 8d and e), and the variance of the encoding distributions depends on the the value of $\mathcal{P}(X_t|\tilde{X})$ (Fig. 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[14, 50, 51]. 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 IB 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 IB 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.

Acknowledgments

This work was supported in part by the US National Science Foundation, through the Center for the Physics of Biological Function (PHY–1734030), and a CAREER award to SEP (1652617); by the National Institutes of Health BRAIN initiative (R01EB026943–01); by a FACCTS grant from the France Chicago Center; and by European Research Council Consolidator Grant (724208).

[1] Barlow HB. Possible Principles Underlying the Transformation of Sensory Messages. In: Sensory communication. MIT Press; 2012. .

[2] Laughlin SB. A Simple Coding Procedure Enhances a Neuron's Information Capacity. Zeitschrift für Naturforschung C. 1981;36:910 – 912.

[3] de Ruyter van Steveninck RR, Laughlin SB. The rate of information transfer at graded-potential synapses. Na-

ture. 1996;379(6566):642–645. Available from: https://doi.org/10.1038/379642a0.

[4] Olshausen BA, Field DJ. Emergence of simple-cell receptive field properties by learning a sparse code for natural images. Nature. 1996;381(6583):607—609. Available from: https://doi.org/10.1038/381607a0.

[5] Bialek W, Nemenman I, Tishby N. Predictability, Complexity, and Learning. Neural Computation. 2001;13(11):2409–2463. Available from: https://doi.org/10.1162/089976601753195969.

[6] Lee TS, Mumford D. Hierarchical Bayesian inference in the visual cortex. J Opt Soc Am A. 2003 Jul;20(7):1434–1448. Available from: http://josaa.osa.org/abstract.cfm?URI=josaa-20-7-1434.

[7] Rao RPN, Ballard DH. Dynamic Model of Visual Recognition Predicts Neural Response Properties in the Visual Cortex. Neural Computation. 1997;9(4):721–763. Available from: https://doi.org/10.1162/neco.1997.9.4.721.

[8] Rao RPN, Ballard DH. Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects. Nature Neuroscience. 1999;2(1):79–87. Available from: https://doi.org/10.1038/4580.

[9] Palmer SE, Marre O, Berry MJ, Bialek W. Predictive information in a sensory population. Proceedings of the National Academy of Sciences. 2015;112(22):6908–6913. Available from: https://www.pnas.org/content/112/22/6908.

[10] Zelano C, Mohanty A, Gottfried J. Olfactory Predictive Codes and Stimulus Templates in Piriform Cortex. Neuron. 2011;72(1):178 – 187. Available from: http://www.sciencedirect.com/science/article/pii/S0896627311007318.

[11] Mayer A, Balasubramanian V, Walczak AM, Mora T. How a well-adapting immune system remembers. Proceedings of the National Academy of Sciences. 2019;116(18):8815–8823. Available from: https://www.pnas.org/content/116/18/8815.

[12] Wang Y, Ribeiro JML, Tiwary P. Past–future information bottleneck for sampling molecular reaction coordinate simultaneously with thermodynamics and kinetics. Nature Communications. 2019;10(1):3573. Available from: https://doi.org/10.1038/s41467-019-11405-4.

[13] Tishby N, Pereira FC, Bialek W. The Information Bottleneck Method; 1999. p. 368–377.

[14] Alemi AA. Variational Predictive Information Bottleneck; 2019.

[15] Gardiner CW. Handbook of stochastic methods for physics, chemistry and the natural sciences. vol. 13 of Springer Series in Synergetics. 3rd ed. Berlin: Springer-Verlag; 2004.

[16] Van Kampen NG. Stochastic Processes in Physics and Chemistry. North-Holland Personal Library. Elsevier Science; 1992. Available from: https://books.google.com/books?id=3e7XbMoJzmoC.

[17] Berg HC, Purcell EM. Physics of chemoreception. Biophysical Journal. 1977;20(2):193 – 219. Available from: http://www.sciencedirect.com/science/article/pii/S0006349577855446.

[18] Bialek W. Biophysics: Searching for Principles. Princeton University Press; 2012. Available from: https://books.google.com/books?id=5In_FKA2rmUC.

[19] Beaudry NJ, Renner R. An intuitive proof of the data processing inequality; 2011.

[20] Sederberg AJ, MacLean JN, Palmer SE. Learning to make external sensory stimulus predictions using internal correlations in populations of neurons. Proceedings of the National Academy of Sciences. 2018;115(5):1105–1110. Available from: https://www.pnas.org/content/115/5/1105.

[21] Salisbury JM, Palmer SE. Optimal Prediction in the Retina and Natural Motion Statistics. Journal of Statistical Physics. 2016 Mar;162(5):1309–1323. Available from: https://doi.org/10.1007/s10955-015-1439-y.

[22] Wright S. The Differential Equation of the Distribution of Gene Frequencies. Proceedings of the National Academy of Sciences of the United States of America. 1945 12;31(12):382–389. Available from: https://pubmed.ncbi.nlm.nih.gov/16588707.

[23] Tataru P, Simonsen M, Bataillon T, Hobolth A. Statistical Inference in the Wright-Fisher Model Using Allele Frequency Data. Systematic biology. 2017 01;66(1):e30–e46. Available from: https://pubmed.ncbi.nlm.nih.gov/28173553.

[24] Mayer A, Balasubramanian V, Mora T, Walczak AM. How a well-adapted immune system is organized. Proceedings of the National Academy of Sciences. 2015;112(19):5950–5955. Available from: https://www.pnas.org/content/112/19/5950.

[25] Chechik G, Globerson A, Tishby N, Weiss Y. Information Bottleneck for Gaussian Variables. In: Thrun S, Saul LK, Schölkopf B, editors. Advances in Neural Information Processing Systems 16. MIT Press; 2004. p. 1213–1220. Available from: http://papers.nips.cc/paper/2457-information-bottleneck-for-gaussian-variables.pdf.

[26] Kalman RE. A New Approach to Linear Filtering and Prediction Problems. Transactions of the ASME–Journal of Basic Engineering. 1960;82(Series D):35–45.

[27] Billock VA, de Guzman GC, Kelso JAS. Fractal time and 1/f spectra in dynamic images and human vision. Physica D: Nonlinear Phenomena. 2001;148(1):136 – 146. Available from: http://www.sciencedirect.com/science/article/pii/S0167278900001743.

[28] Ruderman DL. Origins of scaling in natural images. Vision Research. 1997;37(23):3385 – 3398. Available from: http://www.sciencedirect.com/science/article/pii/S0042698997000084.

[29] Sandev T, Metzler R, Tomovski v. Correlation functions for the fractional generalized Langevin equation in the presence of internal and external noise. Journal of Mathematical Physics. 2014;55(2):023301. Available from: https://doi.org/10.1063/1.4863478.

[30] Mainardi F, Pironi P. The Fractional Langevin Equation: Brownian Motion Revisited; 2008.

[31] Jeon JH, Metzler R. Fractional Brownian motion and motion governed by the fractional Langevin equation in confined geometries. Phys Rev E. 2010 Feb;81:021103. Available from: https://link.aps.org/doi/10.1103/PhysRevE.81.021103.

[32] Łuksza M, Lässig M. A predictive fitness model for influenza. Nature. 2014;507(7490):57–61. Available from: https://doi.org/10.1038/nature13087.

[33] Dolan PT, Whitfield ZJ, Andino R. Mapping the Evolutionary Potential of RNA Viruses. Cell Host & Microbe. 2018;23(4):435 – 446. Available from: http://www.sciencedirect.com/science/

article/pii/S1931312818301410.

[34] Wang S, Mata-Fink J, Kriegsman B, Hanson M, Irvine DJ, Eisen HN, et al. Manipulating the selection forces during affinity maturation to generate cross-reactive HIV antibodies. Cell. 2015 Feb;160(4):785–797.

[35] Sachdeva V, Husain K, Sheng J, Wang S, Murugan A. Tuning environmental timescales to evolve and maintain generalists; 2019.

[36] Nourmohammad A, Eksin C. Optimal evolutionary control for artificial selection on molecular phenotypes; 2019.

[37] Rousseau E, Moury B, Mailleret L, Senoussi R, Palloix A, Simon V, et al. Estimating virus effective population size and selection without neutral markers. PLOS Pathogens. 2017 11;13(11):1–25. Available from: https://doi.org/10.1371/journal.ppat.1006702.

[38] Kimura M. Diffusion Models in Population Genetics. Journal of Applied Probability. 1964;1(2):177–232. Available from: http://www.jstor.org/stable/3211856.

[39] Arimoto S. An algorithm for computing the capacity of arbitrary discrete memoryless channels. IEEE Transactions on Information Theory. 1972 January;18(1):14–20.

[40] Blahut RE. Computation of channel capacity and rate-distortion functions. IEEE Trans Inform Theory. 1972;18:460–473.

[41] Murphy K, Weaver C. Janeway's Immunobiology. CRC Press; 2016. Available from: https://books.google.com/books?id=GmPLCwAAQBAJ.

[42] Wu T, Fischer I, Chuang IL, Tegmark M. Learnability for the Information Bottleneck. Entropy. 2019 Sep;21(10):924. Available from: http://dx.doi.org/10.3390/e21100924.

[43] Wu T, Fischer I. Phase Transitions for the Information Bottleneck in Representation Learning. In: International Conference on Learning Representations; 2020. Available from: https://openreview.net/forum?id=HJloElBYvB.

[44] Husain K, Pittayakanchit W, Pattanayak G, Rust MJ, Murugan A. Kalman-like Self-Tuned Sensitivity in Biophysical Sensing. Cell Systems. 2019;9(5):459 – 465.e6. Available from: http://www.sciencedirect.com/science/article/pii/S2405471219303060.

[45] Walczak AM, Tkačik Gcv, Bialek W. Optimizing information flow in small genetic networks. II. Feed-forward interactions. Phys Rev E. 2010 Apr;81:041905. Available from: https://link.aps.org/doi/10.1103/PhysRevE.81.041905.

[46] Nikitin AP, Stocks NG, Morse RP, McDonnell MD. Neural Population Coding Is Optimized by Discrete Tuning Curves. Phys Rev Lett. 2009 Sep;103:138101. Available from: https://link.aps.org/doi/10.1103/PhysRevLett.103.138101.

[47] Smith JG. The information capacity of amplitude- and variance-constrained scalar gaussian channels. Information and Control. 1971;18(3):203 – 219. Available from: http://www.sciencedirect.com/science/article/pii/S0019995871903469.

[48] Verbeurgt C, Wilkin F, Tarabichi M, Gregoire F, Dumont JE, Chatelain P. Profiling of olfactory receptor gene expression in whole human olfactory mucosa. PloS one. 2014 05;9(5):e96333–e96333. Available from: https://pubmed.ncbi.nlm.nih.gov/24800820.

[49] Dunkel M, Schmidt U, Struck S, Berger L, Gruening B, Hossbach J, et al. SuperScent?a database of flavors and scents. Nucleic Acids Research. 2008 10;37(suppl_1):D291–D294. Available from: https://doi.org/10.1093/nar/gkn695.

[50] Chalk M, Marre O, Tkačik G. Toward a unified theory of efficient, predictive, and sparse coding. Proceedings of the National Academy of Sciences. 2018;115(1):186–191. Available from: https://www.pnas.org/content/115/1/186.

[51] Alemi AA, Fischer I, Dillon JV, Murphy K. Deep Variational Information Bottleneck; 2016.

[52] Nørrelykke SF, Flyvbjerg H. Harmonic oscillator in heat bath: Exact simulation of time-lapse-recorded data and exact analytical benchmark statistics. Phys Rev E. 2011 Apr;83:041103. Available from: https://link.aps.org/doi/10.1103/PhysRevE.83.041103.

[53] Srinivasan MV, Laughlin SB, Dubs A, Horridge GA. Predictive coding: a fresh view of inhibition in the retina. Proceedings of the Royal Society of London Series B Biological Sciences. 1982;216(1205):427–459. Available from: https://royalsocietypublishing.org/doi/abs/10.1098/rspb.1982.0085.

IV. SUPPORTING INFORMATION

Appendix A COMPUTING THE OPTIMAL REPRESENTATION FOR JOINTLY GAUSSIAN PAST-FUTURE DISTRIBUTIONS

We follow Chechik, et al.[25] to analytically construct 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 $\tilde{X}$

$$\tilde{X} = A_\beta X_t + \xi. \tag{21}$$

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[25].

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

$$v_i^T \Sigma_{X_t|X_{t+\Delta t}} \Sigma_{X_t}^{-1} = \lambda_i v_i^T. \tag{22}$$

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

The matrix, A_β, is then given by

$$A_\beta = \begin{bmatrix} \alpha_1 v_1^T \\ \alpha_2 v_2^T \\ \vdots \end{bmatrix}. \tag{23}$$

α_i are scalar values given by

$$\begin{aligned} \alpha_i &= \sqrt{\frac{\beta(1-\lambda_i)-1}{\lambda_i v_i^T \Sigma_{X_t} v_i}} \quad \text{if } \beta > \frac{1}{1-\lambda_i} \\ \alpha_i &= 0 \qquad\qquad\qquad\qquad\quad \text{otherwise.} \end{aligned} \tag{24}$$

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 $\tilde{X}$

$$\Sigma_{\tilde{X}} = \langle \tilde{X}^T \tilde{X} \rangle = \langle \tilde{X}^T A_\beta^T A_\beta \tilde{X} \rangle + \langle \xi^T \xi \rangle, \tag{25}$$

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

$$\Sigma_{\tilde{X}} = A_\beta \Sigma_{X_t} A_\beta^T + I_2. \tag{26}$$

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

$$I(X_t; \tilde{X}) = \frac{1}{2} \log_2(|A_\beta \Sigma_{X_t} A_\beta^T + I_2|), \tag{27}$$

$$I(X_{t+\Delta t}; \tilde{X}) = I(X_t; \tilde{X}) - \frac{1}{2} \sum_{i=1}^{n(\beta)} \log_2(\beta(1-\lambda_i)),$$

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

$$\begin{aligned} \Sigma_{\tilde{X}X_t} &= A_\beta \Sigma_{X_t} \\ \Sigma_{\tilde{X}X_{t+\Delta t}} &= A_\beta \Sigma_{X_{t+\Delta t}X_t}. \end{aligned} \tag{28}$$

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

$$\begin{aligned} \Sigma_{X_t|\tilde{X}} &= \Sigma_{X_t} - \Sigma_{X_t\tilde{X}}\Sigma_{\tilde{X}}^{-1}\Sigma_{X_t\tilde{X}}^T \\ \Sigma_{X_{t+\Delta t}|\tilde{X}} &= \Sigma_{X_{t+\Delta t}} - \Sigma_{X_{t+\Delta t}\tilde{X}}\Sigma_{\tilde{X}}^{-1}\Sigma_{X_{t+\Delta t}\tilde{X}}^T. \end{aligned} \tag{29}$$

Appendix B THE STOCHASTICALLY DRIVEN DAMPED HARMONIC OSCILLATOR

.1 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

$$\begin{aligned} m\frac{dv}{dt} &= -\Gamma v(t) - kx + (2k_BT\Gamma)^{1/2}\xi(t) \\ \frac{dx}{dt} &= v. \end{aligned} \tag{30}$$

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

$$\begin{aligned} \frac{dv}{dt} &= -\frac{x}{4\zeta^2} - v + \frac{\xi(t)}{\sqrt{2}\zeta} \\ \frac{dx}{dt} &= v. \end{aligned} \tag{31}$$

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 $\Sigma_{X_tY_{t+\Delta t}}$. 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. [52], we can compute the requisite autocorrelations to describe the cross-covariance matrix, $\Sigma_{X_tX_{t+\Delta t}}$.

We begin by using the equipartition theorem that states that

$$\begin{aligned} \langle x_0^2 \rangle &= 1 \\ \langle x_0 v_o \rangle &= 0 \\ \langle v_0^2 \rangle &= \frac{1}{4\zeta^2}. \end{aligned} \tag{32}$$

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

$$\Sigma_{X_tX_{t+\Delta t}} = \exp\left(-\frac{\Delta t}{2}\right)\begin{bmatrix} \cos(\omega\Delta t) + \frac{\sin(\omega\Delta t)}{2\omega} & -\frac{\sin(\omega\Delta t)}{4\zeta^2\omega} \\ \frac{\sin(\omega\Delta t)}{4\zeta^2\omega} & \frac{\cos(\omega\Delta t)}{4\zeta^2} - \frac{\sin(\omega\Delta t)}{8\omega\zeta^2} \end{bmatrix} \tag{33}$$

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. [29].

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

$$\begin{aligned}\Sigma_{X_t|X_{t+\Delta t}} &= \Sigma_{X_t} - \Sigma_{X_t X_{t+\Delta t}} \Sigma_{X_t}^{-1} \Sigma_{X_t X_{t+\Delta t}} \\ \Sigma_{X_{t+\Delta t}|X_t} &= \Sigma_{X_t} - \Sigma_{X_t X_{t+\Delta t}}^T \Sigma_{X_t}^{-1} \Sigma_{X_t X_{t+\Delta t}}\end{aligned} \tag{34}$$

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).

.2 Applying the information bottleneck Solution

To apply the information bottleneck solution, we construct the matrix, $\Sigma_{Y_u|Y_{u+\Delta u}} \Sigma_{Y_u}^{-1}$, 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

$$w = \begin{bmatrix} \omega\cot(\omega\Delta t) + \frac{|\csc(\omega\Delta t)|}{2\sqrt{2}\zeta}\sqrt{2-\zeta^2-\zeta^2\cos(2\omega\Delta t)} & \omega\cot(\omega\Delta t) - \frac{|\csc(\omega\Delta t)|}{2\sqrt{2}\zeta}\sqrt{2-\zeta^2-\zeta^2\cos(2\omega\Delta t)} \\ 1 & 1 \end{bmatrix}. \tag{35}$$

The eigenvalues are then

$$\begin{aligned}\lambda_1 &= 1 - \exp(-\Delta t)\left(\frac{1}{4\omega^2\zeta^2} + \frac{\cos(2\omega\Delta t)}{4\omega^2} + \frac{|\sin(\omega\Delta t)|}{2\sqrt{2}\omega^2\zeta}\sqrt{2-\zeta^2-\zeta^2\cos(2\omega\Delta t)}\right) \\ \lambda_2 &= 1 - \exp(-\Delta t)\left(\frac{1}{4\omega^2\zeta^2} + \frac{\cos(2\omega\Delta t)}{4\omega^2} - \frac{|\sin(\omega\Delta t)|}{2\sqrt{2}\omega^2\zeta}\sqrt{2-\zeta^2-\zeta^2\cos(2\omega\Delta t)}\right)\end{aligned} \tag{36}$$

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 $\frac{\exp(-\Delta t)|2\sin(\omega\Delta t)|}{\zeta}$. 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 Figure 5.

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

$$I(X_t; X_{t+\Delta t}) = \frac{1}{2}\log(|\Sigma_{X_t}|) - \frac{1}{2}\log(|\Sigma_{X_t|X_{t+\Delta t}}|). \tag{37}$$

Simplifying this expression yields

$$\begin{aligned}I(X_t; X_{t+\Delta t}) = \Delta t - \frac{1}{2}\log\Bigg(&\exp(2\Delta t) + \cos^4(\omega\Delta t) - \sin^4(\omega\Delta t) \\ &-2\exp(\Delta t)\left(\cos^2(\omega\Delta t) + \frac{1+\zeta^2}{1-\zeta^2}\sin^2(\omega\Delta t)\right) + 2\sin^2(\omega\Delta t)\Bigg)\end{aligned} \tag{38}$$

We can see for very large Δt, this expression becomes

$$I(X_t; X_{t+\Delta t}) \sim \Delta t - \frac{1}{2}\log\left(\exp(2\Delta t) - 2\exp(\Delta t)\right). \tag{39}$$

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,

$$I(X_t; X_{t+\Delta t}) \propto \frac{1}{2}\log(|\Sigma_{X_t}|) \tag{40}$$

where there are some constants that set the units of the information and the reference point. For widths larger than k, the expression becomes:

$$I(X_t; X_{t+\Delta t}) \propto \exp(-\Delta t) \tag{41}$$

.3 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 4b 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 Figure 4b 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.

.4 Comparing the information bottleneck method to Kalman filters

The Kalman filter approach seeks to fuse a prediction of a system's coordinates at time $t + \Delta t$ based on initial coordinates at time t and knowledge of the dynamical system with an observation at time $t+\Delta t$ to increase the certainty in the inference of the coordinates at time $t + \Delta t$[26]. We present Kalman filters here to highlight the differences between the information bottleneck method and Kalman filtering techniques. First, we consider the structure of the Kalman filter,

$$\begin{aligned} X^{(\text{naive})}(\Delta t) &= \mathcal{H}(X_{0:\Delta t}, \Delta t)X(0) + \xi(\Delta t) \\ \tilde{X}^{(\text{measured})}(\Delta t) &= AX(\Delta t) + \chi(\Delta t) \\ X^{(\text{corrected})} &= X^{(\text{naive})}(\Delta t) + K_{\Delta t}(\tilde{X}^{(\text{measured})}(\Delta t) - AX^{(\text{naive})}(\Delta t)). \end{aligned} \tag{42}$$

The first equation here considers an initial condition, $X(0)$, a dynamical system model, $\mathcal{H}(X_{0:\Delta t}, \Delta t)$, and a particular noise condition to construct an estimate of where an observer might expect their system to be after some time, Δt, has passed from initial time 0. The second equation constructs a measurement, $\tilde{X}(\Delta t)$ of the true coordinates, using a known observation model, A, and some measurement noise, $\chi(\Delta t)$. A is analogous to the probabilistic mapping constructed in the information bottleneck scheme, $\tilde{X}$; however, unlike in information bottleneck, in Kalman filtering, A is given to the algorithm and not discovered by any optimization procedure. Finally, the third equation unites the measurement, $\tilde{X}$, and the guess, $X(\Delta u)$, by choosing $K_{\Delta t}$, to be the transform which minimizes the variance between the true coordinates and the corrected coordinates. This correction is a post hoc correction and is not present in the information bottleneck scheme.

We now compare the results from a Kalman filtering technique and the information bottleneck when they both use the same probabilistic mapping, $\tilde{X}$. From Figure B.1(a), we see that even though the Kalman filter has higher levels of I_{past} for the same probabilistic mapping, the Kalman filtering algorithm is not efficient, as it is not extracting the most available predictive information.

.5 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

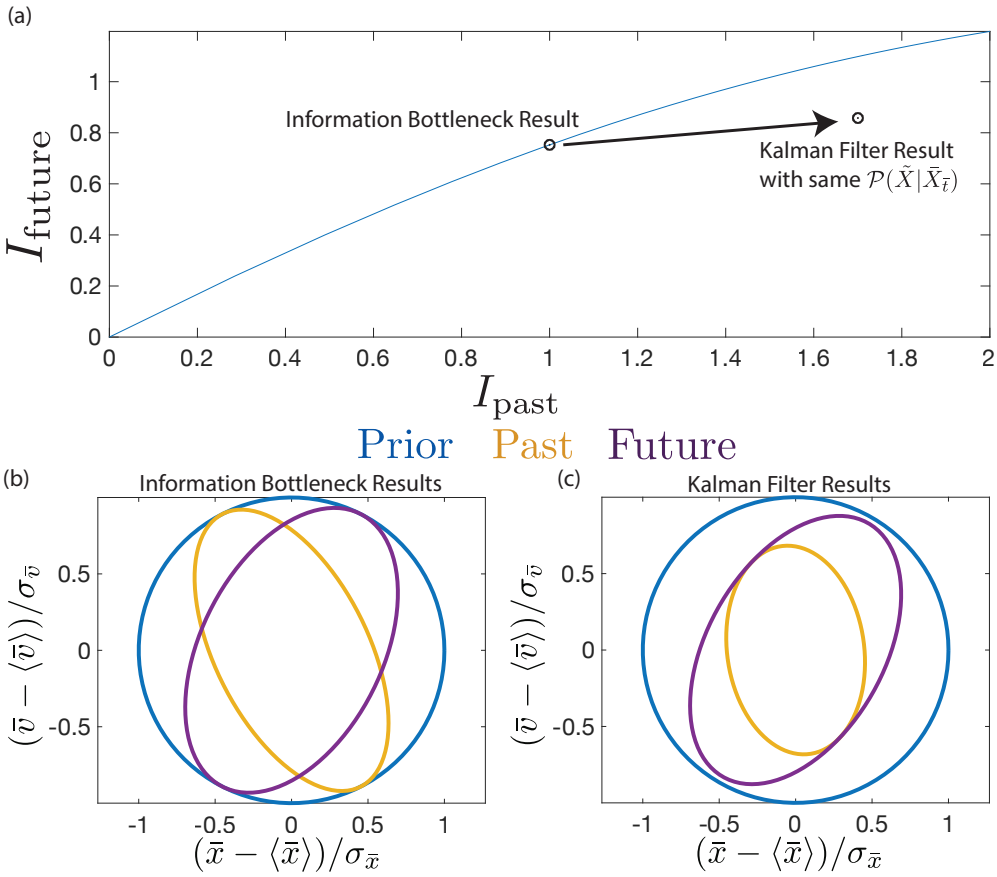

FIG. B.1: Kalman filtering schemes are not efficient coders for a given channel capacity. (a) Here, we present the information bottleneck curve for a stochastically driven damped harmonic oscillator with $\zeta = \frac{1}{2}$ and $\Delta t = 1$. We determine the optimal mapping for $I_{\text{past}} = 1$ bit and plot that point along the information bottleneck curve. Using the same probablistic mapping, we apply a Kalman filtering approach. We see that the Kalman filter approach results in an increase in both I_{past} and I_{future}, but the result does not lie along the curve, indicating the scheme is not efficient. Panels (b) and (c) present the results in terms of uncertainty reduction in each scheme.

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 $I^{\text{future}}_{\text{transfer}}((\zeta, \Delta t), I_{\text{past}} \to (\zeta', \Delta t'))$.

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

$$Q^{\text{transfer}}((\zeta, \Delta t)) = \frac{\int_{\Delta t_{\min}}^{\Delta t_{\max}} \int_{\zeta_{\min}}^{\zeta_{\max}} I^{\text{future}}_{\text{transfer}}((\zeta, \Delta t), I_{\text{past}} \to (\zeta', \Delta t')) d\zeta' d\Delta t'}{\int_{\Delta t_{\min}}^{\Delta t_{\max}} \int_{\zeta_{\min}}^{\zeta_{\max}} I^{\text{future}}_{\text{optimal}}((\zeta', \Delta t'), I_{\text{past}} d\zeta' d\Delta t'} \tag{43}$$

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.

.6 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

$$\begin{aligned}\frac{dv}{dt} &= -\int_0^t \frac{\gamma v}{|t-t'|^\alpha} dt' - \omega_0^2 x + \xi(t) \\ \frac{dx}{dt} &= v,\end{aligned} \tag{44}$$

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

$$\begin{aligned}\mathcal{L}[\langle v(t)v(0)\rangle] &= \frac{s}{s^2 + \gamma s^\alpha + \omega^2} \\ \mathcal{L}[\langle v(t)x(0)\rangle] &= -\frac{1}{s^2 + \gamma s^\alpha + \omega^2} \\ \mathcal{L}[\langle x(t)v(0)\rangle] &= -\mathcal{L}[\langle v(t)x(0)\rangle] \\ \mathcal{L}[\langle x(t)x(0)\rangle] &= \frac{1}{\omega^2 s} - \frac{1}{s^2 + \gamma s^\alpha + \omega^2}.\end{aligned} \tag{45}$$

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 $dt = 2$ and compute correlation functions along the discrete points of time, yielding the full covariance matrices. After this, the recipe is as outlined in Appendix A.

Appendix C 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. 19. We numerically integrate Eq. 19 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 (Fig. 10) using:

$$I(X_t; X_{t+\Delta t}) = -\sum_{X_t} \mathcal{P}(X_t) \log(\mathcal{P}(X_t)) - \sum_{X_{t+\Delta t}} \mathcal{P}(X_{t+\Delta t}) \log(\mathcal{P}(X_{t+\Delta t})) + \sum_{X_t, X_{t+\Delta t}} \mathcal{P}(X_t, X_{t+\Delta t}) \log(\mathcal{P}(X_{t+\Delta t}, X_t)). \tag{46}$$

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[2, 53]. The upper bound on the information the representation variable has about the past state is $I(X_t; \tilde{X}) = \log(m)$.

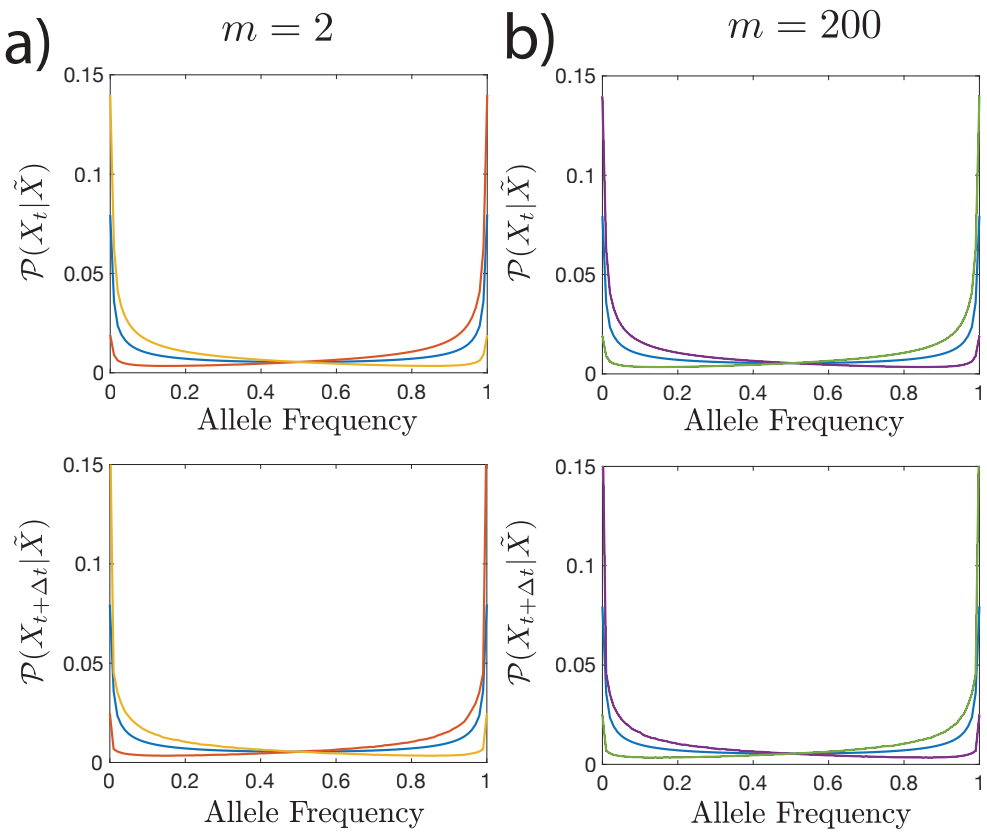

FIG. C.2: The optimal $P(X_t|\tilde{X})$ and $P(X_{t+\Delta t}|\tilde{X})$ for Wright Fisher dynamics with $N = 100$, $N\mu = 0.2$, $Ns = 0.001$, $\Delta t = 1$ with information bottleneck parameters $\beta = 1.01$ ($I(X_t;\tilde{X}) = 0.27$) for $m = 2$ (a) and $m = 200$ (b). Many representations are 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.

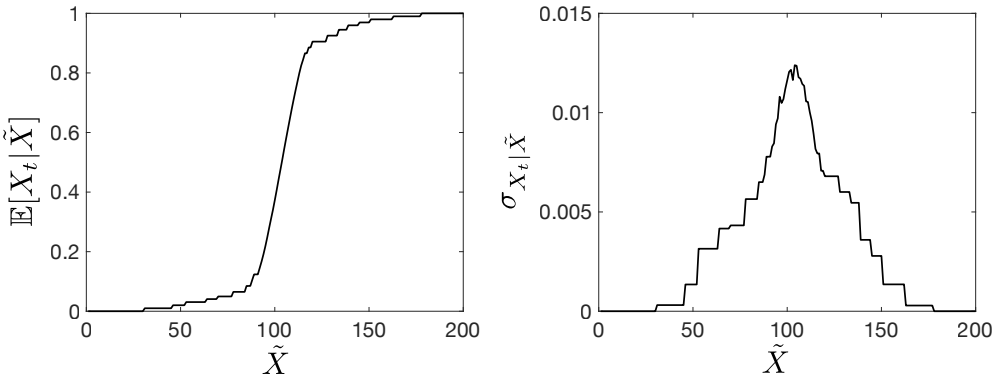

FIG. C.3: Mean (left) and variance (right) of the past allele frequency X_t conditioned on the (categorical) representation variable $\tilde{X}$ (left), for the information bottleneck solution of the Wright-Fisher dynamics with $m = 200$, $N = 100$, $N\mu = 0.2$, $Ns = 0.001$, $\beta = \infty$. The standard deviation is not constant: it is smaller where the prior probability of X_t is large.