

Optimal prediction with resource constraints using the information bottleneck

Vedant Sachdeva, Thierry Mora, Aleksandra Walczak, Stephanie Palmer

▶ **To cite this version:**

Vedant Sachdeva, Thierry Mora, Aleksandra Walczak, Stephanie Palmer. Optimal prediction with resource constraints using the information bottleneck. PLoS Computational Biology, 2021, 17 (3), pp.e1008743. 10.1371/journal.pcbi.1008743 . hal-03367851v2

HAL Id: hal-03367851

https://hal.science/hal-03367851v2

Submitted on 7 Oct 2021

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



Distributed under a Creative Commons Attribution 4.0 International License

PLOS COMPUTATIONAL BIOLOGY

RESEARCH ARTICLE

Optimal prediction with resource constraints using the information bottleneck

Vedant Sachdeva[1], Thierry Mora[2,3,4,5], Aleksandra M. Walczak[2,3,4,5], Stephanie E. Palmer[6,7]*

1 Graduate Program in Biophysical Sciences, University of Chicago, Chicago, Illinois, United States of America, **2** Laboratoire de physique de l'École normale supérieure, Centre National de la Recherche Scientifique, Paris, France, **3** Paris Sciences et Lettres University Paris, Paris, France, **4** Sorbonne Université Paris, Paris, France, **5** Université de Paris, Paris, France, **6** Department of Organismal Biology and Anatomy, University of Chicago, Chicago, Illinois, United States of America, **7** Department of Physics, University of Chicago, Chicago, Illinois, United States of America

* sepalmer@uchicago.edu

Check for updates

OPEN ACCESS

Citation: Sachdeva V, Mora T, Walczak AM, Palmer SE (2021) Optimal prediction with resource constraints using the information bottleneck. PLoS Comput Biol 17(3): e1008743. https://doi.org/10.1371/journal.pcbi.1008743

Editor: Aldo A Faisal, Imperial College London, UNITED KINGDOM

Received: April 29, 2020

Accepted: January 27, 2021

Published: March 8, 2021

Copyright: © 2021 Sachdeva et al. This is an open access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Data Availability Statement: These are theoretical results that can be numerically calculated, without data to share.

Funding: This work was supported in part by the US National Science Foundation, through the Center for the Physics of Biological Function (PHY–1734030) (SEP), and a CAREER award to SEP (1652617); by the National Institutes of Health BRAIN initiative (R01EB026943–01) (SEP); by a FACCTS grant from the France Chicago Center (SEP, AMW, TM, VS); by the Centre National de la Recherche Scientifique (SEP, AMW, VS); and by

Abstract

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

Author summary

From catching a ball to building immunity, we rely on the ability of biological systems to incorporate past observations to make predictions about the future state of the environment. However, the success of these predictions is limited by environmental parameters and encoding capacities of the predictors. We explore these trade-offs in three systems: simple intertial motion, more complex motion with long-tailed temporal correlations, and mutating viral strains. We show that the velocity and position of a moving object should not be equally well-remembered in the biological systems internal representation, and identify the flexible "best-compromise" representations that are not optimal but remain predictable in a wide range of parameters regimes. In the evolutionary context, we find that the optimal predictive representations are discrete, reminiscent of immune strategies that cover the space of potential viruses.

European Research Council Consolidator Grant (724208) (TM & AMW). The funders had no role in the study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: The authors have declared that no competing interests exist.

1 Introduction

How biological systems represent external stimuli is critical to their behavior. The efficient coding hypothesis, which states that neural systems extract as much information as possible from the external world, given basic encoding capacity constraints, has been successful in explaining some early sensory representations in the brain. Barlow suggested sensory circuits may reduce redundancy in the neural code and minimize metabolic costs for signal transmission [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 [5–8], 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 [9–11]. Molecular networks have also been shown to be predictive of future states, suggesting prediction may be one of the fundamental principles of biological computation [12, 13]. 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 [13–15]. 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 [14].

The information bottleneck framework, introduced by Tishby and colleagues [14, 16–18], allows us to define a 'relevance' variable in the encoded sensory stream. We take the relevant piece to be the future behavior of that input. Solving the bottleneck problem allows us to optimally estimate the future state of the external stimulus, given a certain amount of information retained about the past. In general, predicting the future coordinates of a system, $X_{t+\Delta t}$ reduces to knowing the precise historical coordinates of the stimulus X_t and an exact knowledge of the temporal correlations in the system. These rules and temporal correlations can be thought of as arising from two parts: a deterministic portion, described by a function of the previous coordinates, $\mathcal{H}(X_t)$, and the noise internal to the system, $\xi(t)$. Knowing the actual realization of the noise $\xi(t)$ reduces the prediction problem to simply integrating the stochastic equations of motion forward in time. If the exact realization of the noise if not known, we can still perform a probabalistic prediction by calculating the future form of the probability distribution of the variable X_t or its moments [19, 20]. The higher-order moments yield an estimate of X_t and the uncertainty in the estimate. However, biological systems cannot precisely know X_t due to inherently limited readout precision [21, 22], creating a trade-off between representing the past and predicting the future.

We briefly summarize the information bottleneck method to quantify this trade-off here, and provide a more thorough explanation of the case with Gaussian statistics (reproduced from [16]) in S1 Text. The method assumes that the input variable, in our case the signal $X_{t-t_0:t}$, which considers measurements between times $t - t_0$ and t. We will call the past. This can be used to make inferences about the relevance variable, in our case the future signal $X_{t+\Delta t:t+\Delta t+t_0}$, which considers measurements between times $t + \Delta t$ and $t + \Delta t + t_0$. We will call this the future. For convenience, in this introduction, we will take the past as a single point in time, X_t and the future as $X_{t+\Delta t}$. The resource constraints are introduced via a representation variable, $\tilde{X}$, which can have a varying amount of information about the input signal, X_t. This $\tilde{X}$, which

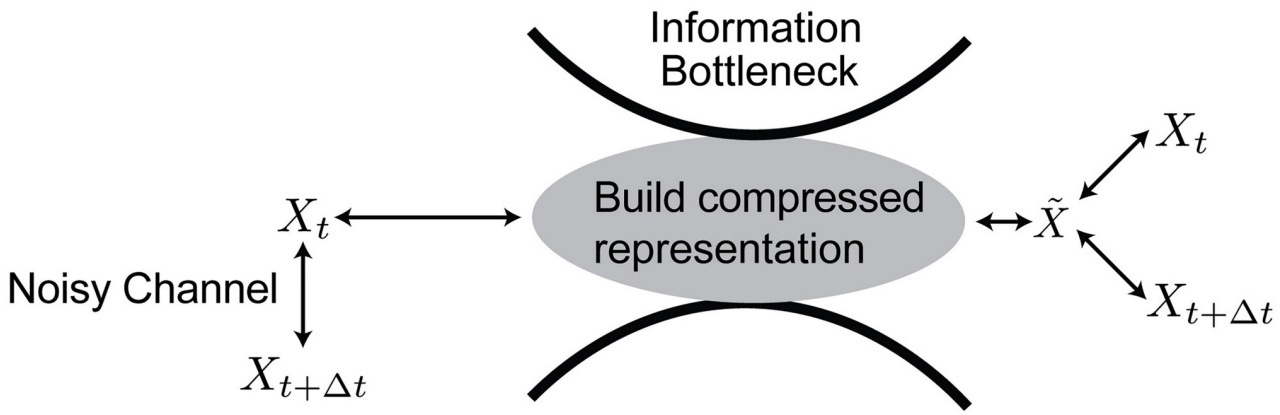


Fig 1. A schematic representation our predictive information bottleneck. On the left hand side, we have coordinates X_t evolving in time, subject to noise to give $X_{t+\Delta t}$. We construct a representation, $\tilde{X}$, that compresses the X_t (minimizes $I(X_t;\tilde{X})$) while retaining as much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

https://doi.org/10.1371/journal.pcbi.1008743.g001

has a dependence on the input, $\mathcal{P}(\tilde{X}|X_t)$, is constrained to be maximally informative of the future signal, subject to a constraint on $I(X_t;\tilde{X})$, the information it has about the past (Fig 1).

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

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

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

The causal relationship between X_t and $X_{t+\Delta t}$ results in a data processing inequality, $I(X_t;\tilde{X}) \geq I(X_{t+\Delta t};\tilde{X})$, meaning that the information generated about $X_{t+\Delta t}$ cannot exceed the amount encoded about X_t [23]. Additionally, the information about X_t that the representation can extract is bounded by the amount of information X_t, itself, contains about the $X_{t+\Delta t}$, $I(\tilde{X};X_{t+\Delta t}) \leq I(X_t;X_{t+\Delta t})$.

We use this framework to study how biological systems can optimally encode external stimuli for downstream decoding, but without any explicit constraints on or specification of that decoder. Here, we assume that the compressed representation variable has a one-time-step output and only has access to a fixed amount of historical information about the stimulus. Here, we assume that the compressed representation variable has a single 'present' time-step output and only has access to a fixed amount of historical information about the stimulus. This reflects, for example, the instantaneous neural output from a retinal ganglion cell population that is passed downstream to the cortex for further processing and readout. We start with a one-time-step past input and then extend this to a longer temporal window into the past. We begin by assuming a one-time-step past input and then later extend it to a more extended temporal window in the past. The optimal predictive encoder does in general favor some aspects of this past information (position information) over others (velocity information). A downstream decoder may be able to recover some of the lower priority information by combining measurements and predictions across time to reduce variance post hoc, but the gain in

precision comes at the cost of additional constraints on the size and complexity of the encoded representation variable. In addition, the gained information about the stimulus that was originally discarded may not provide significant predictive advantages. We do, however, provide a comparison between our information bottleneck framework and the results of a model that performs this kind of prediction combined with measurement and error estimates across time in Section D in S2 Text. There we demonstrate that for a given level of $I(X_t; \tilde{X})$, a Kalman filter achieves lower $I(\tilde{X}; X_{t+\Delta t})$. A question we do not explore here is how to, practically, read out the optimally encoded representation. It has been shown previously that simple perceptrons can read out predictive information from the retinal code [24], which makes biologically plausible readout possible and is a direction of future work.

We use information bottleneck to compute the optimal predictive encoding in two well-studied dynamical systems with ties to biological data: the stochastically-driven damped harmonic oscillator (SDDHO) and the Wright-Fisher model. We look at these two different systems to gain intuition about how different types of dynamics influence the ability of a finite and noisy system to make accurate predictions. We further consider two types of SDDHO processes to study the effects of noise correlations on prediction. Our exploration of the SDDHO system has a two-fold motivation: it is a physical system that describes motion that a visual system might need to process and predict to catch prey or evade predators. It is also the simplest possible continuous stochastic system whose full dynamics can be solved exactly. Previous studies used the SDDHO process to create moving bar stimuli and quantify retinal prediction [10, 24, 25]. Prediction of a time series with Markovian dynamics is not limited to physical motion, of course. The Wright-Fisher model [26] is a canonical model of evolution [27] which has been used to consider how the adaptive immune system predicts the future state of the pathogenic environment [12, 28]. Resource constraints also create trade-offs between representation precision and prediction in the immune system, and finding the general principles that connect prediction in these two contexts can reveal common principles across biological systems and scales.

The results of these information bottleneck calculations in these different dynamical contexts will reveal the form and content of optimally predictive features. These features are matched both to the input parameters and to the level of resource constraints that compress the input. Our results form expectations about what to find in biological systems when the internal representation can be measured (e.g. as in [10]), and the input statistics match the kinds of dynamics studied here. While our results will show what types of feature extraction are expected in systems predicting their inputs optimally, not all systems may be optimized for a broad range of input dynamics. In fact, we assume that natural selection favors encodings that confer just enough predictive capacity to support the organism's behavioral repertoire. That might mean flexibly predicting in many different environments either over an individual or group migratory lifespan. To help quantify the 'transferability' of any optimally predictive encoding scheme, we will develop a metric, Q, that tracks how well one representation performs under other input dynamics, where it might not be the absolute optimal, but still performs well. Of course, we only expect our maximally predictive encodings to match biological filters when the system has an intrinsic behavioral goal that requires prediction. There are computations that do not require prediction, and would presumably result from constraints that prioritize other types of information in the input.

2 Results

2.1 The stochastically driven damped harmonic oscillator

Previous work explored the ability of the retina to construct an optimally predictive internal representation of a dynamic stimulus. Palmer et al [10] 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 [10]. In order for optimal prediction to be possible, the retina should encode the position and velocity as dictated by the information bottleneck solution to the problem, for the retina's given level of compression of the visual input. In that study, the SDDHO was set near critical damping, and only one set of parameters in the model was shown to the retina. Inspired by this experiment, we explore the optimal predictive encoding schemes as a function of the parameters in the dynamics, and we describe the optimal solution across the entire parameter space of the model, over a wide range of desired prediction timescales.

We consider the dynamics of a mass m in a viscous medium attached to a spring receiving noisy velocity kicks generated by a temporally uncorrelated Gaussian process, as depicted in Fig 2A. The dynamics of this model were solved previously [29]. See Section A in S2 Text for details. Equations of motion are introduced in terms of physical variables $\bar{x}$, $\bar{v}$, and $\bar{t}$ (bars will be dropped later when referring to rescaled variables), which evolve according to

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

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 $D = \frac{k_B T}{\Gamma}$. We also introduce a dimensionless parameter, the damping coefficient, $\zeta = 1/(2\omega_0\,\tau)$. When $\zeta < 1$, the system is underdamped and the motion of the mass will be oscillatory. When $\zeta \geq 1$, the system is overdamped and the motion will be non-oscillatory. Additionally, the equipartition theorem tells us that $\langle\bar{x}(\bar{t})^2\rangle \equiv x_0^2 = k_B T/k = D/(\tau\omega_0^2)$. Putting this all together, we obtain

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

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

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

There are now just two parameters that govern a particular solution to our information bottleneck problem: ζ and Δt, the timescale on which we want to retain optimal information about the future. We define $X_t = (x(t), v(t))$ and $X_{t+\Delta t} = (x(t + \Delta t), v(t + \Delta t))$ and seek a representation, $\tilde{X}(\zeta, \Delta t)$, that can provide a maximum amount of information about $X_{t+\Delta t}$ for a fixed amount of information about X_t. By considering position and velocity, our system is Markovian, so extended temporal windows provide no additional information. If we were to ignore velocity in this model, estimates of the future would become suboptimal to the information bottleneck bound. We explore models where extended temporal windows are relevant in Section 2.2. To construct the information bottleneck solution in the case with Gaussian variables, we follow the construction given in [16]. 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.

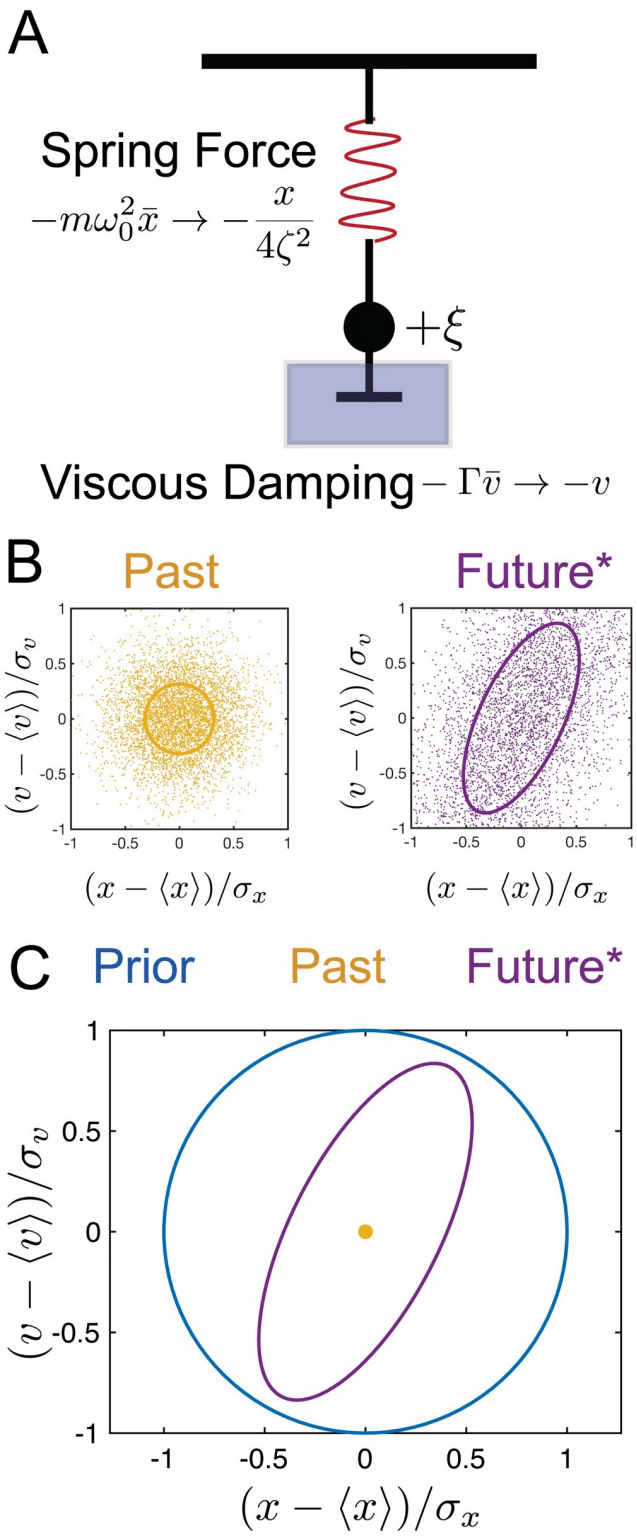


Fig 2. Schematic of the stochastically driven damped harmonic oscillator (SDDHO). (a) The SDDHO consists of a mass attached to a spring undergoing viscous damping and experiencing Gaussian thermal noise of magnitude. There are two parameters to be explored in this model: $\zeta = \frac{1}{2\omega_0\tau}$ and $\Delta t = \frac{\Delta t}{\tau}$. (b) $\zeta = \frac{1}{2}$, $\Delta t = 1$. Here, we show an example distribution of the history (yellow, left) and show its time evolution (purple, right). We take 5000 samples from the distribution, at random, and let these points evolve in time according to the SDDHO equation of motion. We visualize

the evolution of the distribution of points in time via an ellipse representing the $1 - \Sigma$ confidence region of the rescaled position and velocity. (c) We illustrate the limiting case of the information bottleneck method when $\beta \to \infty$. Representations of the past and how that constrains an estimate of the future position and velocity of the object can be compared to the prior be examining the relative size and shape of their respective ellipses. The blue circle represents the prior and its $1 - \Sigma$ confidence region. In yellow, we plot the inferred $1 - \Sigma$ confidence interval associated with the estimate of past, X_t, given by the encoding distribution when $\beta \to \infty$. In this limit, the distribution is reduced to a single point. In purple, we plot the $1 - \Sigma$ confidence region of $X_{t+\Delta t}$ given our knowledge of X_t. Precise knowledge of the past coordinates reduces the our uncertainty about the future position and velocity (as compared to the prior), as depicted by the smaller area of the purple ellipse.

https://doi.org/10.1371/journal.pcbi.1008743.g002

The optimal compressed representation is a noisy linear transform of X_t (see S1 Text) [16],

$$\tilde{X} = A_\beta X_t + \xi. \tag{5}$$

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

We calculate the optimal compression, $\tilde{X}$, and its predictive information (see Section B in S2 Text). The coordinates at time t and time $t + \Delta t$ in the SDDHO bottleneck problem are jointly Gaussian, which means that the optimal compression can be fully described by its first and second-order statistics. We generalize analytically the results that were numerically obtained in Ref. [10] and explore the full parameter space of this dynamical model and examine all predictive bottleneck solutions, including different desired prediction timescales.

We quantify the efficiency of the representation $\tilde{X}$ in terms of the variance of the following four probability distributions: the prior distribution, $\mathcal{P}(X_t)$, the distribution of X_t conditioned on the compression, $\mathcal{P}(X_t|\tilde{X})$, the distribution of $X_{t+\Delta t}$ conditioned on the compressed variable $\mathcal{P}(X_{t+\Delta t}|\tilde{X})$, and the distribution of $X_{t+\Delta t}$ conditioned on X_t $\mathcal{P}(X_{t+\Delta t}|X_t)$. We represent the uncertainty reduction, or the mutual information between these two variables, using two dimensional contour plots that depict the variances of the distributions in the $((x - \langle x \rangle)/\sigma_x, (v - \langle v \rangle)/\sigma_v)$ plane, where σ_x and σ_v are the standard deviations of the signal distribution $\mathcal{P}(X_t)$. We present example distributions of $\mathcal{P}(X_t|\tilde{X})$ and $\mathcal{P}(X_{t+\Delta t}|\tilde{X})$ in Fig 2B (left, right, respectively).

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

Limiting cases along the information bottleneck curve help build intuition about the optimal compression. If $\tilde{X}$ provides no information about the stimulus (e.g. $\beta = 0$), the variances of both of the conditional distributions match that of the prior distribution, $\mathcal{P}(X_t)$, which is depicted as a circle of radius 1 (blue circle in Fig 2C). However, if the encoding contains information about X_t, the variance of $\mathcal{P}(X_t|\tilde{X})$ will be reduced compared to the prior. The maximal amount of predictive information, which is reached when $\beta \to \infty$, can be visualized by examining the variance of $\mathcal{P}(X_{t+\Delta t}|X_t)$ (e.g. the purple contour in Fig 2C), which quantifies the correlations in X, itself, with no compression. Regardless of how precisely the current state of the

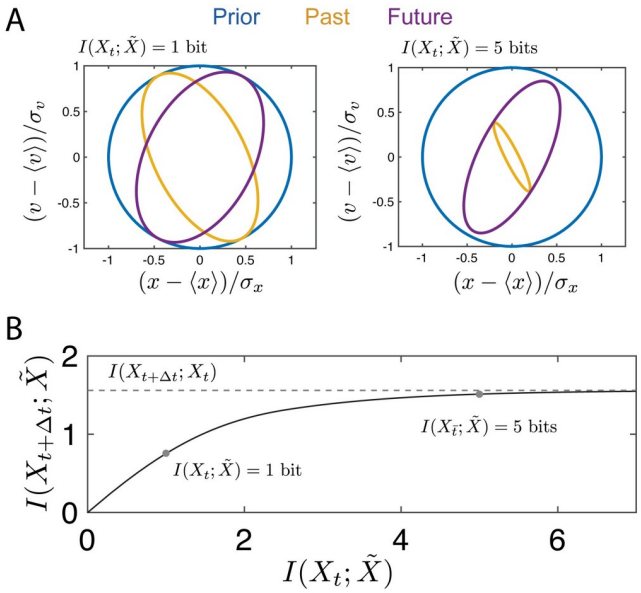


Fig 3. We consider the task of predicting the path of an SDDHO with $\zeta = \frac{1}{2}$ and $\Delta t = 1$. (a) (left) We encode the history of the stimulus, X_t, with a representation generated by the information bottleneck, $\tilde{X}$, that can store 1 bit of information. Knowledge of the coordinates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

https://doi.org/10.1371/journal.pcbi.1008743.g003

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 Fig 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 an information bottleneck curve for this damping coefficient and desired prediction timescale, as shown in Fig 3. Velocity uncertainty in the compressed representation is only reduced (i.e. predictive information that uses past velocity estimates is only useful) as we allow for less compression, as shown in Fig 3A (right). For both of the cases represented in Fig 3A, the illustrated encoding strategy yields a maximal amount of mutual information between the compressed representation, $\tilde{X}$, and the future for the given level of compression, as indicated by the red dots in Fig 3B.

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

$\omega^2 = \frac{1}{4\zeta^2} - \frac{1}{4}$ such that

$$w_1 = \begin{bmatrix} \omega \cot(\omega\Delta t) + \frac{|\csc(\omega\Delta t)|}{2\sqrt{2}\zeta}\sqrt{2 - \zeta^2 - \zeta^2\cos(2\omega\Delta t)} \\ 1 \end{bmatrix}. \quad (6)$$

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

$$\phi = \arctan\left(\left(\omega\cot(\omega\Delta t) + \frac{|\csc(\omega\Delta t)|}{2\sqrt{2}\zeta}\sqrt{2 - \zeta^2 - \zeta^2\cos(2\omega\Delta t)}\right)^{-1}\right). \quad (7)$$

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

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

$$\phi = \arctan\left(\frac{\Delta t}{1 + \omega^2}\right). \quad (8)$$

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

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

$$\phi = \arctan\left(\frac{2\sinh\left(\frac{\Delta t}{2}\right)}{\cosh\left(\frac{\Delta t}{2}\right) + \sqrt{\frac{1+\cosh(\Delta t)}{2}}}\right). \quad (9)$$

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

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

$$\phi = \arctan\left(\frac{2\zeta\sin\left(\frac{\Delta t}{2\zeta}\right)}{\cos\left(\frac{\Delta t}{2\zeta}\right) + \sqrt{2 - \zeta^2 - \zeta^2\cos\left(\frac{\Delta t}{\zeta}\right)}}\right). \quad (10)$$

We observe periodicity in the optimal encoding angle between position and velocity. This means that the optimal tradeoff between representing position or velocity depends on the timescale of prediction. However, the denominator never approaches 0, so the encoding scheme never favors pure velocity encoding. It returns to position-only encoding when $\Delta t/2\zeta = n\pi$.

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

$$1 - \lambda_1 = \exp(-\Delta t)\left(\frac{1}{4\omega^2\zeta^2} + \frac{\cos(2\omega\Delta t)}{4\omega^2} + \frac{|\sin(\omega\Delta t)|}{2\sqrt{2}\omega^2\zeta}\sqrt{2 - \zeta^2 - \zeta^2\cos(2\omega\Delta t)}\right). \quad (11)$$

For large Δt, it is clear that the slope will be constrained by the exponential term, and the information will fall as $\exp(-\Delta t)$ as we attempt to predict farther into the future. For small Δt,

however, we see that the slope goes as $1 - \Delta t^2$, and our predictive information decays more slowly.

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

$$I(X_t; X_{t+\Delta t}) = \frac{1}{2}\log(|\Sigma_{X_t}|) - \frac{1}{2}\log(|\Sigma_{X_t|X_{t+\Delta t}}|). \tag{12}$$

For large Δt, this expression becomes

$$I(X_t; X_{t+\Delta t}) \propto \exp(-\Delta t). \tag{13}$$

For small Δt,

$$I(X_t; X_{t+\Delta t}) \propto \Delta t - \frac{1}{2}\log(\Delta t). \tag{14}$$

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

2.2.1 Optimal representations in all parameter regimes for fixed $I(X_t; \tilde{X})$. We sweep over all possible parameter regimes of the SDDHO keeping $I(X_t; \tilde{X})$ fixed at 5 bits and find the optimal representation for a variety of timescales (Fig 4), keeping a fixed amount of information encoded about X_t for each realization of the stimulus and prediction. More information can be transmitted for shorter delays (Fig 4A, 4D and 4G) between the X_t and $X_{t+\Delta t}$ signal than for longer delays (Fig 4C, 4F and 4I). In addition, at shorter prediction timescales more information about X_t is needed to reach the upper bound, as more information can be gleaned about the future. In particular, for an overdamped SDDHO at short timescales (Fig 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 position information. This can be visualized by noting that the encoding ellipse remains on-axis and mostly compressed along the position dimension. For the underdamped case, in short time predictions (Fig 4G), a similar strategy is effective. However, for longer predictions (Fig 4H and 4I), inertial effects cause position at one time to be strongly predictive of future velocity and vice versa. As a result, the encoding distribution has to take advantage of these correlations to be optimally predictive. These effects can be observed in the rotation of the encoding ellipse, as it indicates that the uncertainty in position-velocity correlated directions are being reduced, at some cost to position and velocity encoding. The critically damped SDDHO (Fig 4D–4F) 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.

2.1.2 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 X_t, $\mathcal{P}(X_t|\tilde{X}_{\text{suboptimal}})$, and the future, $\mathcal{P}(X_{t+\Delta t}|\tilde{X}_{\text{suboptimal}})$. For a fixed amount of information about X_t, $I(X_t; \tilde{X}_{\text{optimal}}) = I(X_t, \tilde{X}_{\text{suboptimal}})$, we compare the predictive information in the optimal (Fig 5A) and the suboptimal representations (Fig 5B). We examine the choice of parameters in the stimulus dynamics for which encoding position alone is an optimal strategy. We note that encoding velocity with high certainty provides very little predictive power, indicating that encoding velocity and position is not

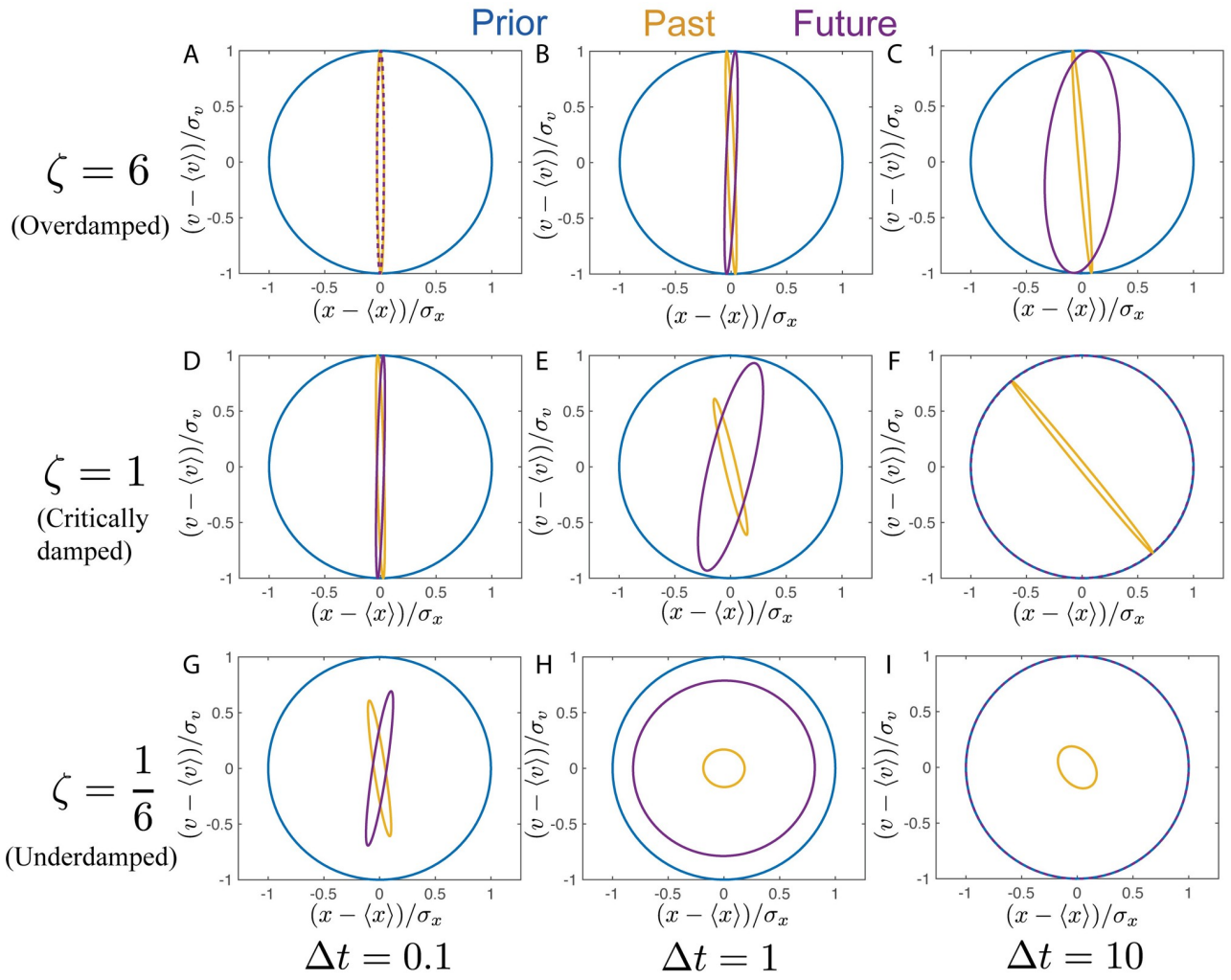


Fig 4. Possible behaviors associated for the SDDHO for a variety of timescales with a fixed $I(X_t; \tilde{X})$ of 5 bits. For an overdamped SDDHO, panel a-c, the optimal representation continues to encode mostly position information, as velocity is hard to predict. For the underdamped case, panels g-i,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

https://doi.org/10.1371/journal.pcbi.1008743.g004

equally important, even for equal compression levels. While the nature of the suboptimal and optimal representations depend on the input dynamics, we see that the encoding schemes discovered by the information bottleneck are, indeed, optimally predictive.

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

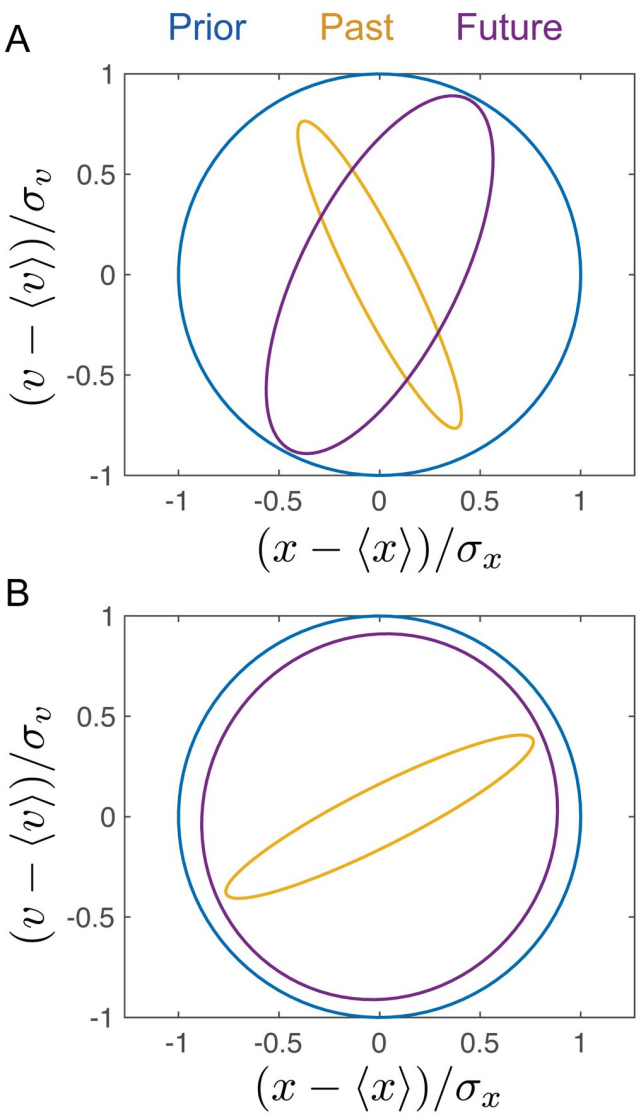


Fig 5. Example of a sub-optimal compression. An optimally predictive, compressed representation, in panel (a) compared to a suboptimal representation, in panel (b) for a prediction at $\Delta t = 1$ in the future, within the underdamped regime ($\zeta = 1/2$). We fix the mutual information between the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

https://doi.org/10.1371/journal.pcbi.1008743.g005

different timescales to avoid predators, and predators may have a different dynamical regime [25, 30]. 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 (ζ, Δt) and ask how efficiently it captures predictive information for a different parameter regime, (ζ', $\Delta t'$).

We identify these global strategies by finding the optimal encoder for a stimulus with parameters (ζ, Δt) 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 (ζ', $\Delta t'$)

and compute the amount of predictive information the previous representation yields for this stimulus. We call this the transferred predictive information $I^{\text{future}}_{\text{transfer}}((\zeta, \Delta t), I_{\text{past}} \to (\zeta', \Delta t'))$. We note that $I^{\text{future}}_{\text{transfer}}((\zeta, \Delta t), I_{\text{past}} \to (\zeta', \Delta t'))$ may sometimes be larger than $I^{\text{future}}_{\text{optimal}}((\zeta, \Delta t), I_{\text{past}})$, because changing $(\zeta, \Delta t)$ may increase both I_{past} and I_{future} (see e.g. Fig 6A).

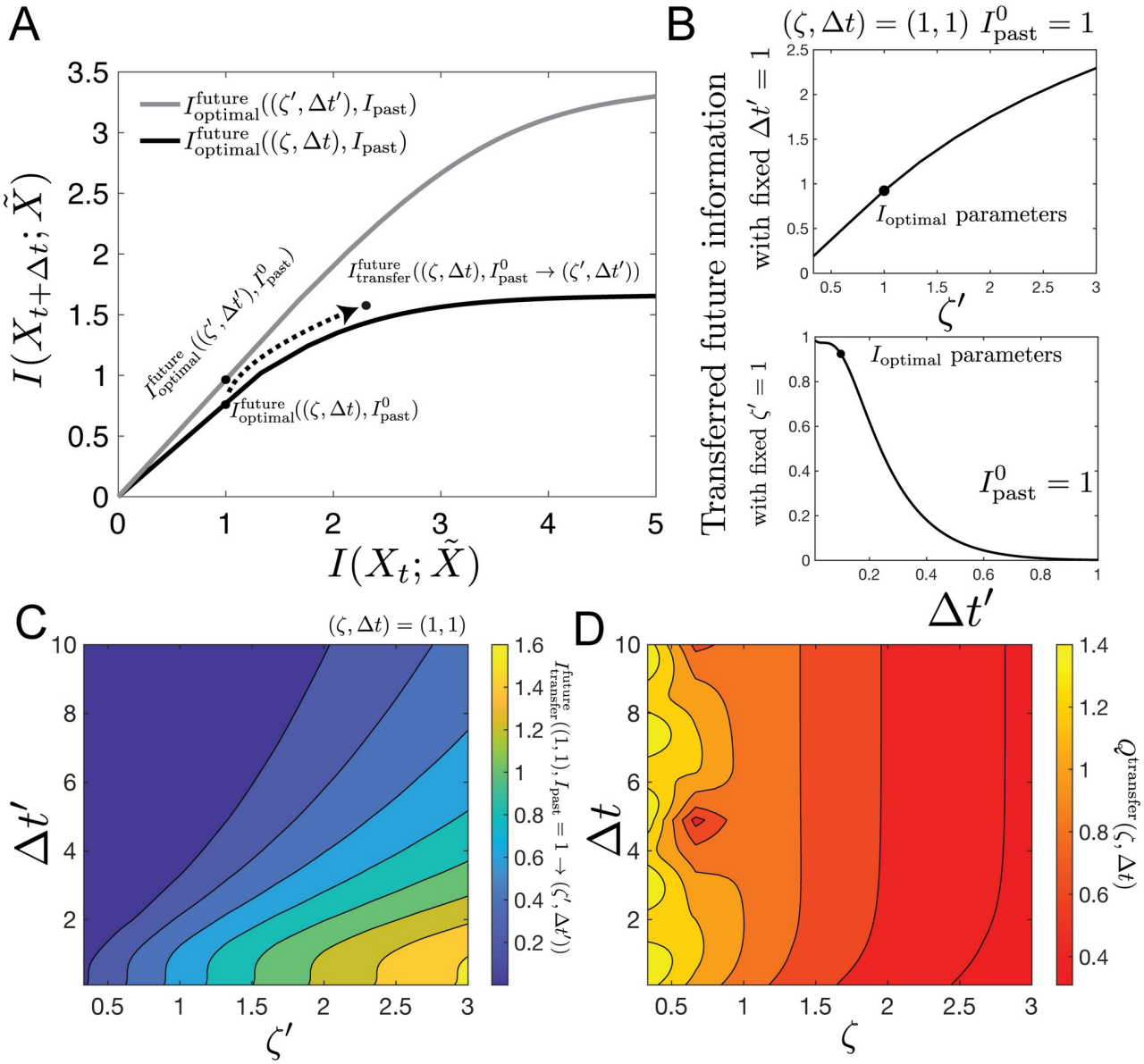


Fig 6. Representations learned on underdamped systems can be transferred to other types of motion, while representations learned on overdamped systems cannot be easily transferred. (a) Here, we consider the information bottleneck bound curve (black) for a stimulus with underlying parameters, $(\zeta,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

https://doi.org/10.1371/journal.pcbi.1008743.g006

For every fixed $(\zeta, \Delta t)$ and I_{past}, we can take the optimal $\tilde{X}$ and transfer it to a wide range of new ζ''s and timescales, $\Delta t'$. For a particular example $(\zeta, \Delta t)$, this is shown in Fig 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 Δt's. What we get, then, is a mapping of $I^{\text{future}}_{\text{transfer}}$ for this representation that was optimized for one particular $(\zeta, \Delta t)$ pair across all new $(\zeta', \Delta t')$'s. This is shown in Fig 6C, (Fig 6B are just two slices through this surface). This surface gives a qualitative picture the transferability of this particular representation.

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

2.2 History-dependent Gaussian stimuli

In the above analysis, we considered stimuli with temporal correlations that fall off exponentially. However, natural scenes, such as leaves blowing in the wind or bees moving in their hives, are shown to have heavy-tailed statistics [25, 31, 32]. To extend our results to such stimuli, we consider prediction where the statistics of the motion model may feature long-ranged temporal correlations and by increasing the dimensionality of the input and output to the information bottleneck, we demonstrate that the information bottleneck continues to provide useful predictive encoding schemes for such stimuli. We show this through the use of the Generalized Langevin equation [33–35]:

$$\frac{dv}{dt} = -\int_0^t \frac{\gamma v}{|t-t'|^z}\, dt - \omega_0^2 x + \xi(t) \tag{15}$$

$$\frac{dx}{dt} = v \tag{16}$$

Here, we have returned to unscaled definitions of v, and t. The damping force has a power-law kernel. In order for the system to obey the fluctuation-dissipation theorem, we note that $\langle \xi(t) \rangle = 0$, and $\langle \xi(t')\xi(t) \rangle \propto \frac{1}{|t-t'|^z}$. 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 S1 Text, we compute the optimal representation and determine how informative the past is about the future. The objective function for this extended information bottleneck problem is,

$$\mathcal{L} = \min_{\mathcal{P}(\tilde{X}|X_{t-t_0:t})} I(X_{t-t_0:t}; \tilde{X}) - \beta I(X_{t+\Delta t:t+\Delta t+t_0}; \tilde{X}). \tag{17}$$

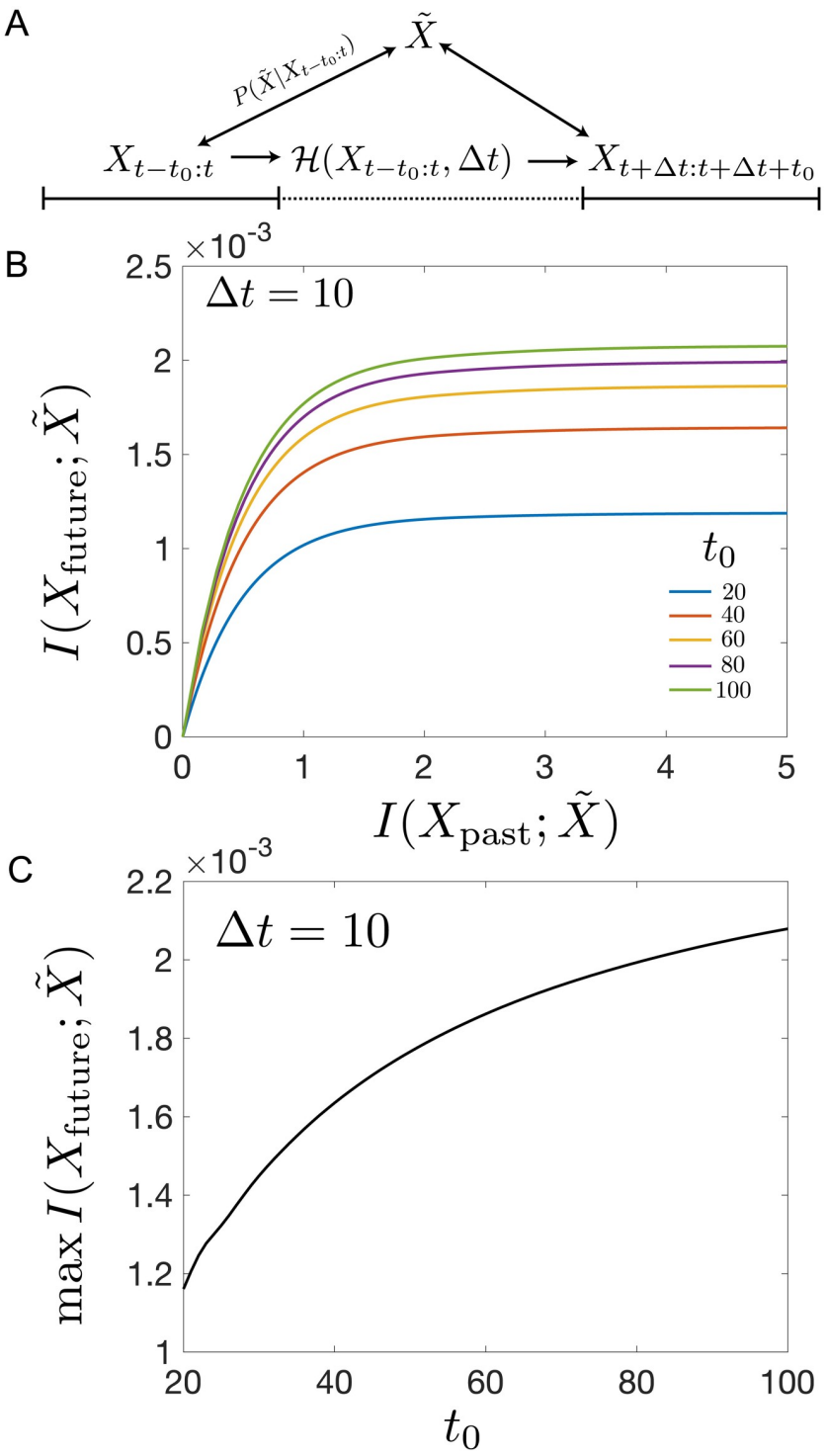


Fig 7. The ability of the information bottleneck Method to predict history-dependent stimuli. (a) The prediction problem, using an extended history and a future. This problem is largely similar to the one set up for the SDDHO but the past and the future are larger composites of observations within a window of time $t-t_0$: t, expressed as X_{past} for the past and $t + \Delta t$: $t + \Delta t + t_0$, expressed as X_{future} for the future. (b) Predictive information $I(X_{t+\Delta t:t+\Delta t+t_0}, \tilde{X})$ with lag Δt. (c) The maximum available predictive information saturates as a function of the historical information used t_0.

https://doi.org/10.1371/journal.pcbi.1008743.g007

We demonstrate the impacts of the discretization of time in S2. The information bottleneck curves show more predictive information as the prediction process uses more past information (larger t_0 in Fig 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.

2.3 Evolutionary dynamics

Exploiting temporal correlations to make predictions is not limited to vision. Another aspect of the prediction problem appears in the adaptive immune system, where temporal correlations in pathogen evolution may be exploited to help an organism build and maintain immunity in a changing environment. Exploiting these correlations can be done at a population level, in terms of vaccine design [36–39], and has been postulated as a means for the immune system to adapt to future threats [12, 40]. Here, we present efficient predictive coding strategies for the Wright-Fisher model, which is commonly used to describe viral evolution [41]. In contrast to the two models studied so far, Wright-Fisher dynamics are not Gaussian, though they are still Markovian. This implies that predictive information can reside in higher-order moments of the joint distribution, thus the optimal compressed representation variable can no longer be Gaussian. The Wright-Fisher model allows us to explore how the results obtained in the previous sections generalize to non-Gaussian statistics of the past and future distributions. To make this computationally tractable, we will take the representation variable to be discrete, though later allow its cardinality to be large to approximate the continuous solution. There exist methods to approximate continuous compressed representations directly [42–44], though we do not use those here.

Wright-Fisher models of evolution assume a constant population size of N. We consider a single mutating site with each individual in the population having either a wild-type or a mutant allele at this site. The allele choice of subsequent generations depends on the frequency of the mutant allele in the ancestral generation at time t, X_t, the selection pressure on the mutant allele, s, and the mutation rate from the wild-type to the mutant allele and back, μ, as depicted as Fig 8A. For large enough N, the update rule of the allele frequencies is given through the diffusion approximation interpreted with the Ito convention [45]:

$$\frac{dX_t}{dt} = sX_t(1 - X_t) + \mu(1 - 2X_t) + \sqrt{X_t(1 - X_t)/N}\eta(t), \tag{18}$$

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

For this model, defining the representation $\tilde{X}$ as a noisy linear transformation of X_t, the allele frequency at time t, as we did for the Gaussian case in S1 Text. Eq 1 does not capture all of the dependences between the past and future allele frequencies, because correlations exist beyond second order. This arises because of the non-linear form of Eq 18. Instead, we determine the mapping of X_t to $\tilde{X}$ numerically using the Blahut-Arimoto algorithm [46, 47]. In general, for a discrete representation variable $\tilde{X}$, the true cardinality of $\tilde{X}$ is unknown for a given

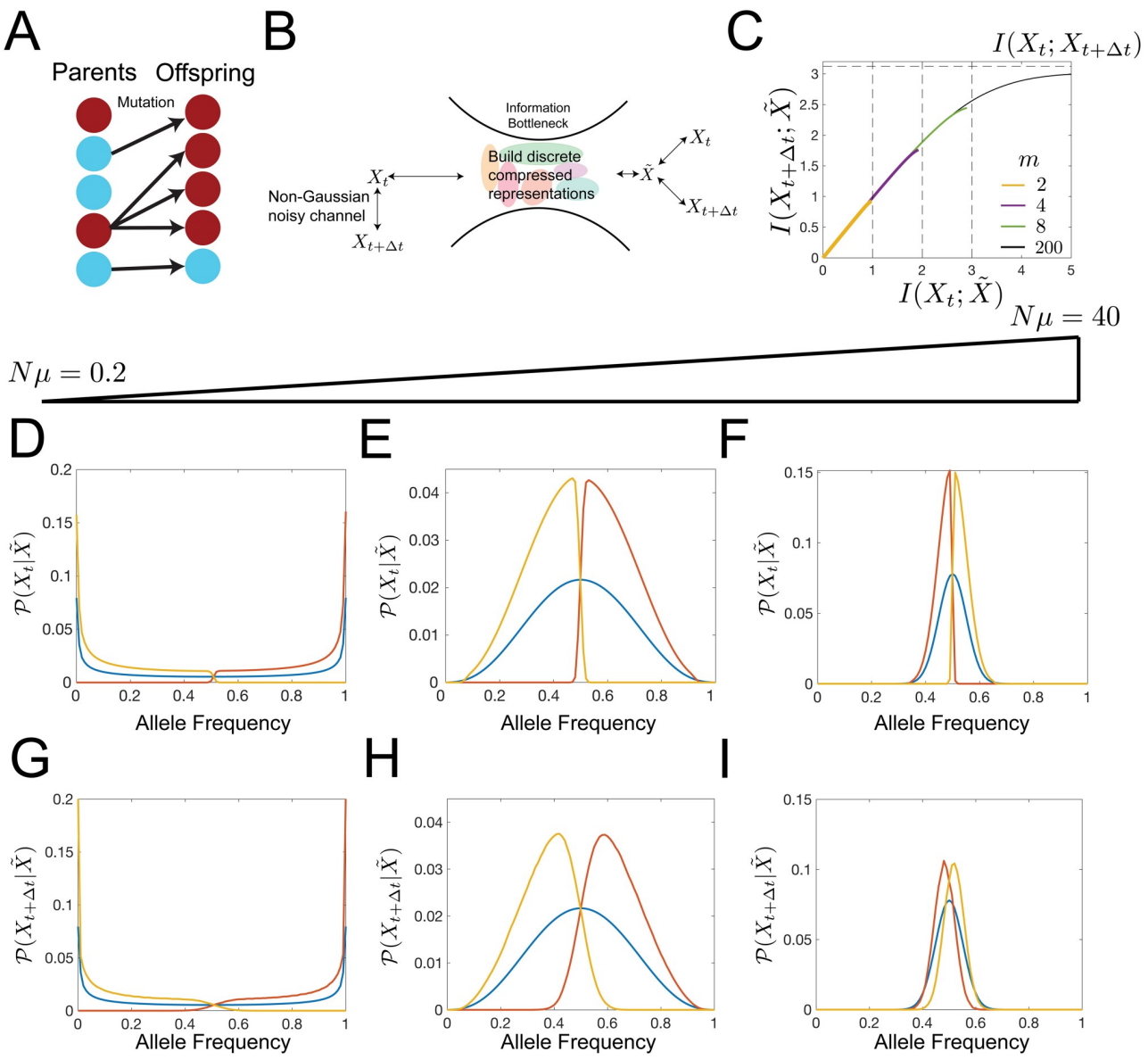


Fig 8. The information bottleneck solution for a Wright Fisher process. (a) The Wright-Fisher model of evolution can be visualized as a population of N parents giving rise to a population of N offspring. Genotypes of the offspring are selected as a function of the parents' generation genotypes subject to mutation rates, μ, and selective pressures s. (b) Information bottleneck schematic with a discrete (rather than continuous) representation variable, $\tilde{X}$. (c) Predictive information as a function of compression level. Predictive information increases with the cardinality, m, of the representation variable. The amount of predictive information is limited by $\log(m)$ (vertical dashed lines) for small m, and the mutual information between allele frequencies at time $t + \Delta t$ and time t, $I(X_{t+\Delta t};X_t)$ (horizontal dashed line), for large m. Bifurcations occur in the amount of predictive information. For small $I(X_t;\tilde{X})$, the encoding strategies for different m are degenerate and the degeneracy is lifted as $I(X_t;\tilde{X})$ increases, with large m schemes accessing higher $I(X_t;\tilde{X})$ ranges. Parameters: $N = 100$, $N\mu = 0.2$, $N\mu = 0.2$, $Ns = 0.001$, $\Delta t = 1$. (d-i) We explore information bottleneck solutions to Wright-Fisher dynamics under the condition that the cardinality of $\tilde{X}$, m, is 2 and take β to be large enough that $I(X_t;\tilde{X}) \approx 1$, $\beta \approx 4$. Parameters: $N = 100$, $Ns = 0.001$, $\Delta t = 1$, and $N\mu = 0.2$, $N\mu = 2$, and $N\mu = 40$ (from left to right). (d-f) In blue, we plot the steady state distribution. In yellow and red, we show the inferred historical distribution of alleles based on the observed value of $\tilde{X}$. Note that each distribution is corresponds to roughly non-overlapping portions of allele frequency space. (g-i) Predicted distribution of alleles based on the value of $\tilde{X}$. We observe that as mutation rate increases, the timescale of relaxation to steady state decreases, so historical information is less useful and the predictions becomes more degenerate with the steady state distribution.

https://doi.org/10.1371/journal.pcbi.1008743.g008

β. Our approach is to first fix the cardinality of $\tilde{X}$ to a given value m (Fig 8C) and compute the information curve for the given m by sweeping over β. We then repeat this for larger values of m. We note that for small β, the solutions for different values of m are degenerate, while at higher values of β, bifurcations emerge between encoding schemes for the solutions with cardinality m and $m - 1$. This is because the true cardinality of the optimal solution undergoes transitions to higher and higher values as β increases [14]. The discreteness of $\tilde{X}$ results in each realization of the representation tiling a distinct part of frequency space. This encoding scheme can be thought of a different types of immune defenses: innate, adaptive, and different lymphocyte phenotypes acting at different stages or for different types of immune responses [48]. Accordingly, m would correspond to the number of distinct cell types mobilized against pathogens of various frequencies. The concept of discrete tiling of space is also analogous to ideas of immune coverage, whereby a finite number of distinct antigen receptors cover the entire "shape space" of possible antigens [49]. However, to make this analogy more precise would require to study an effective theory of phenotypic evolution [50].

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,

$$P_s(X) \propto [X(1-X)]^{N\mu-1} e^{NsX} \tag{19}$$

(blue line in Fig 8D) is peaked around the frequency boundaries, indicating that at long times, an allele either fixes or goes extinct. In this case, one value of the representation variable corresponds to the range of high allele frequencies and the other corresponds to low allele frequencies (Fig 8D, yellow and red lines). These encoding schemes can be used to make predictions, whether it be by an observer or the immune system, via determining the future probability distribution of the alleles conditioned on the value of the representation variables, $\mathcal{P}(X_{t+\Delta t}|\tilde{X})$. We present these predictions in Fig 8G. The predictive information conferred by the representation variable is limited by the information it has about X_t as in the Gaussian case (Fig 8C).

For larger mutation rates, the steady state distribution becomes centered around the equal probability of observing either one of the two alleles, but the two representation variables still cover the frequency domain in way that minimizes overlap (Fig 8E and 8F). We observe a sharp drop in $P(X_t|\tilde{X})$ at the boundary between the two representations. The future distribution of allele frequencies in this region (Fig 8H and 8I), however, displays large overlap. The degree of this overlap increases as the mutation rate gets larger, suggesting prediction is harder in the strong mutation limit. The optimal encoding of the distribution of X_t biases the representation variable towards frequency space regions with larger steady state probability mass.

In Fig 9, we explore the consequence of transferring a mapping, $\mathcal{P}(\tilde{X}|X_t)$, from a high mutation model to a low mutation model and vice versa. We observe that the weak mutation representation is more transferrable than the strong mutation representation. One reason for this is that the strong mutation limit provides little predictive information, as seen in Fig 10A. In addition, high mutation representations focus on $X = 1/2$, while the population more frequently occupies allele frequencies near 0 and 1 in other regimes. Comparatively, representations learned on weak mutation models can provide predictive information, because they cover more evenly the spectrum of allele frequencies.

We can extend the observations in Fig 8 to see how the predictive information depends on the strength of the selection and mutation rates (Fig 10A and 10C). Prediction is easiest in the weak mutation and selection limit, as population genotype change occur slowly and the steady state distribution is localized in one regime of the frequency domain. For evolutionary forces acting on faster timescales, prediction becomes harder since the relaxation to the steady state is fast. Although the mutation result might be expected, the loss of predictive information in the

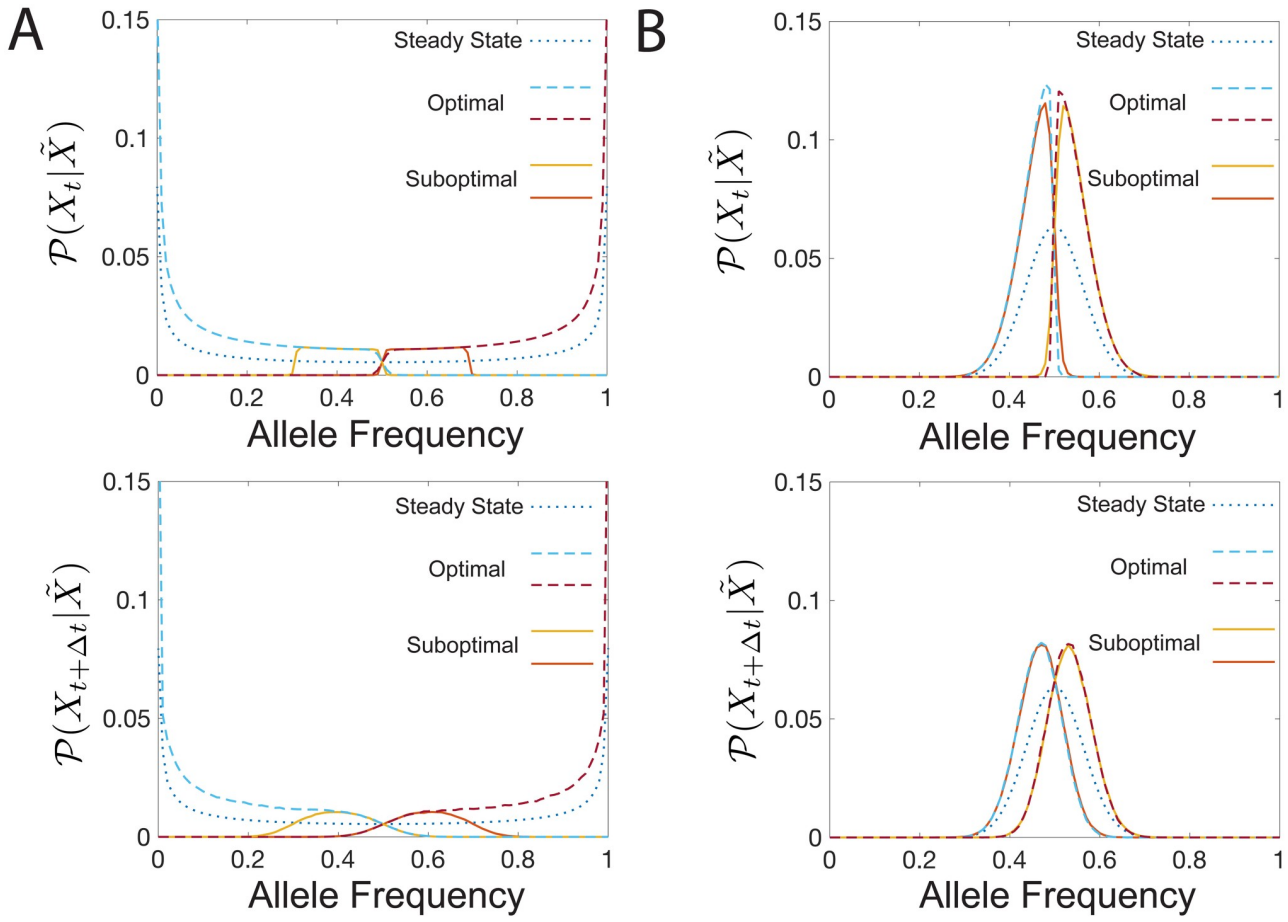


Fig 9. Transferability of prediction schemes in Wright-Fisher dynamics. We transfer a mapping, $\mathcal{P}(\tilde{X}|X_t)$, trained on one set of parameters and apply it to another. We consider transfers between two choices of mutability, $N\mu_1 = 0.2$ (low) and $N\mu_2 = 20$ (high), with $N = 100$, $Ns = 0.001$, $\Delta t = 1$. 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.

https://doi.org/10.1371/journal.pcbi.1008743.g009

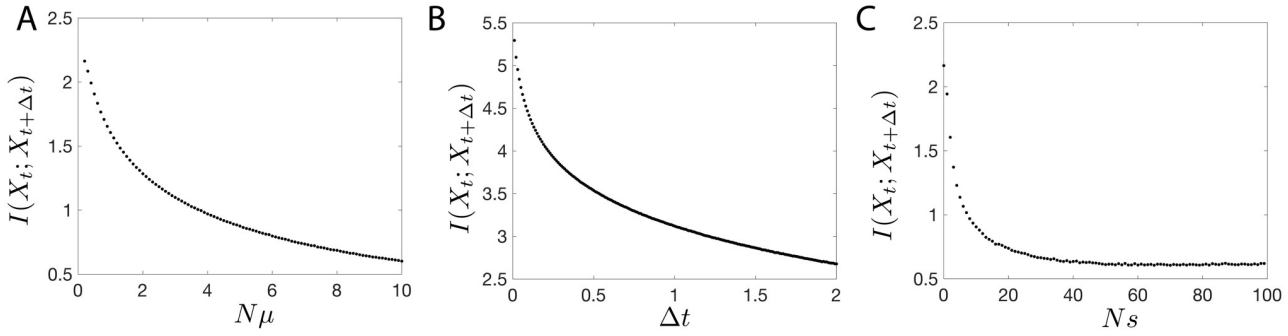


Fig 10. Amount of predictive information in the Wright Fisher dynamics as a function of model parameters. (a-c), Value of the asymptote of the information bottleneck curve, $I(X_t;X_{t+\Delta t})$ with: (a) $N = 100$, $Ns = 0.001$, $\Delta t = 1$ as a function of μ; (b) $N = 100$, $N\mu = 0.2$, $Ns = 0.001$ as a function of Δt; and (c) $N = 100$, $N\mu = 0.2$, and $\Delta t = 1$ as a function of s.

https://doi.org/10.1371/journal.pcbi.1008743.g010

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 10B).

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 the allele frequency at time t. However, at high β values the amount of information the representation variable can hold saturates, and the predictive information reaches a maximum value (1 bit for the $m = 2$ yellow line in Fig 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 8C)). Comparing the $m = 2$ and $m = 3$ representations for maximum values of β for each of them (Fig 11A and 11B), shows that larger numbers of representations tile allele frequency space more finely, allowing for more precise encodings of the past and future distributions. The maximum amount of information about the past goes as $\log(m)$ (Fig 8C). The predictive information curves for different m values are the same, until the branching point $\lesssim \log(m)$ for each m (Fig 8C).

We analyze the nature of this branching by taking $m \gg 1$, $m = 200$ (Fig 11C and 11D). At small β (and corresponding small $I(X_t;\tilde{X})$) the optimal encoding scheme is the same if we had imposed a small m (Fig 11C), with additional degenerate representations (S3 Fig). By increasing β (and $I(X_t;\tilde{X})$), the degeneracy is lifted and additional representation cover non-overlapping regimes of allele frequency space. This demonstrates the existence of a critical β for each predictive coding scheme, above which m needs to be increased to extract more predictive information and below which additional values of the representation variable encode redundant portions of allele frequency space. While we do not estimate the critical β, approaches to estimating them are presented in [51, 52].

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 and S4 Fig). 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.

3 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.

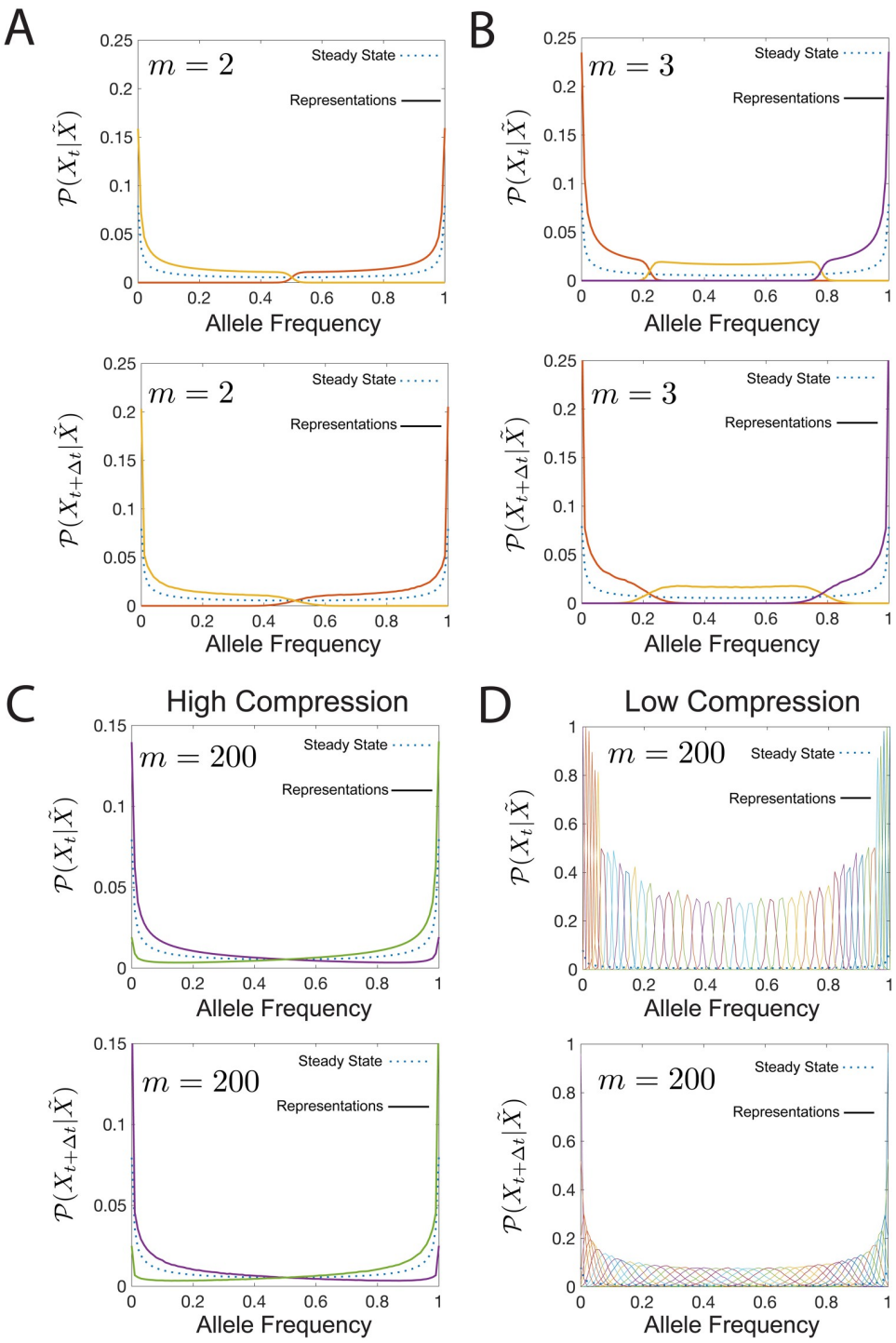


Fig 11. Encoding schemes with $m > 2$ representation variables. The steady state is plotted as a dotted line and the representation for each realization of the value of $\tilde{X}$ are plotted as solid lines. The representations which carry maximum predictive information for (a) $m = 2$ at $I(X_t;\tilde{X}) \approx \log(m) = 1$ bit, and (b) $m = 3$ at $I(X_t;\tilde{X}) \approx \log(m) \approx 1.5$ bits. The optimal representations at large m tile space more finely and have higher predictive information. The optimal representations for $m = 200$ at fixed $\beta = 1.01$ ($I(X_t;\tilde{X}) = 0.28$, $I(X_{t+\Delta t};\tilde{X}) = 0.27$) (c) and $\beta = 20$ ($I(X_t;\tilde{X}) = 2.77$, $I(X_{t+\Delta t};\tilde{X}) = 2.34$). (d) At low $I(X_t;\tilde{X})$, many of the representations are redundant and do not confer more predictive information than the $m = 2$ scheme. A more explicit comparison is given in S3 Fig. At high $I(X_t;\tilde{X})$, the degeneracy is lifted. All computations done at $N = 100$, $N\mu = 0.2$, $Ns = 0.001$, $\Delta t = 1$.

https://doi.org/10.1371/journal.pcbi.1008743.g011

The key principle is that not all input dimensions are equally relevant for prediction; information encoding systems must be able to parse which dimensions are relevant when coding capacity is small relative to the available predictive information. Hence, the biological (or engineered) system must navigate a tradeoff between reducing the overall uncertainty in its prediction while only being able to make measurements with some fixed uncertainty.

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

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

Superficially, our framework may seem similar to a Kalman filter [53]. 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 [54]. In this framework, the Kalman filters seek to maximize information by minimizing the variance in estimating the true coordinates of an external input. The estimate is, then, a prediction of the next time step, and is iteratively updated. Our information bottleneck approach extracts past information, but explicitly includes another constraint: resource limitations. The tuning of I_{past} is the main difference between our approach and a Kalman filter. Another major difference is that we do not assume the underlying encoder has any explicit representation of the 'physics' of the input. There is no internal model of the input stimulus, apart from our probabilistic mapping from the input to our compressed representation of that input. A biological system could have such an internal model, but that would add significant coding costs that would have to be treated by another term in our framework to draw a precise equivalence between the approaches. We show in the S1 Fig that the Kalman filter approach is not as efficient, in general, as the predictive information bottleneck approach that we present here.

Our results on systems with Wright-Fisher input dynamics reveal that discrete representations that tile input space are optimally predictive encoders. Although we impose discrete internal representations, their non-overlapping character remains even it the limit of a large number of representations. These kinds of solutions are reminiscent of the Laughlin solution for information maximization of input and output in the visual system given a nonlinear noisy channel [2], in which the input space is covered proportionally to the steady state distribution

at a given frequency, in chunks given by the size of the noise in the system. Tiling solutions have also been described when optimizing information in gene regulatory networks with non-linear input-output relations, when one input regulates many gene outputs [55]. 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 [56, 57].

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 [28]. 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 [58, 59], 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 8E), 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 [15, 42, 60]. In these deep networks, variational methods approximate the information quantities in the bottleneck, and have proven their practical utility in many machine learning contexts. These approaches do not always provide intuition about how the networks achieve this performance and what the information bottleneck approach creates in the hidden encoding layers. Here, we have worked through a set of analytically tractable examples, laying the groundwork for building intuition about the structure of information bottleneck solutions and their generalizations in more complex problems.

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

By providing a set of dissected solutions to the predictive information bottleneck problem, we hope to show that not only is the approach feasible for biological encoding questions, it

also illuminates connections between seemingly disparate systems (such as visual processing and the immune system). In these systems the overarching goal is the same, but the microscopic implementation might be very different. Commonalities in the optimally predictive solutions as well as the most generalizable ones can provide clues about how to best design experimental probes of this behavior, at both the molecular and cellular level or in networks.

Methods

Computational methods are as described in the Results and in S1–S3 Text.

Supporting information

S1 Text. Computing the optimal representation for jointly Gaussian past-future distributions. We present the results of [16], which is a a derivation for the solution to the information bottleneck in the limit of jointly Gaussian variables. This formalism is used throughout the text to analytically produce the results presented.
(PDF)

S2 Text. The stochastically driven damped harmonic oscillator. We provide full derivations for our results involving the harmonic oscillator, including extensions to generalized frictional kernels. We also provide some comparison to another popular scheme, the Kalman filter.
(PDF)

S3 Text. Wright-Fisher dynamics. We provide some detail about the parameters in our simulation for Wright-Fisher dynamics and a short derivation for the maximum amount of encoded information for a given m.
(PDF)

S1 Fig. Kalman filtering schemes are not efficient coders for a given channel capacity. We compare the amount of information conferred about the future for a given encoding level and find that Kalman Filter-based approaches do not maximize the amount of predictive information conferred, suggesting they are not efficient predictive coding schemes.
(TIF)

S2 Fig. We plot the information curve for $\Delta t = 10$, $t_0 = 20$ for different values of dt. We note that there are diminishing returns for increasingly small dt. However, we cannot make dt arbitrarily small, as this introduces numerical errors.
(TIF)

S3 Fig. 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.
(TIF)

S4 Fig. 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.
(TIF)

Acknowledgments

We acknowledge Ilya Nemenman for insight on the structure of information bottleneck solutions. We acknowledge Wim Van Drogen for helping compare information bottleneck solutions to the Kalman filter. We acknowledge Jared Salisbury for early work leading to the development of this project.

Author Contributions

Conceptualization: Vedant Sachdeva, Thierry Mora, Aleksandra M. Walczak, Stephanie E. Palmer.

Formal analysis: Vedant Sachdeva.

Funding acquisition: Thierry Mora, Aleksandra M. Walczak, Stephanie E. Palmer.

Investigation: Vedant Sachdeva.

Methodology: Vedant Sachdeva, Thierry Mora, Aleksandra M. Walczak, Stephanie E. Palmer.

Software: Vedant Sachdeva.

Supervision: Thierry Mora, Aleksandra M. Walczak, Stephanie E. Palmer.

Visualization: Vedant Sachdeva.

Writing – original draft: Vedant Sachdeva, Thierry Mora, Aleksandra M. Walczak, Stephanie E. Palmer.

Writing – review & editing: Vedant Sachdeva, Thierry Mora, Aleksandra M. Walczak, Stephanie E. Palmer.

References

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. https://doi.org/10.1515/znc-1981-9-1040
3. de Ruyter van Steveninck RR, Laughlin SB. The rate of information transfer at graded-potential synapses. Nature. 1996; 379(6566):642–645. 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. https://doi.org/10.1038/381607a0
5. Bialek W, Nemenman I, Tishby N. Predictability, Complexity, and Learning. Neural Computation. 2001; 13(11):2409–2463. https://doi.org/10.1162/089976601753195969
6. Lee TS, Mumford D. Hierarchical Bayesian inference in the visual cortex. J Opt Soc Am A. 2003; 20 (7):1434–1448. https://doi.org/10.1364/JOSAA.20.001434
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. https://doi.org/10.1162/neco.1997.9.4.721
8. 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.
9. 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. https://doi.org/10.1038/4580
10. 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. https://doi.org/10.1073/pnas.1506855112
11. Zelano C, Mohanty A, Gottfried J. Olfactory Predictive Codes and Stimulus Templates in Piriform Cortex. Neuron. 2011; 72(1):178–187. https://doi.org/10.1016/j.neuron.2011.08.010 PMID: 21982378

12. 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. https://doi.org/10.1073/pnas.1812810116
13. 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. https://doi.org/10.1038/s41467-019-11405-4
14. Tishby N, Pereira FC, Bialek W. The Information Bottleneck Method; 1999. p. 368–377.
15. Alemi AA. Variational Predictive Information Bottleneck; 2019.
16. 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.
17. Friedman, N, Mosenzon, O, Slonim, N, Tishby, N. Multivariate Information Bottleneck arXiv. 2013; 1301.2270.
18. Slonim, N. The Information Bottleneck: Theory and Applications. 2002.
19. 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.
20. 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.
21. Berg HC, Purcell EM. Physics of chemoreception. Biophysical Journal. 1977; 20(2):193–219. https://doi.org/10.1016/S0006-3495(77)85544-6 PMID: 911982
22. Bialek W. Biophysics: Searching for Principles. Princeton University Press; 2012. Available from: https://books.google.com/books?id=5In_FKA2rmUC.
23. Beaudry NJ, Renner R. An intuitive proof of the data processing inequality; 2011.
24. 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. https://doi.org/10.1073/pnas.1710779115
25. Salisbury JM, Palmer SE. Optimal Prediction in the Retina and Natural Motion Statistics. Journal of Statistical Physics. 2016; 162(5):1309–1323. https://doi.org/10.1007/s10955-015-1439-y
26. 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; 31(12):382–389. https://doi.org/10.1073/pnas.31.12.382
27. Tataru P, Simonsen M, Bataillon T, Hobolth A. Statistical Inference in the Wright-Fisher Model Using Allele Frequency Data. Systematic biology. 2017; 66(1):e30–e46.
28. 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. https://doi.org/10.1073/pnas.1421827112
29. 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; 83:041103. https://doi.org/10.1103/PhysRevE.83.041103
30. Ruderman DL, Bialek W. Statistics of natural images: Scaling in the woods. Physical Review Letters. 1994; 73:814–817 https://doi.org/10.1103/PhysRevLett.73.814 PMID: 10057546
31. 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. https://doi.org/10.1016/S0167-2789(00)00174-3.
32. Ruderman DL. Origins of scaling in natural images. Vision Research. 1997; 37(23):3385–3398. https://doi.org/10.1016/S0042-6989(97)00008-4 PMID: 9425551
33. 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. https://doi.org/10.1063/1.4863478
34. Mainardi F, Pironi P. The Fractional Langevin Equation: Brownian Motion Revisited; 2008.
35. Jeon JH, Metzler R. Fractional Brownian motion and motion governed by the fractional Langevin equation in confined geometries. Phys Rev E. 2010; 81:021103. https://doi.org/10.1103/PhysRevE.81.021103
36. Łuksza M, Lässig M. A predictive fitness model for influenza. Nature. 2014; 507(7490):57–61. https://doi.org/10.1038/nature13087

37. Dolan PT, Whitfield ZJ, Andino R. Mapping the Evolutionary Potential of RNA Viruses. Cell Host & Microbe. 2018; 23(4):435–446. https://doi.org/10.1016/j.chom.2018.03.012 PMID: 29649440

38. 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; 160(4):785–797. https://doi.org/10.1016/j.cell.2015.01.027 PMID: 25662010

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

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

41. 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; 13(11):1–25. https://doi.org/10.1371/journal.ppat.1006702 PMID: 29155894

42. 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. https://doi.org/10.1073/pnas.1711114115

43. Painsky A, Tishby N. Gaussian Lower Bound for the Information Bottleneck Limit. Journal of Machine Learning Research. 2018; 18:1–29

44. Kolchinsky A, Tracey BD, Wolpert DH. Nonlinear Information Bottleneck. Entropy. 2019; 21(12) https://doi.org/10.3390/e21121181

45. Kimura M. Diffusion Models in Population Genetics. Journal of Applied Probability. 1964; 1(2):177–232. https://doi.org/10.2307/3211856

46. Arimoto S. An algorithm for computing the capacity of arbitrary discrete memoryless channels. IEEE Transactions on Information Theory. 1972; 18(1):14–20. https://doi.org/10.1109/TIT.1972.1054753

47. Blahut RE. Computation of channel capacity and rate-distortion functions. IEEE Trans Inform Theory. 1972; 18:460–473. https://doi.org/10.1109/TIT.1972.1054855

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

49. Perelson A. S. Immune Network Theory. Immunological Reviews. 1989; 110:5–36. https://doi.org/10.1111/j.1600-065X.1989.tb00025.x PMID: 2477327

50. Nourmohammad A, Schiffels S, Lässig M. Evolution of molecular phenotypes under stabilizing selection. Journal of Statistical Mechanics: Theory and Experiment. 2013; 01:1–12.

51. Wu T, Fischer I, Chuang IL, Tegmark M. Learnability for the Information Bottleneck. Entropy. 2019; 21(10):924. https://doi.org/10.3390/e21100924

52. 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.

53. 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. https://doi.org/10.1115/1.3662552

54. 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. https://doi.org/10.1016/j.cels.2019.08.008 PMID: 31563474

55. Walczak AM, Tkačik Gcv, Bialek W. Optimizing information flow in small genetic networks. II. Feed-forward interactions. Phys Rev E. 2010; 81:041905. https://doi.org/10.1103/PhysRevE.81.041905

56. Nikitin AP, Stocks NG, Morse RP, McDonnell MD. Neural Population Coding Is Optimized by Discrete Tuning Curves. Phys Rev Lett. 2009; 103:138101. https://doi.org/10.1103/PhysRevLett.103.138101

57. Smith JG. The information capacity of amplitude- and variance-constrained scalar gaussian channels. Information and Control. 1971; 18(3):203–219. https://doi.org/10.1016/S0019-9958(71)90346-9

58. 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; 9(5):e96333–e96333. https://doi.org/10.1371/journal.pone.0096333

59. 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; 37(suppl_1):D291–D294. https://doi.org/10.1093/nar/gkn695 PMID: 18931377

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