Thermodynamics of Modularity: Structural Costs Beyond the Landauer Bound

Alexander B. Boyd,[1,*] Dibyendu Mandal,[2,†] and James P. Crutchfield[1,‡]
[1]Complexity Sciences Center and Physics Department, University of California at Davis, One Shields Avenue, Davis, California 95616, USA
[2]Department of Physics, University of California, Berkeley, California 94720, USA

(Received 15 August 2017; revised manuscript received 9 April 2018; published 3 August 2018)

Information processing typically occurs via the composition of modular units, such as the universal logic gates found in discrete computation circuits. The benefit of modular information processing, in contrast to globally integrated information processing, is that complex computations are more easily and flexibly implemented via a series of simpler, localized information processing operations that only control and change local degrees of freedom. We show that, despite these benefits, there are unavoidable thermodynamic costs to modularity—costs that arise directly from the operation of localized processing and that go beyond Landauer's bound on the work required to erase information. Localized operations are unable to leverage global correlations, which are a thermodynamic fuel. We quantify the minimum irretrievable dissipation of modular computations in terms of the difference between the change in global nonequilibrium free energy, which captures these global correlations, and the local (marginal) change in nonequilibrium free energy, which bounds modular work production. This modularity dissipation is proportional to the amount of additional work required to perform a computational task modularly, measuring a *structural* energy cost. It determines the thermodynamic efficiency of different modular implementations of the same computation, and so it has immediate consequences for the architecture of physically embedded transducers, known as information ratchets. Constructively, we show how to circumvent modularity dissipation by designing internal ratchet states that capture the information reservoir's global correlations and patterns. Thus, there are routes to thermodynamic efficiency that circumvent globally integrated protocols and instead reduce modularity dissipation to optimize the architecture of computations composed of a series of localized operations.

DOI: 10.1103/PhysRevX.8.031036

Subject Areas: Computational Physics, Statistical Physics

I. INTRODUCTION

Physically embedded information processing operates via thermodynamic transformations of the supporting material substrate. The thermodynamics is best exemplified by Landauer's principle [1]: Erasing a single bit of stored information at temperature T must be accompanied by the expense of at least $k_B T \ln 2$ amount of heat released into the substrate. While the Landauer cost is only time asymptotic and not yet the most significant energy demand in everyday computations—in our cell phones, tablets, laptops, and cloud computing—there is a clear trend and desire to increase thermodynamic efficiency. Digital technology is expected, for example, to reach the vicinity of the Landauer cost in the near future [2]. This seeming inevitability forces us to ask if the Landauer bound can be achieved for more complex information processing tasks than writing or erasing a single bit of information.

In today's massive computational tasks, in which vast arrays of bits are processed in sequence and in parallel, a task is often broken into modular components to add flexibility and comprehensibility to hardware and software design. This holds far beyond the arenas of today's digital computing. Rather than tailoring processors to do only the task specified, there is great benefit in modularly deploying elementary, but universal functional components—e.g., NAND, NOR, and perhaps Fredkin [3] logic gates, biological neurons [4], or similar units appropriate to other domains [5]—that can be linked together into circuits that perform any functional operation. This leads naturally to hierarchical design, perhaps across many organizational levels. In these ways, the principle of modularity reduces the challenges of designing, monitoring, and diagnosing efficient processing considerably [6,7]. Designing each modular

*abboyd@ucdavis.edu
†dibyendu.mandal@berkeley.edu
‡chaos@ucdavis.edu

component of a complex computation to be efficient is vastly simpler than designing and optimizing the whole. Even biological evolution seems to have commandeered prior innovations, remapping and reconnecting modular functional units to form new organizations and new organisms of increasing survivability [8].

There is, however, a potential thermodynamic cost to modular information processing. For concreteness, recall the stochastic computing paradigm in which an input (a sequence of symbols) is sampled from a given probability distribution and the symbols are correlated to each other. In this setting, a modularly designed computation processes only the *local* component of the input, ignoring the latter's *global* structure. This inherent locality is a physical control restriction and, thus, can lead to thermodynamic inefficiency [9,10]. Local control in modular systems necessarily leads to irretrievable loss of global correlations during computing. Since such correlations are a thermodynamic resource [11,12], their loss implies an energy cost—a thermodynamic *modularity dissipation*.

Employing stochastic thermodynamics and information theory, we show how modularity dissipation arises by deriving an exact expression for dissipation in a generic localized information processing operation. We emphasize that this dissipation is above and beyond the Landauer bound for losses in the operation of single logical gates. The mechanism responsible for modularity dissipation is distinct from that underlying Landauer's principle—state-space contraction due to mesoscopic control that implements logically irreversible operations. It arises solely from the modular state-space architecture of complex computations. One immediate consequence is that the additional dissipation requires investing additional work to drive computation forward.

The additional work corresponds to the Universe's entropy production, much like the reduction in possible entropy extraction for open-loop feedback control when compared to closed-loop feedback [13,14]. In the special case where all correlations between the local modular component and the rest of the system are destroyed, the reduction in entropy extraction for open-loop feedback is the same as the additional work dissipated in modular operations. However, open-loop and closed-loop feedback specify *different* computations. This contrasts with our focus on different structural implementations of the *same* computation, meaning the same input-to-output channel. For a particular computation, the stochastic thermodynamics of control [15] provides tools to evaluate the energetic efficiency of different types of Hamiltonian control: local modular versus globally integrated.

In general, to minimize work invested in performing a computation, we must leverage the global correlations in a system's environment. Globally integrated computations can achieve the minimum dissipation by simultaneous control of the whole system, manipulating the joint system-environment Hamiltonian to follow the desired joint distribution. Not only is this level of control difficult to implement physically, but designing the required protocol poses a considerable computational challenge in itself, with so many degrees of freedom and a potentially complex state space. Genetic algorithm methods have been proposed, though, for approximating the optimum [16]. Tellingly, they can find unusual solutions that break conventional symmetries and take advantage of the correlations between the many different components of the entire system [17,18]. However, as we will show, it is possible to rationally design local information processors that, by accounting for these correlations, minimize modularity dissipation.

The following shows how to design optimal modular computational schemes such that useful global correlations are not lost, but stored in the structure of the computing mechanism. Since the global correlations are not lost in these optimal schemes, the net processing can be thermodynamically reversible (dissipationless). Utilizing the tools of information theory and computational mechanics—Shannon information measures and optimal hidden Markov generators—we identify the informational system structures that can mitigate and even nullify the potential thermodynamic cost of modular computation.

A brief tour of our main results will help orient the reader. It can even serve as a complete, but approximate description for the approach and technical details, should this be sufficient for the reader's interests.

Section II considers the thermodynamics of a composite information reservoir [19], in which only a subsystem is amenable to external control. Information reservoirs, which do not change energy with state changes, are relatively new thermodynamic constructs used for information storage and manipulation [19]. The composite information reservoir described in the following gives a basis for general localized thermodynamic information processing. We assume that the information reservoir is coupled to an ideal heat bath, as a source of randomness and energy—ideal in that it has infinite heat capacity and no memory of past interactions with the information reservoir. Thus, (i) external control of the information reservoir yields random Markovian dynamics over its *informational states*, as we call them, (ii) heat flows into the heat bath, and (iii) work investment comes from the controller. Statistical correlations may exist between the controlled and uncontrolled subsystems, either due to initial or boundary conditions or due to an operation's history.

To highlight the information-theoretic origin of the dissipation and to minimize the energetic aspects, we assume that the informational states have equal internal (free) energies. Appealing to stochastic thermodynamics and information theory, we then show that the minimum irretrievable *modularity dissipation* over the duration of an operation due to the locality of control is proportional to

the reduction in mutual information between the controlled and uncontrolled subsystems; see Eq. (8). We deliberately refer to "operation" here instead of "computation," since the result holds whether the desired task is interpreted as computation or not. The result holds so long as free-energy uniformity is satisfied at all times, a condition natural in computation and other information processing settings.

Section IV applies this analysis to information engines, an active subfield within the thermodynamics of computation in which information effectively acts as the fuel for driving physically embedded information processing [20–24]. The particular implementations of interest—information ratchets—process an input symbol string by interacting with each symbol in order, sequentially transforming it into an output symbol string, as shown in Fig. 3. This kind of information transduction [21,25] is information processing in a very general sense: With properly designed dynamics over an infinite reservoir of internal states, the devices can implement a universal Turing machine [26]. Since information engines rely on localized information processing, reading in and manipulating one symbol at a time in their original design [20], the measure of irretrievable dissipation applies directly. The exact expression for their modularity dissipation is given in Eq. (17).

Sections V and VI specialize information transducers further to the cases of pattern extractors and pattern generators. Section V's pattern extractors use structure in their environment to produce work and pattern generators use stored work to create structure from an unstructured environment. The irreversible relaxation of correlations in information transduction can then be curbed by intelligently designing these computational processes. While there are not yet general principles for designing implementations for arbitrary computations, the measure of modularity dissipation that we develop shows how to construct energy-efficient extractors and generators. For example, efficient extractors consume complex patterns and turn them into sequences of independent and identically distributed (IID) symbols.

We show that extractor transducers whose states are predictive of their inputs are optimal, with zero minimal modularity dissipation. This makes immediate intuitive sense since, by design, such transducers can anticipate the next input and adapt accordingly. This observation also emphasizes the principle that thermodynamic agents should requisitely match the structural complexity of their environment to leverage those informational correlations as a thermodynamic fuel [23]. We illustrate this result in the case of the golden mean pattern in Fig. 4.

Conversely, Sec. VI shows that, when generating patterns from unstructured IID inputs, transducers whose states are retrodictive of their output are most efficient—i.e., have minimal modularity dissipation. This is also intuitively appealing in that pattern generation may be viewed as the time reversal of pattern extraction. Since predictive transducers are efficient for pattern extraction, retrodictive transducers are expected to be efficient pattern generators; see Fig. 6. This also allows one to appreciate that pattern generators previously thought to be asymptotically efficient are actually quite dissipative [27]. Taken altogether, these results provide guideposts for designing efficient, modular, and complex information processors—guideposts that go substantially beyond Landauer's principle for localized processing.

II. GLOBAL VERSUS LOCALIZED PROCESSING

If a physical system, denote it $\mathcal{Z}$, stores information as it behaves, it acts as an information reservoir. Then, a wide range of physically embedded computational processes can be achieved by connecting $\mathcal{Z}$ to an ideal heat bath at temperature T and externally controlling the system's physical parameters, its Hamiltonian. Coupling with the heat bath allows for physical phase-space compression and expansion, which are necessary for useful computations and which account for the work investment and heat dissipation dictated by Landauer's bound. However, the bound is only achievable when the external control is precisely designed to harness the changes in phase space. This may not be possible for modular computations. The modularity here implies that control is localized and potentially ignorant of global correlations in $\mathcal{Z}$. This leads to uncontrolled changes in phase space.

Most computational processes unfold via a sequence of local operations that update only a portion of the system's informational state. A single step in such a process can be conveniently described by breaking the whole informational system $\mathcal{Z}$ into two constituents: the informational states $\mathcal{Z}^{\text{int}}$ that are controlled and evolving and the informational states $\mathcal{Z}^{\text{stat}}$ that are not part of the local operation on $\mathcal{Z}^{\text{int}}$. We call $\mathcal{Z}^{\text{int}}$ the "interacting" subsystem and $\mathcal{Z}^{\text{stat}}$ the "stationary" subsystem. As shown in Fig. 1, the dynamic over the joint state space $\mathcal{Z} = \mathcal{Z}^{\text{int}} \otimes \mathcal{Z}^{\text{stat}}$ is the product of the identity over the stationary subsystem and a Markov channel over the interacting subsystem. We refer to the latter as a "local Markov channel," since it only updates the local interacting degrees of freedom. The informational states of the noninteracting stationary subsystem $\mathcal{Z}^{\text{stat}}$ are fixed over the immediate computational task, since this information should be preserved for use in later computational steps.

Such classical computations are described by a global Markov channel over the joint state space:

$$M^{\text{global}}_{z^{\text{i}}_t, z^{\text{s}}_t \to z^{\text{i}}_{t+\tau}, z^{\text{s}}_{t+\tau}} = \Pr(Z^{\text{i}}_{t+\tau} = z^{\text{i}}_{t+\tau}, Z^{\text{s}}_{t+\tau} = z^{\text{s}}_{t+\tau} | Z^{\text{i}}_t = z^{\text{i}}_t, Z^{\text{s}}_t = z^{\text{s}}_t), \tag{1}$$

where $Z_t = Z^{\text{i}}_t \otimes Z^{\text{s}}_t$ and $Z_{t+\tau} = Z^{\text{i}}_{t+\tau} \otimes Z^{\text{s}}_{t+\tau}$ are the random variables for the informational state of the joint system before and after the computation, with the random variable

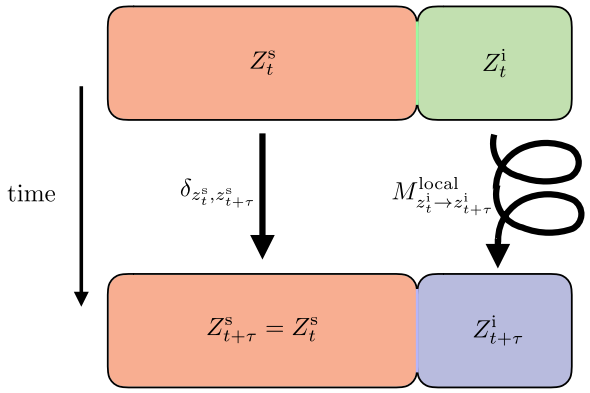

FIG. 1. Local computations operate on only an interacting subset $\mathcal{Z}^{\mathrm{int}}$ of the entire information reservoir $\mathcal{Z} = \mathcal{Z}^{\mathrm{int}} \otimes \mathcal{Z}^{\mathrm{stat}}$ described by random variable $Z = Z^i \otimes Z^s$. The Markov channel that describes the global dynamic is the product of a local operation with the identity operation: $M^{\mathrm{global}}_{(z_t^{\mathrm{i}},z_t^{\mathrm{s}})\to(z_{t+\tau}^{\mathrm{i}},z_{t+\tau}^{\mathrm{s}})} = M^{\mathrm{local}}_{z_t^{\mathrm{i}}\to z_{t+\tau}^{\mathrm{i}}}\delta_{z_t^{\mathrm{s}},z_{t+\tau}^{\mathrm{s}}}$, such that the stationary noninteracting portion Z^{s} of the information reservoir remains invariant, but the interacting portion Z^{i} changes.

Z^i describing the $\mathcal{Z}^{\mathrm{int}}$ subspace and the random variable Z^s the $\mathcal{Z}^{\mathrm{stat}}$ subspace, respectively. From here on, we often refer to the random variables Z^i and Z^s rather than their state spaces $\mathcal{Z}^{\mathrm{int}}$ and $\mathcal{Z}^{\mathrm{stat}}$ when describing the system. (Lowercase variables denote values their associated random variables realize.) The right-hand side of Eq. (1) gives the transition probability over the time interval $(t, t+\tau)$ from joint state $(z_t^{\mathrm{i}}, z_t^{\mathrm{s}})$ to state $(z_{t+\tau}^{\mathrm{i}}, z_{t+\tau}^{\mathrm{s}})$. The fact that $\mathcal{Z}^{\mathrm{stat}}$ is fixed means that the global dynamic can be expressed as the product of a local Markov computation on $\mathcal{Z}^{\mathrm{int}}$ with the identity over $\mathcal{Z}^{\mathrm{stat}}$,

$$M^{\mathrm{global}}_{(z_t^{\mathrm{i}},z_t^{\mathrm{s}})\to(z_{t+\tau}^{\mathrm{i}},z_{t+\tau}^{\mathrm{s}})} = M^{\mathrm{local}}_{z_t^{\mathrm{i}}\to z_{t+\tau}^{\mathrm{i}}}\delta_{z_t^{\mathrm{s}},z_{t+\tau}^{\mathrm{s}}}, \tag{2}$$

where the local Markov computation is the conditional marginal distribution:

$$M^{\mathrm{local}}_{z_t^{\mathrm{i}}\to z_{t+\tau}^{\mathrm{i}}} = \Pr(Z_{t+\tau}^{\mathrm{i}} = z_{t+\tau}^{\mathrm{i}} | Z_t^{\mathrm{i}} = z_t^{\mathrm{i}}). \tag{3}$$

When the processor is in contact with a heat bath at temperature T, the average entropy production $\langle\Sigma_{t\to t+\tau}\rangle$ of the Universe over the time interval $(t, t+\tau)$ can be expressed in terms of the work done minus the change in nonequilibrium free energy F^{neq}:

$$\langle\Sigma_{t\to t+\tau}\rangle = \frac{\langle W_{t\to t+\tau}\rangle - (F_{t+\tau}^{\mathrm{neq}} - F_t^{\mathrm{neq}})}{T}.$$

In turn, the nonequilibrium free energy F_t^{neq} at any time t can be expressed as the weighted average of the internal (free) energy U_z of the joint informational states minus the uncertainty in those states:

$$F_t^{\mathrm{neq}} = \sum_z \Pr(Z_t = z)U_z - (k_{\mathrm{B}}T\ln 2)H[Z_t]. \tag{4}$$

Here, $H[Z]$ is the Shannon information of the random variable Z that realizes the state of the joint system $\mathcal{Z}$ [15]. When the information-bearing degrees of freedom support an information reservoir, we take all states z and z' to have the same internal energy $U_z = U_{z'}$. This is the situation we consider in the following. Under this assumption, the first term on the right of Eq. (4) does not change even when there is a change in the probability distribution $\Pr(Z_t = z)$. The entropy production of the Universe $\langle\Sigma_{t\to t+\tau}\rangle$ then reduces to the work minus a change in Shannon information of the information-bearing degrees of freedom [15,28]:

$$\langle\Sigma_{t\to t+\tau}\rangle = \frac{\langle W_{t\to t+\tau}\rangle}{T} + k_{\mathrm{B}}\ln 2(H[Z_{t+\tau}] - H[Z_t]). \tag{5}$$

Essentially, this is an expression of a generalized Landauer principle: Increasing entropy of the Universe guarantees that work production is bounded by the change in Shannon entropy of the informational variables [1].

Appendix A describes an isothermal protocol that implements a Markov channel, in this case either M^{global} or M^{local}. By controlling the energy landscape, we exactly specify the form of the computation from input to output. Thus, if one is concerned with implementing deterministic logical operations, we can exponentially reduce any thermal randomness in the computation by making linear changes in energies. Our framing, however, is closer in spirit to modern random computation [29–31], where the outcome of a computation is not a deterministic variable but a random one. In the natural (e.g., biological or molecular) setting, information processing in the presence of noise and stochasticity is the rule, not the exception. Rarely are noise-free discrete computation theory concepts applicable there.

In point of fact, a more general perspective on the current setting would see it as a study of computation in thermodynamic systems, much in the spirit of computational mechanics itself—the mechanics of computation [32]. That is, our approach considers the generation, storage, dissipation, and transmission of information as a thermodynamic system evolves. This provides a broader perspective of which the thermodynamics of computation forms a major component.

For the particular case of a globally integrated isothermal operation, the energy landscape over the whole system space $\mathcal{Z}$ is controlled simultaneously. This achieves zero entropy production. Also, the globally integrated work done on the system achieves the theoretical minimum:

$$\langle W^{\mathrm{global}}_{t\to t+\tau}\rangle_{\mathrm{min}} = -k_{\mathrm{B}}T\ln 2(H[Z_{t+\tau}] - H[Z_t]).$$

The process is reversible since the change in system Shannon entropy balances the change in the reservoir's physical entropy due to heat dissipation.

Note that we do not assume the initial and final microstate probabilities before and after a thermodynamic operation obey equilibrium distributions. Indeed, for any meaningful computation, the system must transition between *nonequilibrium* distributions. This is because equilibrium distributions are uniform distributions, since we assume the internal energies of the information-bearing degrees of freedom are uniform. Because of this, we consider transitions between nonequilibrium, metastable states with a decay time much longer than the experimental timescale. This timescale separation is necessary if information must be stored reliably over long periods of time.

We achieve reversibility between nonequilibrium metastable states if the control timescale is much longer than that of the metastable states' internal dynamics, but much shorter than the timescale of the global equilibration dynamics. This is the regime we consider. Since the internal energy is uniform, the system cannot store the work and must dissipate it as heat to the surrounding environment. This may not hold for a generic modular operation.

There are two consequences of the locality of control. First, since Z^{s} is kept fixed, meaning that $Z^{\mathrm{s}}_t = Z^{\mathrm{s}}_{t+\tau}$, the change in uncertainty $H[Z^{\mathrm{i}}_{t+\tau}, Z^{\mathrm{s}}_{t+\tau}] - H[Z^{\mathrm{i}}_t, Z^{\mathrm{s}}_t]$ of the joint information-bearing variables during the operation—the second term in the left-hand side of Eq. (5)—simplifies to

$$H[Z_{t+\tau}] - H[Z_t] = H[Z^{\mathrm{i}}_{t+\tau}, Z^{\mathrm{s}}_t] - H[Z^{\mathrm{i}}_t, Z^{\mathrm{s}}_t]. \tag{6}$$

Second, Appendix C shows that if the joint system Z is an information reservoir with control limited to subsystem Z^i, then there is no energetic coupling between Z^i and Z^s. The lack of energetic coupling to stationary subsystem Z^s implies that the interacting subsystem is effectively isolated from the stationary subsystem. Thus, on its own, the interacting subsystem matches the framework for an open driven system described in Ref. [21], and so the entropy production $\langle \Sigma^i \rangle = \langle W \rangle - \Delta F^i$ estimated from the interacting system alone must be non-negative. In this,

$$F^i_t = \sum_{z \in \mathcal{Z}^i} \Pr(Z^i_t = z) U_z - (k_B T \ln 2) H[Z^i_t]$$

is the marginalized estimate of the nonequilibrium free energy isolated to the interacting system [28]. As a result, the work investment is bounded by the change in the marginalized estimate of the nonequilibrium free energy. This implies, in turn, a generalized Landauer principle corresponding to the change in marginal distribution over Z^{i}, which is determined by the local Markov channel shown in Eq. (2). In other words, absent control over the noninteracting subsystem Z^s, which remains stationary over the local computation on Z^{i}, the work done $\langle W_{t \to t+\tau} \rangle$ on Z^i is bounded below:

$$\langle W_{t \to t+\tau} \rangle \geq \langle W^{\mathrm{local}}_{t \to t+\tau} \rangle_{\min} = k_{\mathrm{B}} T \ln 2 (H[Z^{\mathrm{i}}_t] - H[Z^{\mathrm{i}}_{t+\tau}]). \tag{7}$$

This information-theoretic bound on the work is achievable, as described in Appendix A, by an isothermal process composed of slow manipulations of the energy landscape of the interacting subsystem, which evolves the entire system between nonequilibrium metastable distributions.

Combining the last two relations with the expression for entropy production in Eq. (5) gives the *modularity dissipation* Σ^{mod}, which is the minimum irretrievable dissipation of a modular computation that comes from local interactions:

$$\begin{aligned} \frac{\langle \Sigma^{\mathrm{mod}}_{t \to t+\tau} \rangle_{\min}}{k_{\mathrm{B}} \ln 2} &= \frac{\langle W^{\mathrm{local}}_{t \to t+\tau} \rangle_{\min}}{k_B T \ln 2} + H[Z^i_{t+\tau}, Z^s_t] - H[Z^i_t, Z^s_t] \\ &= H[Z^i_t] - H[Z^i_{t+\tau}] + H[Z^i_{t+\tau}, Z^s_t] \\ &\quad - H[Z^i_t, Z^s_t] + (H[Z^s_t] - H[Z^s_t]) \\ &= I[Z^i_t; Z^s_t] - I[Z^i_{t+\tau}; Z^s_t], \end{aligned} \tag{8}$$

where $I[X;Y] = H[X] + H[Y] - H[X,Y]$ is the mutual information between the random variables X and Y. While this bound on dissipation was established assuming that the energetically uncoupled and uncontrolled portion Z^s of the system is stationary, it also applies to modular computations where the uncontrolled system evolves under its own dynamics, independent of control. This follows from the facts that the uncontrolled system's evolution can only lead it to dissipate nonequilibrium free energy, as there is no work done on it, and the uncontrolled stationary subsystem can only increase the Universe's entropy production, if allowed to change [28]. We require the uncontrolled system to be stationary Z^s, as this puts the strictest bound on dissipation, and since an efficient computation holds elements fixed when they are not being actively changed.

This is our central result: a thermodynamic cost for modular operations above and beyond the Landauer bound for logically irreversible operations. It is an additional cost above the bound that arises from a distinct mechanism beyond Landauer's microscopic state-space contraction. Differing from Landauer's principle, it arises from a computation's implementation architecture. Specifically, the minimum entropy production is proportional to the minimum additional work that must be done to execute a computation modularly:

$$\langle W^{\mathrm{local}}_{t \to t+\tau} \rangle_{\min} - \langle W^{\mathrm{global}}_{t \to t+\tau} \rangle_{\min} = T \langle \Sigma^{\mathrm{mod}}_{t \to t+\tau} \rangle_{\min}.$$

Appendix A describes how to achieve this minimum dissipation through isothermal protocols. Because of the bound set on the work by the local entropy change shown in

Eq. (7), any alternative protocol, perhaps done in finite time [33] or with unobserved coarse-grained variables [34], would necessarily require more work to implement. The following draws out the implications.

Using the fact that the local operation M^{local} ignores Z^{s}, we see that the joint distribution over all three variables Z_t^{i}, Z_t^{s}, and $Z_{t+\tau}^{\text{i}}$ can be simplified to

$$\begin{aligned}&\Pr(Z_{t+\tau}^{\text{i}} = z_{t+\tau}^{\text{i}}, Z_t^{\text{i}} = z_t^{\text{i}}, Z_t^{\text{s}} = z_t^{\text{s}}) \\ &\quad = \Pr(Z_{t+\tau}^{\text{i}} = z_{t+\tau}^{\text{i}} | Z_t^{\text{i}} = z_t^{\text{i}}) \Pr(Z_t^{\text{i}} = z_t^{\text{i}}, Z_t^{\text{s}} = z_t^{\text{s}}).\end{aligned}$$

Thus, Z_t^{i} *shields* $Z_{t+\tau}^{\text{i}}$ from Z_t^{s}. A consequence is that the mutual information between $Z_{t+\tau}^{\text{i}}$ and Z_t^{s} conditioned on Z_t^{i} vanishes. This is shown in Fig. 2 via an information diagram—a tool that lays out informational interdependencies between random variables [35] and has been particularly useful in analyzing temporal information processing [36,37]. Figure 2 also shows that the modularity dissipation, highlighted by a dashed red outline, can be reexpressed as the mutual information between the noninteracting stationary system Z^{s} and the interacting system Z^{i} before the computation that is not shared with Z^{i} after the computation:

$$\begin{aligned}\langle \Sigma_{t\to t+\tau}^{\text{mod}} \rangle_{\text{min}} &= k_B \ln 2 (\text{I}[Z_t^i; Z_t^s] - I[Z_{t+\tau}^i; Z_t^s]) \\ &= k_B \ln 2 (I[Z_t^i; Z_t^s | Z_{t+\tau}^i] + I[Z_t^i; Z_t^s; Z_{t+\tau}^i] \\ &\quad - I[Z_{t+\tau}^i; Z_t^s | Z_t^i] - I[Z_t^i; Z_t^s; Z_{t+\tau}^i]) \\ &= (k_B \ln 2) \text{I}[Z_t^i; Z_t^s | Z_{t+\tau}^i], \end{aligned} \tag{9}$$

where, in the second line, we used the expression for three-variable mutual information $I[X;Y;Z] = I[X;Y] - I[X;Y|Z]$ and, to get to our final result, we appealed to the shielding $I[Z_{t+\tau}^i; Z_t^s | Z_t^i] = 0$. This is our second main result. The conditional mutual information on the right bounds how much entropy is produced when performing a local computation. It quantifies the irreversibility of modular information processing.

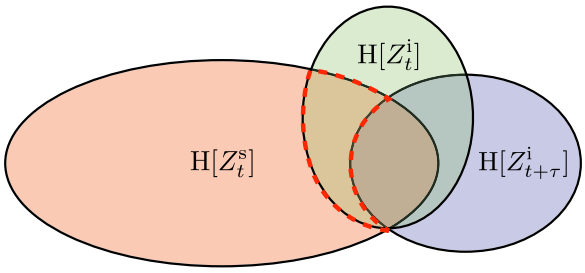

FIG. 2. Information diagram for a local computation: Information atoms of the noninteracting subsystem $H[Z_t^s]$ (red ellipse), the interacting subsystem before the computation $H[Z_t^i]$ (green ellipse), and the interacting subsystem after the computation $H[Z_{t+\tau}^i]$ (blue ellipse). The initial state of the interacting subsystem shields the final state from the noninteracting subsystem; graphically the blue and red ellipses only overlap within the green ellipse. The modularity dissipation is proportional to the difference between information atoms $I[Z_t^i; Z_t^s]$ and $I[Z_{t+\tau}^i; Z_t^s]$. Because of statistical shielding, it simplifies to the information atom $I[Z_t^i; Z_t^s | Z_{t+\tau}^i]$, highlighted by a red dashed outline.

III. PRIOR THERMODYNAMICS OF CORRELATION

The thermodynamics of modularity lets us revisit prior results in a new light. The cost in Eq. (9) was recognized in the context of copying and measurement [38] and is relevant to biological push-pull systems [39]. While, in principle, the logical operations performed by biological systems can be performed reversibly if done quasistatically, Ref. [39] showed that these biological processes have control restrictions that lead to thermodynamic inefficiencies. First of all, biochemical systems are often constrained to hold chemical potentials constant. They perform logical operations instead by removal of barriers and so dissipate potential sources of work. This is particularly relevant to decorrelating readouts from sensory receptors, which is a source of thermodynamic dissipation in a biological implementation of Szilard's engine. Treating the readout as the interacting subsystem and the receptor as the stationary subsystem, this inefficiency is predicted from modularity dissipation. If the noninteracting stationary subsystem is uniformly distributed, such that $H[Z_t^s] = \log_2 |\mathcal{Z}^s|$, and the interacting subsystem is a perfect copy of that system, then all structure in the information reservoir comes in the form of correlations between the subsystems, such that $I[Z_t^i; Z_t^s] = H[Z_t^s]$. If we perform a decorrelation operation, mapping the interacting system to a uniform distribution and decorrelating the two subsystems such that $I[Z_{t+\tau}^i; Z_t^s] = 0$, we potentially can recover $(k_B T \ln 2) H[Z_t^s]$ of work from the system with globally integrated control and energetic coupling between subsystems. However, if the control is local, all those correlations are dissipated in the decorrelation operation, as reflected by the modularity dissipation:

$$\begin{aligned}\langle \Sigma_{t\to t+\tau}^{\text{mod}} \rangle_{\text{min}} &= k_{\text{B}} \ln 2 (I[Z_t^i; Z_t^s] - I[Z_{t+\tau}^i; Z_t^s]) \\ &= (k_{\text{B}} \ln 2) H[Z_t^s],\end{aligned}$$

since energetic coupling is impossible in modular computations. The modularity dissipation imposes an energetic cost on thermodynamic systems when they decorrelate with their environment. The cost applies to a wide variety of information-processing physical agents, including Maxwell's demon.

Modularity dissipation can be tested experimentally with implementations of Szilard's engine—a two-bit Maxwell's demon. The authors of Ref. [40] showed that the Szilard engine can be explicitly implemented by a two-dimensional system, with one degree of freedom corresponding to its

environment (system under study) and the other corresponding to the demon's memory. The step of the engine's functioning where the demon extracts work, shown in Fig. 1 of Ref. [40] as the "control" step, decorrelates the demon with its environment. According to modularity dissipation, then, the correlations must be dissipated if the system under study is controlled modularly. Thus, the only way for the demon to extract work from its environment is to go beyond modular control, dynamically changing the energetic coupling between its memory and environment. Controllable bistable thermodynamic systems, such as Bose-Einstein condensates [41], nanoelectromechanical systems [42], flux qubits [43,44], and feedback traps [45], can store information bistably and so are candidates for experimentally implementing both the demon and its environment in a Szilard engine.

We can probe modularity dissipation in experimental implementations of information reservoirs by comparing the work generated with local control to the work generated with globally integrated control. For the feedback operation that decorrelates the system under study (SUS) and the demon in Szilard's engine, the SUS is the interacting subsystem $Z^i = Z^{\mathrm{SUS}}$, and the demon's memory is fixed, so that it is the stationary subsystem $Z^s = Z^{\mathrm{demon}}$. For this feedback step, modular control means that the externally controlled Hamiltonian [9] is the SUS's Hamiltonian:

$$\mathcal{H}^{\mathrm{ext}}(t) = \mathcal{H}^{\mathrm{SUS}}(t). \tag{10}$$

On the one hand, the local version of Landauer's bound means that the work invested in the decorrelation step should be bounded below by 0, since the marginal state entropy does not change—$H[Z_t^{\mathrm{SUS}}] = H[Z_{t+\tau}^{\mathrm{SUS}}]$—despite the global state changing. Thus, with modular control, Szilard's engine cannot function as it was designed by extracting work during its decorrelation/feedback step. On the other hand, if we use globally integrated control, where $H^{\mathrm{ext}}(t)$ includes coupling terms between the SUS and the demon, then there are protocols that can extract the free energy stored in correlations:

$$\begin{aligned}\Delta F^{\mathrm{neq}} &= (k_B T \ln 2) I[Z_t^{\mathrm{SUS}}; Z_t^{\mathrm{demon}}] \\ &= (k_B T \ln 2) H[Z_t^{\mathrm{demon}}].\end{aligned}$$

The modularity dissipation is the difference between this work, extracted with globally integrated control, and that extracted with optimal local control.

The form of modularity dissipation shown in Eq. (8)—a difference of mutual information—has arisen before in a different context and with different meaning [46,47]. These works show that the unutilized change in free energy corresponds to dissipated work. In the setting of data representations, Eq. (8)'s bound is analogous to the expression for the minimum work required for data representation, with Z_t^i being the work medium, $Z_{t+\tau}^i$ the work extraction device, and Z_t^s the data representation device [47].

Given this parallel, a study of the thermodynamics of prediction in a system driven by an input signal [46] shows that the irretrievable work dissipation,

$$\beta\langle W_{\mathrm{diss}}[X_t \to X_{t+1}]\rangle = I[S_t; X_t] - I[S_t; X_{t+1}],$$

is proportional to the modularity dissipation, if the driving signal X_t is treated as the interacting subsystem Z_t^i and the driven system S_t is treated as the stationary subsystem Z_t^s. While formally similar, the setup is importantly different from the cost of local modular control of information processing. Most practically, the frameworks lead to different results. This is especially clear for signal transduction.

The next section draws out the implications of Eqs. (8) and (9) for information transducers–information processing architectures, in which the processor sequentially takes one input symbol at a time and performs localized computation on it, much as a Turing machine operates. To continue the comparison, these devices respond to an input signal much as the driven systems discussed in Ref. [46]. The irretrievable dissipation the latter derives for its driven systems can be minimized by ensuring that the driven system not store any unwarranted information about the input, beyond that required to predict [46], meaning that the instantaneous memory $I_{\mathrm{mem}} = I[S_t; X_t]$ and instantaneous predictive power $I_{\mathrm{pred}} = I[S_t; X_{t+1}]$ are the same. This means that thermodynamic efficiency can be achieved when the driven system has no memory $H[S_t] = 0$. In this case, the system neither stores nor predicts any information about the input: $I[S_t; X_t] = I[S_t; X_{t+1}] = 0$. For structured inputs to an information transducer, in contrast, we see distinctly different thermodynamics. Section V not only demonstrates that memoryless systems are thermodynamically inefficient, but also proves that predictive memory states are necessary for efficient extractors.

Moreover, the transducer framework allows one to move beyond the task of prediction. We see results reminiscent of Ref. [46] with pattern generators in Sec. VI, in that generators are thermodynamically efficient when they store as little unwarranted information as possible. However, these devices are retrodicting rather than predicting—a different task. In short, though the language and mathematics of the thermodynamics of modularity seem to parallel that of driven systems, modularity dissipation more directly speaks to the design of efficient controllers. By analyzing transducers in Secs. IV–VI as a concrete and flexible form of input-driven information processing, we find results that circumvent Ref. [46]'s interpretation that memoryless driven systems are optimal. More to the point, the results specify the memoryful mechanism by which the physical information processor predicts its input.

Various analyses of other information-driven processes have been developed previously. For example, the authors

of Ref. [48] considered a setting in which a system is subject to an external force that depends on the system's instantaneous position—the system obeys a modified second law of thermodynamics [49]. The authors of Ref. [50] considered a situation where a chemical force replaces the role of information in driving the system out of equilibrium and extracting work. The authors of Ref. [51] analyzed how the mutual information between two spatial degrees of freedom in a biochemical context acts as a thermodynamic resource. Drawing out the current results' implications for these previous settings must be left for the future.

While we identified the inherent dissipation due to modular computations, suggesting that globally integrated control leads to more thermodynamically efficient computations, we see in the transducer context that there are alternative paths to thermodynamic efficiency. For example, modularity dissipation can be minimized by designing a computation such that the modular components themselves store the relevant global correlations, preventing dissipation. Locality and modularity are natural parts of complex computations, so rather than rely on the ability to simultaneously control the global energy landscape, we use modularity dissipation as a structural guide to design modular computational architectures that are thermodynamically efficient.

Modular design is not the sole province of computation *in silico*. Modularity appears in the structure and function of biological organisms as well [52]. Our results can be viewed as providing the information-theoretic and thermodynamic backdrop with which to understand modular biological functions such as memory [23], self-correction [24], and pattern formation [53], among others.

IV. INFORMATION TRANSDUCERS: LOCALIZED PROCESSORS

Information ratchets [21,54] are thermodynamic implementations of information transducers [25] that sequentially transform an input symbol string into an output string. As generalized input-output machines, these devices have been used as autonomous information engines or erasers [20,21], refrigerators [55], pattern generators [27,53], random number generators [31], and self-correcting correlation-powered engines [24]. They have an incredibly wide variety of functionality in turning an input into an output. The ratchet traverses the input symbol string (random variables $Y_{0:\infty} = Y_0 Y_1 Y_2 \ldots$) unidirectionally, processing each symbol in turn to yield the output sequence (random variables $Y'_{0:\infty} = Y'_0 Y'_1 Y'_2 \ldots$). (Here, $Y_{a:b}$ denotes the string of random variables from a to b, $Y_a Y_{a+1} \ldots Y_{b-2} Y_{b-1}$, including a but excluding b.)

As shown in Fig. 3, at time $t = N\tau$, the information reservoir is described by the joint distribution over the ratchet state X_N and the symbol string $\mathbf{Y}_N = Y'_{0:N} Y_{N:\infty}$, the concatenation of the first N symbols of the output string and the remaining symbols of the input string. (This differs slightly from previous treatments [24], in which only the symbol string is the information reservoir. The information processing and energetics are the same, however.) Including the ratchet state in the present definition of the information reservoir allows us to directly determine the modularity dissipation of information transduction.

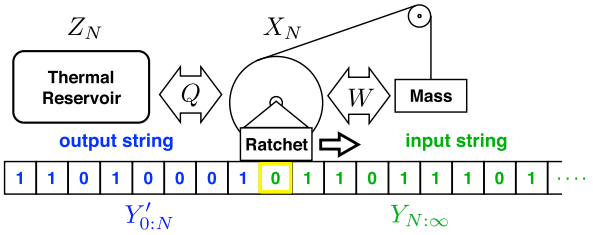

FIG. 3. Information ratchet consists of three interacting reservoirs—work, heat, and information. The work reservoir is depicted as gravitational mass suspended by a pulley. The thermal reservoir keeps the entire system thermalized to temperature T. At time $N\tau$, the information reservoir consists of (i) a string of symbols $\mathbf{Y}_N = Y'_0 Y'_1 \ldots Y'_{N-1} Y_N Y_{N+1} \ldots$, each cell storing an element from the same alphabet $\mathcal{Y}$, and (ii) the ratchet's internal state X_N. The ratchet moves unidirectionally along the string, exchanging energy between the heat and the work reservoirs. The ratchet reads the value of a single cell (highlighted in yellow) at a given time from the input string (green, right); interacts with it; and writes a symbol to the cell in the output string (blue, left) of the information reservoir. Overall, the ratchet transduces the input string $Y_{0:\infty} = Y_0 Y_1 \ldots$ into an output string $Y'_{0:\infty} = Y'_0 Y'_1 \ldots$. (Reprinted from Ref. [21] with permission.)

Operations from time $t = N\tau$ to $t + \tau = (N+1)\tau$ preserve the state of the current output history $Y'_{0:N}$ and the input future, excluding the Nth symbol $Y_{N+1:\infty}$, while changing the Nth input symbol Y_N to the Nth output symbol Y'_N and the ratchet from its current state X_N to its next X_{N+1}. In terms of the previous section, this means the noninteracting stationary subsystem $\mathcal{Z}^{\text{stat}}$ is the entire semi-infinite symbol string *without* the Nth symbol:

$$Z_t^{\text{s}} = (Y_{N+1:\infty}, Y'_{0:N}). \tag{11}$$

The ratchet and the Nth symbol constitute the interacting subsystem $\mathcal{Z}^{\text{int}}$ so that, over the time interval $(t, t+\tau)$, only two variables change:

$$Z_t^{\text{i}} = (X_N, Y_N) \tag{12}$$

and

$$Z_{t+\tau}^{\text{i}} = (X_{N+1}, Y'_N). \tag{13}$$

Despite the fact that only a small portion of the system changes on each time step, the physical device is able to perform a wide variety of physical and logical operations. Ignoring the probabilistic processing aspects, Turing showed that a properly designed finite-state transducer

can compute any input-output mapping [56,57]. Such machines, even those with as few as two internal states and a sufficiently large symbol alphabet [58] or with as few as a dozen states but operating on a binary-symbol strings, are *universal* in that sense [59].

Information ratchets—physically embedded, probabilistic Turing machines—are able to facilitate energy transfer between a thermal reservoir at temperature T and a work reservoir by processing information in symbol strings. In particular, they can function as an eraser by using work to create structure in the output string [20,21] or act as an engine by using the structure in the input to turn thermal energy into useful work energy [21]. They are also capable of much more, including detecting, adapting to, and synchronizing to environment correlations [23,53] and correcting errors [24].

Information transducers are a novel form of information processor from a different perspective, that of communication theory's channels [25]. They are memoryful channels that map input stochastic processes to output processes using internal states that allow them to store information about the past of both the input and the output. With sufficient hidden states, as just noted from the view of computation theory, information transducers are Turing complete and so able to perform any computation on the information reservoir [60]. Similarly, the physical steps that implement a transducer as an information ratchet involve a series of modular local computations.

The ratchet operates by interacting with one symbol at a time in sequence, as shown in Fig. 3. The Nth symbol, highlighted in yellow to indicate that it is the interacting symbol, is changed from the input Y_N to output Y'_N over time interval $[N\tau, (N+1)\tau]$. The ratchet and interaction symbol change together according to the local Markov channel over the ratchet-symbol state space:

$$M^{\text{local}}_{(x,y)\to(x',y')} = \Pr(X_{N+1} = x', Y'_N = y' | X_N = x, Y_N = y).$$

This determines how the ratchet transduces inputs to outputs [21].

Each of these localized operations keeps the remaining noninteracting symbols in the information reservoir fixed. If the ratchet only has energetic control of the degrees of freedom it manipulates, then, as discussed in the previous section and Appendix A, the ratchet's work production in the Nth time step is bounded by the change in uncertainty of the ratchet state and interaction symbol:

$$\langle W^{\text{local}}_N \rangle_{\text{min}} = k_B T \ln 2 (H[X_N, Y_N] - H[X_{N+1}, Y'_N]). \quad (14)$$

This bound has appeared in previous investigations of information ratchets [20,61]. Here, we make a key, but important and compatible observation: If we relax the condition of local control of energies to allow for global control of all symbols simultaneously, then it is possible to extract more work.

That is, foregoing localized operations—abandoning modularity—allows for (and acknowledges the possibility of) globally integrated interactions. Then, we can account for the change in Shannon information of the information reservoir—the ratchet and the entire symbol string. This yields a looser upper bound on work production that holds for both modular and globally integrated information processing. Assuming that all information reservoir configurations have the same free energies, the change in the nonequilibrium free energy during one step of a ratchet's computation is proportional to the global change in Shannon entropy:

$$\Delta F^{\text{neq}}_{N\tau\to(N+t)\tau} = k_B T \ln 2 (H[X_N, \mathbf{Y}_N] - H[X_{N+1}, \mathbf{Y}_{N+1}]).$$

Recalling the definition of entropy production $\langle \Sigma \rangle = (\langle W \rangle - \Delta F^{\text{neq}})/T$ reminds us that, for entropy to increase, the minimum work investment must match the change in free energy:

$$\langle W^{\text{global}}_N \rangle_{\text{min}} = k_B T \ln 2 (H[X_N, \mathbf{Y}_N] - H[X_{N+1}, \mathbf{Y}_{N+1}]). \quad (15)$$

This is the work production that can be achieved through globally integrated isothermal information processing. Also, in turn, it can be used to bound the asymptotic work production in terms of the entropy rates of the input and output processes [21]:

$$\lim_{N\to\infty} \langle W_N \rangle \geq k_B T \ln 2 (h_\mu - h'_\mu), \quad (16)$$

where the entropy rate h_μ is the uncertainty per input and h'_μ is the uncertainty per output [62]. This is known as the "information processing second law" (IPSL) [21].

The authors of Ref. [23] already showed that this bound is not necessarily achievable by information ratchets. This is due to ratchets operating locally. The local bound on work production of modular implementations in Eq. (14) is less than or equal to the global bound on integrated implementations in Eq. (15), since the local bound ignores correlations between the interacting system $\mathcal{Z}^{\text{int}}$ and noninteracting elements of the symbol string in $\mathcal{Z}^{\text{stat}}$. Critically, though, if we design the ratchet such that its states store the relevant correlations in the symbol string, then we can achieve the global bounds. This was hinted at in the fact that the gap between the work done by a ratchet and the global bound can be closed by designing a ratchet that matches the input process's structure [24]—the Principle of Requisite Complexity [23]. However, comparing the two bounds now allows us to be more precise.

The difference between the two bounds represents the amount of additional work that could have been performed by a ratchet, if it was not modular and limited to local

interactions. If the computational device is globally integrated, with full access to all correlations between the information-bearing degrees of freedom, then all of the nonequilibrium free energy can be converted to work, zeroing out the entropy production. Thus, the minimum entropy production for a modular transducer (or information ratchet) at the Nth time step can be expressed in terms of the difference between Eq. (14) and the entropic bounds in Eq. (15):

$$\begin{aligned}\frac{\langle \Sigma_N^{\text{mod}} \rangle_{\text{min}}}{k_B \ln 2} &= \frac{\langle W_N^{\text{local}} \rangle_{\text{min}} - \Delta F^{\text{neq}}_{N\tau \to (N+1)\tau}}{k_B T \ln 2} \\ &= I[Y_{N+1:\infty}, Y'_{0:N}; X_N, Y_N] \\ &\quad - I[Y_{N+1:\infty}, Y'_{0:N}; X_{N+1}, Y'_N] \\ &= I[Y_{N+1:\infty}, Y'_{0:N}; X_N, Y_N | X_{N+1}, Y'_N]. \end{aligned} \tag{17}$$

This can also be derived directly by substituting our interacting variables $(X_N, Y_N) = Z_t^i$ and $(X_{N+1}, Y'_N) = Z_{t+\tau}^i$ and stationary variables $(Y_{N+1:\infty}, Y'_{0:N}) = Z^s$ into the expression for the modularity dissipation in Eqs. (8) and (9) in Sec. II. Even if the energy levels are controlled so slowly that entropic bounds are reached, Eq. (17) quantifies the amount of lost correlations that cannot be recovered. Also, this leads to the entropy production and irreversibility of the transducing ratchet. This has immediate consequences that limit the most thermodynamically efficient information processors.

While previous bounds, such as the IPSL, demonstrated that information in the symbol string can be used as a thermodynamic fuel [20,21]—leveraging structure in the inputs symbols to turn thermal energy into useful work—they largely ignore the structure of information ratchet states X_N. The transducer's hidden states, which can naturally store information about the past, are critical to taking advantage of structured inputs. Until now, we only used informational bounds to predict transient costs to information processing [27,53]. With the expression for the modularity dissipation of information ratchets in Eq. (17), however, we now have bounds that apply to the ratchet's asymptotic functioning. In short, this provides the key tool for designing thermodynamically efficient transducers. We will now show that it has immediate implications for pattern generation and pattern extraction.

V. PREDICTIVE EXTRACTORS

A pattern extractor is a transducer that takes in a structured process $\Pr(Y_{0:\infty})$, with correlations among the symbols, and maps it to a series of independent identically distributed (IID), uncorrelated output symbols. An output symbol can be distributed however we wish individually, but each must be distributed with an identical distribution and independently from all others. The result is that the joint distribution of the output process symbols is the product of the individual marginals:

$$\Pr(Y'_{0:\infty}) = \prod_{i=0}^{\infty} \Pr(Y'_i). \tag{18}$$

If implemented efficiently, this device can use temporal correlations in the input as a thermodynamic resource to produce work. The modularity dissipation of an extractor $\langle \Sigma_N^{\text{ext}} \rangle_{\text{min}}$ can be simplified by noting that the output symbols are uncorrelated with any other variable and, thus, the Y' terms fall out of the mutual information expression for dissipation in Eq. (17), yielding

$$\frac{\langle \Sigma_N^{\text{ext}} \rangle_{\text{min}}}{k_B \ln 2} = I[Y_{N+1:\infty}; X_N, Y_N] - I[Y_{N+1:\infty}; X_{N+1}]. \tag{19}$$

Minimizing this irreversibility, as shown in Appendix B, leads directly to a fascinating conclusion that relates thermodynamics to prediction: The states of maximally thermodynamically efficient extractors are perfect predictors of the input process. Other work anticipates the need for predictive agents to leverage temporal correlations [24,63] and even discusses memoryful agents that can extract additional work from temporal correlations by using predictive states of the input [24,63,64]. Our development of modularity dissipation, however, provides the first proof of the need for predictive states. Moreover, it can be applied to any extractor to determine the dissipation of an imperfect predictor.

To take full advantage of the temporal structure of an input process, the ratchet's states X_N must be able to predict the future of the input $Y_{N:\infty}$ from the input past $Y_{0:N}$. Thus, the ratchet shields the input past from the output future such that there is no information shared between the past and future that is not captured by the ratchet's states:

$$\mathrm{I}[Y_{N:\infty}; Y_{0:N} | X_N] = 0. \tag{20}$$

Additionally, transducers cannot anticipate the future of the inputs beyond their correlations with past inputs [25]. This means that there is no information shared between the ratchet and the input future when conditioned on the input past:

$$\mathrm{I}[Y_{N:\infty}; X_N | Y_{0:N}] = 0. \tag{21}$$

As shown in Appendix B, Eqs. (20) and (21), which together are equivalent to the state X_N being predictive, can be used to prove that the modularity dissipation vanishes: $\langle \Sigma_N^{\text{ext}} \rangle_{\text{min}} = 0$. Moreover, setting the modularity dissipation to zero guarantees that the state shields the past input and the future input from each other, as shown in Eq. (20). Thus, since Eq. (21) is a given for transducers, this establishes that the ratchet's being predictive is equivalent

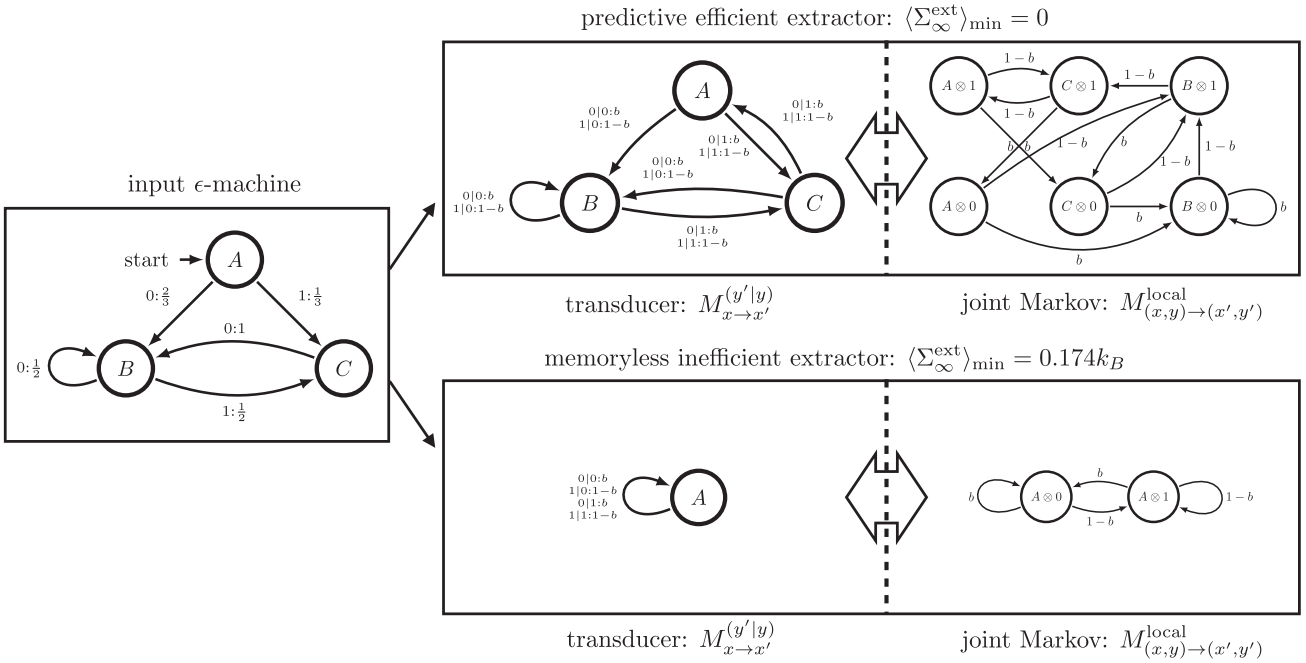

FIG. 4. Multiple ways to transform the golden mean process input, whose ϵ-machine generator is shown in the far left box, into a sequence of uncorrelated symbols. The ϵ-machine is a Mealy hidden Markov model that produces outputs along the edges, with $y:p$ denoting that the edge emits symbol y and is taken with probability p. Top row: Ratchet whose internal states match the ϵ-machine states and so it is able to minimize dissipation ($\langle\Sigma_\infty^{\text{ext}}\rangle_{\min} = 0$) by making transitions such that the ratchet's states are synchronized to the ϵ-machine's states. The transducer representation to the left shows how the states remain synchronized: Its edges are labeled $y'|y:p$, which means that, if the input is y, then the edge is taken with probability p and outputs y'. The joint Markov representation on the right depicts the corresponding physical dynamic over the joint state space of the ratchet and the interaction symbol. The label p along an edge from the state $x \otimes y$ to $x' \otimes y'$ specifies the probability of transitioning between those states according to the local Markov channel $M_{(x,y)\to(x',y')}^{\text{local}} = p$. Bottom row: In contrast to the efficient predictive ratchet, the memoryless ratchet shown is inefficient, since its memory cannot store the predictive information within the input ϵ-machine, much less synchronize to it.

to zero modularity dissipation and, thus, to perfect thermodynamic efficiency. The efficiency of predictive ratchets suggests that predictive generators, such as the ϵ-machine [62], are useful in designing efficient information engines that can leverage temporal structure in an environment.

Consider, for example, an input string that is structured according to the golden mean process, which consists of binary strings in which 1s always occur in isolation, surrounded by 0s. Figure 4 gives two examples of ratchets, described by different local Markov channels $M_{(x,y)\to(x',y')}^{\text{local}}$, that each map the golden mean process to a biased coin. The input process's ϵ-machine, shown in the left box, provides a template for how to design a thermodynamically efficient local Markov channel, since its states are predictive of the process. The Markov channel is a transducer [21]:

$$M_{x\to x'}^{(y'|y)} \equiv M_{(x,y)\to(x',y')}^{\text{local}}. \tag{22}$$

By designing transducer states that stay synchronized to the states of the input process's ϵ-machine, we minimize the modularity dissipation to zero. For example, the efficient transducer shown in Fig. 4 has almost the same topology as the golden mean ϵ-machine, with an added transition between states C and A corresponding to a disallowed word in the input. This transducer is able to harness all structure in the input since it synchronizes to the input process and so is able to optimally predict the next input.

The efficient ratchet shown in Fig. 4 (top row) comes from a general method for constructing an optimal extractor given the input's ϵ-machine. The ϵ-machine is represented by a Mealy hidden Markov model (HMM) [65] with the symbol-labeled state-transition matrices:

$$T_{s\to s'}^{(y)} = \Pr(Y_N = y, S_{N+1} = s'|S_N = s), \tag{23}$$

where S_N is the random variable for the hidden state reading the Nth input Y_N. If we design the ratchet to have the same state space as the input process's hidden state space ($\mathcal{X} = \mathcal{S}$), and if we want the IID output to have bias $\Pr(Y_N = 0) = b$, then we set the local Markov channel over the ratchet and interaction symbol to be

$$M^{\text{local}}_{(x,y)\to(x',y')} = \begin{cases} b, & \text{if } T^{(y)}_{x\to x'} \neq 0 \quad \text{and} \quad y' = 0 \\ 1-b, & \text{if } T^{(y)}_{x\to x'} \neq 0 \quad \text{and} \quad y' = 1. \end{cases}$$

This channel, combined with normalized transition probabilities, does not uniquely specify M^{local}, since there can be forbidden words in the input that, in turn, lead to ϵ-machine causal states that always emit a single symbol. This means that there are joint ratchet-symbol states (x, y) such that $M_{(x,y)\to(x',y')}$ is unconstrained. For these states, we may make any choice of transition probabilities from (x, y), since this state will never be reached by the combined dynamics of the input and ratchet. The end result is that, with this design strategy, we construct a ratchet whose memory stores all information in the input past that is relevant to the future, since the ratchet remains synchronized to the input's causal states.

In this way, the ratchet leverages all temporal order in the input. This is characteristic of any efficient extractor and confirms the thermodynamic Principle of Requisite Variety [23]. The fact that the ratchet states must synchronize to the ϵ-machine's causal states implies that the uncertainty in the ratchet's memory must at least match the uncertainty in the causal states of the input, which is its statistical complexity:

$$H[X_N] \geq H[S_N] \tag{24}$$

$$= C_\mu. \tag{25}$$

Thus, this not only proves the thermodynamic Principle of Requisite Variety in general, but also refines it to a Principle of Requisite Complexity—the structure of a thermodynamically efficient ratchet must match that of the environment.

By way of contrast, consider a memoryless transducer, such as that shown in Fig. 4 (bottom row). It has only a single state and so cannot store any information about the input past. As discussed in previous explorations, ratchets without memory are insensitive to correlations [23,24]. This result for stationary input processes is subsumed by the measure of modularity dissipation. Since there is no uncertainty in X_N, the asymptotic dissipation of memoryless ratchets simplifies to

$$\langle \Sigma^{\text{ext}}_\infty \rangle_{\text{min}} = \lim_{N\to\infty} k_B \ln 2 I[Y_{N+1:\infty}; Y_N] = k_B \ln 2 (H_1 - h_\mu),$$

where in the second step we used input stationarity—every symbol has the same marginal distribution—and so the same single-symbol uncertainty $H_1 = H[Y_N] = H[Y_M]$. Thus, the modularity dissipation of a memoryless ratchet is proportional to the length-1 redundancy $H_1 - h_\mu$ [62]. This is the amount of additional uncertainty that comes from ignoring temporal correlations.

As Fig. 4 shows, this means that a memoryless extractor driven by the golden mean process asymptotically dissipates an average of $\langle \Sigma^{\text{ext}}_\infty \rangle_{\text{min}} \approx 0.174 k_B$ per input symbol. This is in stark contrast, for example, with the claim in Ref. [46] that "unwarranted retention of past information is fundamentally equivalent to energetic inefficiency," since such a memoryless ratchet minimizes the instantaneous nonpredictive information—the measure of dissipation in a driven system [46].

Moreover, we can split the states of the predictive model shown in Fig. 4 to introduce duplicates that have the same histories and same future distributions, such that the states are still predictive of the input. This larger machine, with duplicate states, is still predictive and maximally efficient. This is a further conflict with Ref. [46]. Despite the fact that both of these ratchets perform the same computational process—converting the golden mean process into a sequence of IID symbols—the simpler model requires more energy investment to function, because of its irreversibility.

VI. RETRODICTIVE GENERATORS

Pattern generators are rather like time-reversed pattern extractors, in that they take in an uncorrelated input process,

$$\Pr(Y_{0:\infty}) = \prod_{i=0}^{\infty} \Pr(Y_i), \tag{26}$$

and turn it into a structured output process $\Pr(Y'_{0:\infty})$ that has correlations among the symbols. The modularity dissipation of a generator $\langle \Sigma^{\text{gen}}_N \rangle_{\text{min}}$ can also be simplified by removing the uncorrelated input symbols:

$$\frac{\langle \Sigma^{\text{gen}}_N \rangle_{\text{min}}}{k_B \ln 2} = I[Y'_{0:N}; X_N] - I[Y'_{0:N}; X_{N+1} Y'_N].$$

Paralleling extractors, Appendix B shows that retrodictive ratchets minimize the modularity dissipation to zero.

Retrodictive generator states carry as little information about the output past as possible. Since this ratchet generates the output, it must carry all the information shared between the output past and future. Thus, it shields output past from output future just as a predictive extractor does for the input process:

$$I[Y'_{N:\infty}; Y'_{0:N} | X_N] = 0.$$

However, unlike the predictive states, the output future shields the retrodictive ratchet state from the output past:

$$I[X_N; Y'_{0:N} | Y'_{N:\infty}] = 0. \tag{27}$$

These two conditions mean that X_N is retrodictive and imply that the modularity dissipation vanishes. While we have not established the equivalence of retrodictiveness and efficiency for pattern generators, as we have for predictive pattern extractors, there are easy-to-construct examples

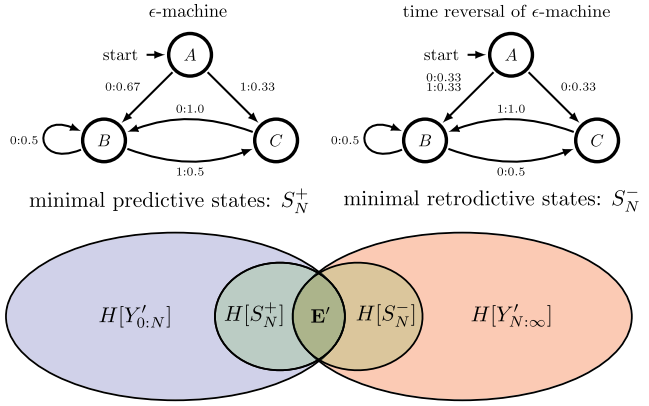

FIG. 5. Alternate minimal generators of the golden mean process: predictive and retrodictive. (Left) The ϵ-machine has the minimal set of causal states S^+ required to predictively generate the output process. As a result, the uncertainty $H[S_N^+]$ is contained by the uncertainty $H[Y'_{0:N}]$ in the output past. (Right) The time reversal of the reverse-time ϵ-machine has the minimal set of states required to retrodictively generate the output. Its states are a function of the output future. Thus, its uncertainty $H[S_N^-]$ is contained by the output future's uncertainty $H[Y'_{N:\infty}]$.

demonstrating that diverging from efficient retrodictive implementations leads to modularity dissipation at every step.

Consider once again the golden mean process. Figure 5 shows that there are alternate ways to generate such a process from a hidden Markov model. The ϵ-machine, shown on the left, is the minimal predictive model, as discussed earlier. It is unifilar, which means that the current hidden state S_N^+ and current output Y'_N uniquely determine the next hidden state S_{N+1}^+ and that, once synchronized to the hidden states, one stays synchronized to them by observing only output symbols. Thus, its states are a function of past outputs. This is illustrated in Fig. 5 by the fact that the information atom $H[S_N^+]$ is contained by the information atom for the output past $H[Y'_{0:N}]$.

The other hidden Markov model generator shown in Fig. 5 (right) is the time reversal of the ϵ-machine that generates the reverse process. This is much like the ϵ-machine, except that it is retrodictive instead of predictive. The recurrent states B and C are counifilar as opposed to unifilar. This means that the next hidden state S_{N+1}^- and the current output Y'_N uniquely determine the current state S_N^-. The hidden states of this minimal retrodictive model are a function of the semi-infinite future. Also, this can be seen from the fact that the information atom for $H[S_N^-]$ is contained by the information atom for the future $H[Y'_{N:\infty}]$.

These two different hidden Markov generators both produce the golden mean process, and they provide a template for constructing ratchets to generate that process. For a hidden Markov model described by a symbol-labeled transition matrix $\{T^{(y)}\}$, with hidden states in $\mathcal{S}$ as described in Eq. (23), the analogous generative ratchet has the same states $\mathcal{X} = \mathcal{S}$ and is described by the joint Markov local interaction:

$$M^{\text{local}}_{(x,y)\to(x',y')} = T^{(y')}_{x\to x'}.$$

Such a ratchet effectively ignores the IID input process and obeys the same informational relationships between the ratchet states and outputs as the hidden states of hidden Markov model with its outputs.

Figure 6 shows both the transducer and joint Markov representation of the minimal predictive generator and minimal retrodictive generator. The retrodictive generator is potentially perfectly efficient, since the process's minimal modularity dissipation vanishes: $\langle \Sigma_N^{\text{gen}} \rangle_{\text{min}} = 0$ for all N.

However, despite being a standard tool for generating an output, the predictive ϵ-machine is necessarily irreversible and dissipative. The ϵ-machine-based ratchet, as shown in Fig. 6 (bottom row), approaches an asymptotic dynamic where the current state X_N stores more than it needs to about the output past $Y'_{0:N}$ in order to generate the future $Y'_{N:\infty}$. As a result, it irretrievably dissipates:

$$\begin{aligned}\langle \Sigma_\infty^{\text{gen}} \rangle_{\text{min}} &= k_B \ln 2 \lim_{N\to\infty} (I[Y'_{0:N}; X_N] - I[Y'_{0:N}; X_{N+1}, Y'_N]) \\ &= \frac{2}{3} k_B \ln 2 \\ &\approx 0.462\ k_B.\end{aligned}$$

This can be calculated with greater ease by noting that X_N and X_{N+1} are predictive functions of their output past. That is, all information in the current ratchet state is shared with the past $I[Y'_{0:N}; X_N] = H[X_N]$, and all future behavior that is predictable from the output past is also predictable from the ratchet state; so, $I[Y'_{0:N}; X_{N+1}, Y'_N] = I[X_N; X_{N+1}, Y'_N]$. These latter can both be calculated directly from the ϵ-machine symbol-labeled transition matrices $T^{(y_N)}_{x_N \to x_{N+1}} = \Pr(Y_N = y_N, X_{N+1} = x_{N+1} | X_N = x_N)$, which give

$$\begin{aligned}\lim_{N\to\infty} \Pr(Y_N = 0, X_{N+1} = B, X_N = B) &= \frac{1}{3} \\ \lim_{N\to\infty} \Pr(Y_N = 1, X_{N+1} = C, X_N = B) &= \frac{1}{3} \\ \lim_{N\to\infty} \Pr(Y_N = 0, X_{N+1} = B, X_N = C) &= \frac{1}{3},\end{aligned}$$

and, consequently, we see that

$$\begin{aligned}&\lim_{N\to\infty} (H[X_N] - I[X_N; X_{N+1}, Y'_N]) \\ &\quad = \lim_{N\to\infty} H[X_N | X_{N+1}, Y'_N] \\ &\quad = 2/3.\end{aligned}$$

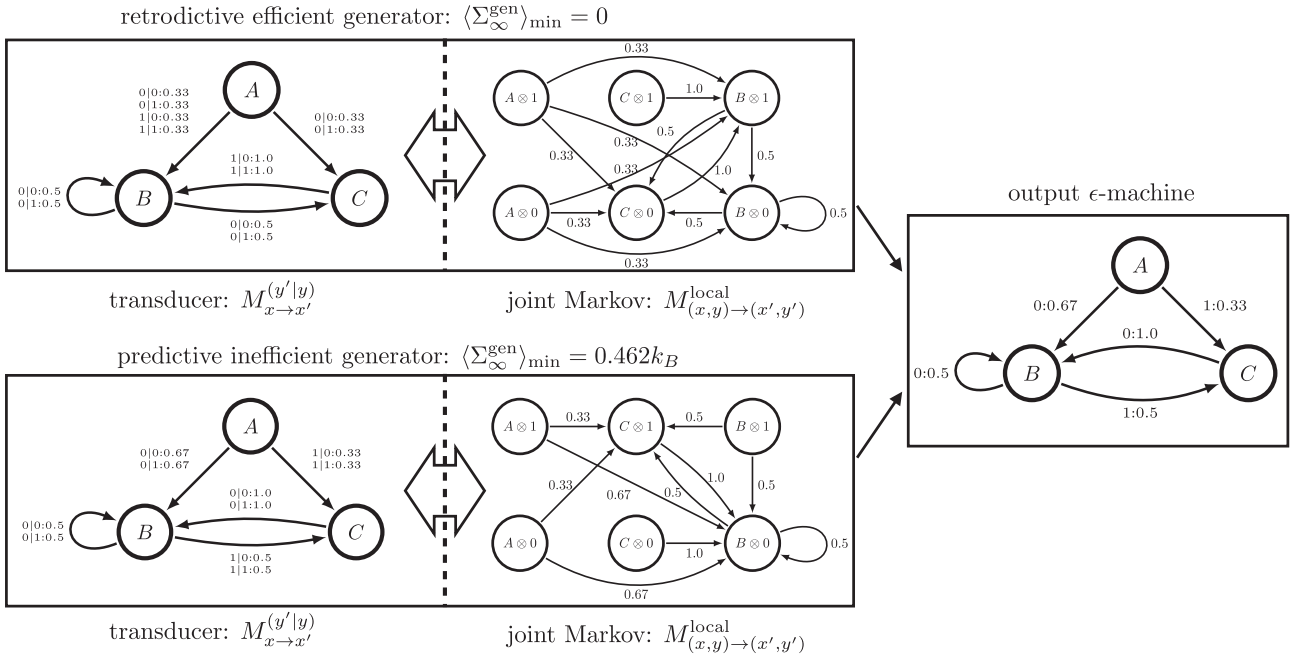

FIG. 6. Alternative generators of the golden mean process. Right: The process's ϵ-machine. Top row: Optimal generator designed using the topology of the minimal retrodictive generator. It is efficient, since it stores as little information about the past as possible, while still storing enough to generate the output. Bottom row: The predictive generator stores far more information about the past than necessary, since it is based off the predictive ϵ-machine. As a result, it is far less efficient. It dissipates at least $\frac{2}{3}k_BT\ln 2$ extra heat per symbol and requires that much more work energy per symbol emitted.

Thus, with every time step, this predictive ratchet stores information about its past, but it also erases information, dissipating $2/3$ of a bit worth of correlations without leveraging them. Those correlations could have been used to reverse the process if they had been turned into work. They are used by the retrodictive ratchet, though, which stores just enough information about its past to generate the future.

It was previously shown that storing unnecessary information about the past leads to additional transient dissipation when generating a pattern [27,53]. This cost also arises from implementation. However, our measure of modularity dissipation shows that there are implementation costs that persist through time. The two locally operating generators of the golden mean process perform the same computation, but have different bounds on their dissipation per time step. Thus, the additional work investment required to generate the process grows linearly with time for the ϵ-machine implementation, but is zero for the retrodictive implementation.

Moreover, we can consider generators that fall in between these extremes using the parametrized HMM shown in Fig. 7 (top). This HMM, parametrized by z, produces the golden mean process at all $z \in [.5, 1]$, but the hidden states share less and less information with the output past as z increases, as shown by Ref. [36]. The extreme at $z = 0.5$ corresponds to the minimal predictive generator, the ϵ-machine. The other at $z = 1$ corresponds to the minimal retrodictive generator, the time reversal of the reverse-time ϵ-machine. The graph there plots the modularity dissipation as a function of z.

The modularity dissipation decreases with z, suggesting that the unnecessary memory of the past leads to additional dissipation. This echoes the claim that "unwarranted retention of past information is fundamentally equivalent to energetic inefficiency" in the particular context of pattern generation [46]. So, while we have only proved that retrodictive generators are maximally efficient, this demonstrates that extending beyond that class can lead to unnecessary dissipation and that there may be a direct relationship between unnecessary memory and dissipation.

Taken altogether, we see that the thermodynamic consequences of localized information processing lead to direct principles for efficient information transduction. Analyzing the most general case of transducing arbitrary structured processes into other arbitrary structured processes remains a challenge. That said, pattern generators and pattern extractors have elegantly symmetric conditions for efficiency that give insight into the range of possibilities. Pattern generators are effectively the time reversal of pattern extractors, which turn structured inputs into structureless outputs. As such, they are most efficient when retrodictive, which is the time reversal of being predictive. Figure 5 illustrates graphically how the predictive ϵ-machine captures past correlations and stores the necessary information about the past, while the retrodictive ratchet's states are

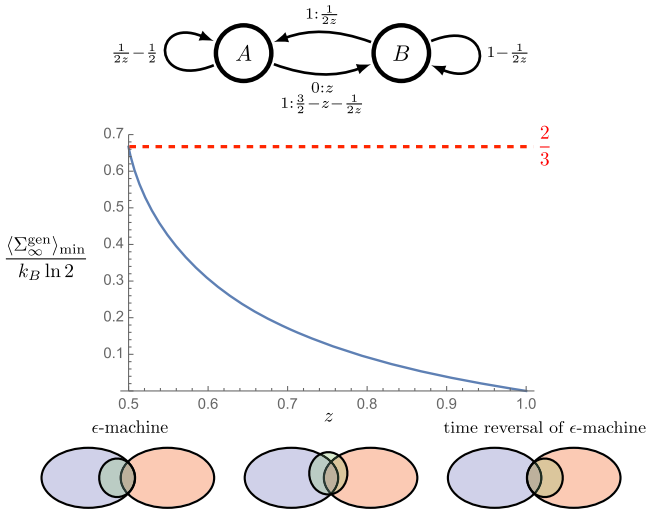

FIG. 7. Top: A parametrized family of HMMs that generate the golden mean process for $z \in [.5, 1]$. Middle: As parameter z increases, the information stored in the hidden states about the output past decreases. At $z = 0.5$, the HMM is the ϵ-machine, whose states are a function of the past. At $z = 1.0$, the HMM is the time reversal of the reverse-time ϵ-machine, whose states are a function of the future. The modularity dissipation decreases monotonically as z increases and the hidden states' memory of the past decreases. Bottom: Information diagrams corresponding to the end cases and a middle case. Labeling is the same as in Fig. 5.

analogous, but store information about the future instead. This may seem unphysical—as if the ratchet is anticipating the future. However, since the ratchet generates the output future, this anticipation is entirely physical, because the ratchet controls the future, as opposed to mysteriously predicting it, as an oracle would.

VII. CONCLUSION

Modularity is a key design theme in physical information processing, since it gives the flexibility to stitch together many elementary logical operations to implement a much larger computation. Any classical computation can be composed from local operations on a subset of information reservoir observables. Modularity is also key to biological organization, its functioning, and our understanding of these [5].

However, there is an irretrievable thermodynamic cost, the modularity dissipation, to this localized computing, which we quantified in terms of the global entropy production. This modularity-induced entropy production is proportional to the reduction of global correlations between the local and interacting portion of the information reservoir and the fixed, noninteracting portion. This measure forms the basis for designing thermodynamically efficient information processing. It is proportional to the additional work investment required by the modular form of the computation, beyond the work required by a globally integrated and reversible computation.

While our main result on modularity dissipation might be viewed as a cousin of Landauer's principle, it is very different—an essential but complementary principle. To close, we should remove any lingering confusion on this score, by contrasting the microscopic mechanisms underlying each.

Recall that Landauer's principle identifies an inescapable dissipation that arises from the collapse of microscopic state-space volume as the "information-bearing degrees of freedom" implement erasing a bit of mesoscopically stored information. This follows directly from Liouville's theorem that guarantees state-space volume conservation: If the mesoscale operation collapses state space, then the surrounding environment's state space must expand, resulting in a transfer of entropy and so dissipation.

The thermodynamic costs due to modularity, in contrast, arise from state-space componentwise organization—technically, the conditional-independence structure of the microscopic state space—used to implement a given information processing operation. Since modularity removes systemwide correlations, one throws away a thermodynamic resource.

Thus, there is indeed a parallel between Landauer's principle and modularity dissipation, as they together identify thermodynamic costs of information processing. The similarity ends there, though. Modularity dissipation arises from a completely different mechanism from Landauer's—one that is also dissipative and also leads to irreducible entropy production. This is why modularity dissipation is a distinct and essential mechanism in a full accounting of the thermodynamic costs of information processing. One concludes that Landauer's principle is incomplete; the fuller theory requires both it and modularity dissipation.

Turing-machine-like information ratchets provide a natural application for this new measure of efficient information processing, since they process information in a symbol string through a sequence of local operations. The modularity dissipation allows us to determine which implementations are able to achieve the asymptotic bound set by the IPSL, which, substantially generalizing Landauer's bound, says that any type of structure in the input can be used as a thermodynamic resource and any structure in the output has a thermodynamic cost. There are many different ratchet implementations that perform a given computation, in that they map inputs to outputs in the same way. However, if we want an implementation to be thermodynamically efficient, the modularity dissipation, monitored by the global entropy production, must be minimized. Conversely, we now appreciate why there are many implementations that dissipate and are thus irreversible. This establishes modularity dissipation as a new thermodynamic cost, due purely to an implementation's architecture, that complements Landauer's bound on isolated logical operations.

We noted that there are not yet general principles for designing devices that minimize modularity dissipation and thus minimize work investment for arbitrary information transduction. However, for the particular cases of pattern generation and pattern extraction, we find that there are prescribed classes of ratchets that are guaranteed to be dissipationless, if operated isothermally. These devices' ratchet states are able to store and leverage the global correlations among the symbol strings. This means, in turn, that it is possible to achieve the reversibility of globally integrated information processing but with modular computational design. Thus, while modular computation often results in dissipating global correlations, this inefficiency can be avoided when designing processors using the computational-mechanics tools outlined here.

ACKNOWLEDGMENTS

As an External Faculty member, J. P. C. thanks the Santa Fe Institute for its hospitality during visits. This material is based upon work supported by, or in part by, John Templeton Foundation Grant No. 52095, Foundational Questions Institute Grant No. FQXi-RFP-1609, and the U.S. Army Research Laboratory and the U.S. Army Research Office under Contract No. W911NF-13-1-0390.

APPENDIX A: ISOTHERMAL MARKOV CHANNELS

To meet the information-theoretic bounds on work dissipation, we describe an isothermal channel in which we change system energies in slow steps to manipulate the distribution over $\mathcal{Z}$'s states. The challenge in this is to evolve an input distribution $\Pr(Z_t = z_t)$ over the time interval $(t, t+\tau)$ according to the Markov channel M, so that the system's conditional probability at time $t+\tau$ is

$$\Pr(Z_{t+\tau} = z_{t+\tau} | Z_t = z_t) = M_{z_t \to z_{t+\tau}}.$$

The Markov channel M specifies the form of the computation, as it probabilistically maps inputs to outputs. While we need not know the input distribution $\Pr(Z_t = z_t)$ to implement a computation, this is necessary to design a thermodynamically efficient computation. Making this as efficient as possible in a thermal environment at temperature T means ensuring that the work invested in the evolution achieves the lower bound:

$$\langle W \rangle \geq k_B T \ln 2 (H[Z_t] - H[Z_{t+\tau}]).$$

This expresses the second law of thermodynamics for the system in contact with a heat bath.

To ensure the appropriate conditional distribution, we introduce an ancillary system $\mathcal{Z}'$, which is a copy of $\mathcal{Z}$, as proposed in Ref. [27]. This is necessary since continuous-time Markov chains—the probabilistic rate equations underlying stochastic thermodynamics—have restrictions on the logical functions they can implement. Some logical functions, such as flipping a bit ($0 \to 1$ and $1 \to 0$), must be implemented with ancillary or hidden states [66]. Including an ancillary system that is a copy of $\mathcal{Z}$ allows implementing any probabilistic channel $M_{z_t \to z_{t+\tau}}$ and, thus, any logical operation on $\mathcal{Z}$.

For efficiency, we take τ to be large with respect to the system's relaxation time scale and break the overall process into three steps that occur over the time intervals $(t, t+\tau_0)$; $(t+\tau_0, t+\tau_1)$; and $(t+\tau_1, t+\tau)$, where $0 < \tau_0 < \tau_1 < \tau$.

Our method of manipulating $\mathcal{Z}$ and $\mathcal{Z}'$ is to control the energy $E(t, z, z')$ of the joint state $z \otimes z' \in \mathcal{Z} \otimes \mathcal{Z}'$ at time t. We also control whether or not probability is allowed to flow in $\mathcal{Z}$ or $\mathcal{Z}'$. This corresponds to raising or lowering energy barriers between system states.

At the beginning of the control protocol, we choose $\mathcal{Z}'$ to be in a uniform distribution uncorrelated with $\mathcal{Z}$. This means the joint distribution can be expressed as

$$\Pr(Z_t = z_t, Z'_t = z'_t) = \frac{\Pr(Z_t = z_t)}{|\mathcal{Z}'|}. \tag{A1}$$

Since we are manipulating an energetically mute information reservoir, we also start with the system in a uniformly zero-energy state over the joint states of $\mathcal{Z}$ and $\mathcal{Z}'$:

$$E(t, z, z') = 0. \tag{A2}$$

While this energy and the distribution change when executing the protocol, we return $\mathcal{Z}'$ to the independent uniform distribution and the energy to zero at the end of the protocol. This means that the starting and ending distributions are typically out of equilibrium. However, since we limit the flow between informational states, they are metastable and do not relax to the uniform equilibrium distribution. In this way, the information reservoir reliably stores and processes many different nonequilibrium states.

The three evolution steps that isothermally implement the Markov channel M are as follows:

(1) Over the time interval $(t, t+\tau_0)$, continuously change the energy such that the energy at the end of the interval $E(t+\tau_0, z, z')$ obeys the relation

$$e^{-(E(t+\tau_0, z, z') - F(t+\tau_0))/k_B T} = \Pr(Z_t = z) M_{z \to z'},$$

while allowing state space and probability to flow in $\mathcal{Z}'$, but not in $\mathcal{Z}$. Since the protocol is done slowly, $\mathcal{Z}'$ follows the Boltzmann distribution and, at time $t+\tau_0$, the distribution over $\mathcal{Z} \otimes \mathcal{Z}'$ is

$$\Pr(Z_{t+\tau_0} = z, Z'_{t+\tau_0} = z') = \Pr(Z_t = z) M_{z \to z'}.$$

This yields the conditional distribution of the current ancillary variable $Z'_{t+\tau}$ on the initial system variable Z_t:

$$\Pr(Z'_{t+\tau_0} = z' | Z_t = z) = M_{z \to z'},$$

since the system variable Z_t remains fixed over the interval. This protocol effectively applies the Markov channel M to evolve from $\mathcal{Z}$ to $\mathcal{Z}'$. However, we want the Markov channel to apply strictly to $\mathcal{Z}$.

Since the protocol is slow and isothermal, there is no entropy production, and the work flow is simply the change in nonequilibrium free energy:

$$\langle W_1 \rangle = \Delta F^{\text{neq}} = \Delta \langle E \rangle - T \Delta S[Z, Z'],$$

where $S[X] = -k_B \sum_{x \in \mathcal{X}} \Pr(X = x) \ln \Pr(X = x)$ is the thermodynamic entropy, which is proportional to the Shannon information $S[X] = k_B \ln 2 H[X]$. Since the average initial energy is uniformly zero, the change in average energy is the average energy at time $t + \tau_0$. So, we can express the work done:

$$\begin{aligned} \langle W_1 \rangle &= \langle E(t + \tau_0) \rangle - T \Delta S[Z, Z'] \\ &= \langle E(t + \tau_0) \rangle + k_B T \ln 2 (H[Z_t, Z'_t] \\ &\quad - H[Z_{t+\tau_0}, Z'_{t+\tau_0}]). \end{aligned}$$

(2) Now, swap the states of $\mathcal{Z}$ and $\mathcal{Z}'$ over the time interval $(t + \tau_0, t + \tau_1)$. This is logically reversible. Thus, it can be done without any work investment over the second time interval:

$$\langle W_2 \rangle = 0. \tag{A3}$$

The result is that the energies and probability distributions are flipped with regard to exchange of the system $\mathcal{Z}$ and ancillary system $\mathcal{Z}'$:

$$\begin{aligned} E(t + \tau_1, z, z') &= E(t + \tau_0, z', z) \\ \Pr(Z_{t+\tau_1} = z, Z'_{t+\tau_1} = z') &= \Pr(Z_{t+\tau_0} = z', Z'_{t+\tau_0} = z). \end{aligned}$$

Most importantly, however, this means that the conditional probability of the current system variable is given by M:

$$\Pr(Z_{t+\tau_1} = z' | Z_t = z) = \Pr(Z'_{t+\tau_0} = z' | Z_t = z) = M_{z \to z'}.$$

The ancillary system must still be reset to a uniform and uncorrelated state and the energies must be reset.

(3) Finally, we again hold $\mathcal{Z}$'s state fixed while allowing $\mathcal{Z}'$ to change over the time interval $(t + \tau_1, t + \tau)$ as we change the energy, ending at $E(t + \tau, z, z') = 0$. This isothermally brings the joint distribution to one where the ancillary system is uniform and independent of $\mathcal{Z}$:

$$\Pr(Z_{t+\tau} = z, Z'_{t+\tau} = z') = \frac{\Pr(Z_{t+\tau} = z)}{|\mathcal{Z}'|}. \tag{A4}$$

Again, the invested work is the change in average energy plus the change in thermodynamic entropy of the joint system:

$$\begin{aligned} \langle W_3 \rangle &= \langle \Delta E \rangle + k_B T \ln 2 (H[Z_{t+\tau_1}, Z'_{t+\tau_1}] \\ &\quad - H[Z_{t+\tau}, Z'_{t+\tau}]) \\ &= -\langle E(t + \tau_1) \rangle + k_B T \ln 2 (H[Z_{t+\tau_1}, Z'_{t+\tau_1}] \\ &\quad - H[Z_{t+\tau}, Z'_{t+\tau}]). \end{aligned}$$

This ends this three-step protocol.

Summing up the work terms gives the total dissipation:

$$\begin{aligned} \langle W_{\text{total}} \rangle &= \langle W_t \rangle + \langle W_2 \rangle + \langle W_3 \rangle \\ &= k_B T \ln 2 (H[Z_t, Z'_t] - H[Z_{t+\tau_0}, Z'_{t+\tau_0}]) \\ &\quad + k_B T (H[Z_{t+\tau_1}, Z'_{t+\tau_1}] - H[Z_{t+\tau}, Z'_{t+\tau}]) \\ &\quad + \langle E(t + \tau_0) \rangle - \langle E(t + \tau_1) \rangle. \end{aligned}$$

Recall that the distributions $\Pr(Z_{t+\tau_1}, Z'_{t+\tau_1})$ and $\Pr(Z_{t+\tau_0}, Z'_{t+\tau_0})$, as well as $E(t + \tau_0, z, z')$ and $E(t + \tau_1, z, z')$, are identical under exchange of $\mathcal{Z}$ and $\mathcal{Z}'$, so $H[Z_{t+\tau_1}, Z'_{t+\tau_1}] = H[Z_{t+\tau_0}, Z'_{t+\tau_0}]$ and $\langle E(t + \tau_0) \rangle = \langle E(t + \tau_1) \rangle$. Additionally, we know that both the starting and ending distributions have a uniform and uncorrelated ancillary system, so their entropies can be expressed:

$$H[Z_t, Z'_t] = H[Z_t] + \log_2 |\mathcal{Z}'| \tag{A5}$$

$$H[Z_{t+\tau}, Z'_{t+\tau}] = H[Z_{t+\tau}] + \log_2 |\mathcal{Z}'|. \tag{A6}$$

Substituting this into the above expression for total invested work, we find that we achieve the lower bound with this protocol:

$$\langle W_{\text{total}} \rangle = k_B T \ln 2 (H[Z_t] - H[Z_{t+\tau}]). \tag{A7}$$

Thus, the protocol implements a Markov channel that achieves the information-theoretic bounds. It is similar to that described in Ref. [27].

The basic principle underlying the thermodynamic efficiency of this protocol is that, when manipulating system energies to change state space, we choose the energies so that there is no instantaneous probability flow. That is, if one interrupts the protocol and holds the energy landscape fixed, the distribution will not continue to change. If it did, this change would correspond to relaxation to equilibrium, dissipation of nonequilibrium free energy, and thus, an increase in the Universe's entropy. By guiding the distribution via an energy landscape such that the system remains stationary if the protocol is stopped, we

are able to achieve the information-theoretic bounds set by the second law of thermodynamics in Eq. (A7). However, there are situations in which it is impossible to prevent instantaneous flow, even when slowly manipulating the energies, due to limits of control imposed by the system, such as in the case of local control. Then, there are necessarily inefficiencies that arise from the dissipation of the distribution evolving out of equilibrium.

APPENDIX B: TRANSDUCER DISSIPATION

1. Predictive extractors

For a pattern extractor, being reversible means that the transducer is predictive of the input process. More precisely, an extracting transducer that produces zero entropy is equivalent to it being a predictor of its input.

As shown earlier in Eq. (19), a reversible extractor satisfies

$$I[Y_{N+1:\infty}; X_{N+1}] = I[Y_{N+1:\infty}; X_N Y_N],$$

for all N, since it must be reversible at every step to be fully reversible. The physical ratchet being predictive of the input means two things. It means that X_N shields the past $Y_{0:N}$ from the future $Y_{N:\infty}$. This is equivalent to the mutual information between the past and future vanishing when conditioned on the ratchet state:

$$I[Y_{0:N}; Y_{N:\infty}|X_N] = 0.$$

Note that this also implies that any subset of the past or future is independent of any other subset conditioned on the ratchet state:

$$I[Y_{a:b}; Y_{c:d}|X_N] = 0, \quad \text{where } b \leq N \quad \text{and} \quad c \geq N.$$

The other feature of a predictive transducer is that the past shields the ratchet state from the future:

$$I[X_N; Y_{N:\infty}|Y_{0:N}] = 0.$$

This is guaranteed by the fact that transducers are nonanticipatory: They cannot predict future inputs outside of their correlations with past inputs.

We start by showing that, if the ratchet is predictive, then the entropy production vanishes. It is useful to note that for transducers, which are nonanticipatory of their input, being predictive is equivalent to storing as much information about the future as is predictable from the past:

$$I[X_N; Y_{N:\infty}] = I[Y_{0:N}; Y_{N:\infty}],$$

which can be seen by subtracting $I[Y_{0:N}; Y_{N:\infty}; X_N]$ from each side of the immediately preceding expression. Thus, it is sufficient to show that the mutual information between the partial input future $Y_{N+1:\infty}$ and the joint distribution of the predictive variable X_N and next output Y_N is the same as mutual information with the joint variable $(Y_{0:N}, Y_N) = Y_{0:N+1}$ of the past inputs and the next input:

$$I[Y_{N+1:\infty}; X_N, Y_N] = I[Y_{N+1:\infty}; Y_{0:N}, Y_N].$$

To show this for a predictive variable, we use Fig. 8, which displays the information diagram for all four variables with the information atoms of interest labeled.

Assuming that X_N is predictive zeros out a number of information atoms, as shown below:

$$\begin{aligned} I[X_N; Y_N, Y_{N+1:\infty}|Y_{0:N}] &= b + c + h = 0 \\ I[X_N; Y_N|Y_{0:N}] &= b + h = 0 \\ I[Y_{0:N}; Y_N, Y_{N+1:\infty}|X_N] &= i + f + g = 0 \\ I[Y_{0:N}; Y_N|X_N] &= i + f = 0. \end{aligned}$$

These four equations make it clear that $g = c = 0$. Thus, substituting $I[Y_{N+1:\infty}; X_N, Y_N] = a + b + c + d + e + f$ and $I[Y_{N+1:\infty}; Y_{0:N}, Y_N] = a + b + d + e + f + g$, we find that their difference vanishes:

$$I[Y_{N+1:\infty}; X_N, Y_N] - I[Y_{N+1:\infty}; Y_{0:N}, Y_N] = c - g = 0.$$

There is zero dissipation, since X_{N+1} is also predictive, meaning $I[Y_{N+1:\infty}; Y_{0:N}, Y_N] = I[Y_{N+1:\infty}; X_{N+1}]$, so

$$\begin{aligned} \frac{\langle \Sigma_N^{\text{ext}} \rangle_{\min}}{k_B T \ln 2} &= I[Y_{N+1:\infty}; X_N, Y_N] - I[Y_{N+1:\infty}; X_{N+1}] \\ &= I[Y_{N+1:\infty}; X_N, Y_N] - I[Y_{N+1:\infty}; Y_{0:N+1}] \\ &= 0. \end{aligned}$$

Going the other direction, using zero entropy production to prove that X_N is predictive for all N is now simple.

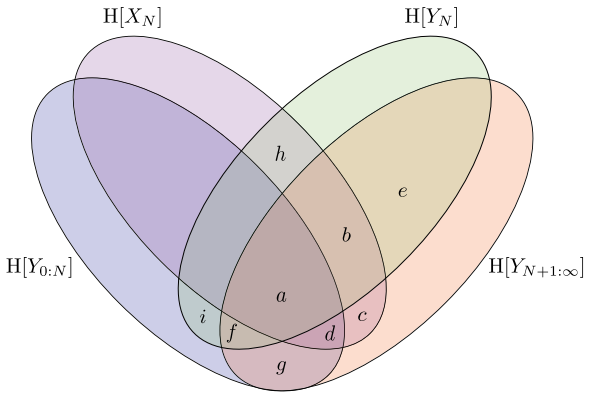

FIG. 8. Information diagram for dependencies between the input past $Y_{0:N}$, next input Y_N, current ratchet state X_N, and input future $Y_{N+1:\infty}$, excluding the next input. We label certain information atoms to help illustrate the algebraic steps in the associated proof.

We already showed that $I[Y_{N+1:\infty}; X_N, Y_N] = I[Y_{N+1:\infty}; Y_{0:N}, Y_N]$ if X_N is predictive. Combining with zero entropy production ($I[Y_{N+1:\infty};X_{N+1}]=I[Y_{N+1:\infty};X_N,Y_N]$) immediately implies that X_{N+1} is predictive, since $I[Y_{N+1:\infty}; X_{N+1}] = I[Y_{N+1:\infty}; Y_{0:N}, Y_N]$; plus, the fact that X_{N+1} is nonanticipatory is equivalent to X_{N+1} being predictive.

With this recursive relation, all that is left to establish is the base case, that X_0 is predictive. Applying zero entropy production again, we find the relation necessary for prediction:

$$I[Y_{1:\infty}; X_1] = I[Y_{1:\infty}; X_0, Y_0] = I[Y_{1:\infty}; Y_0].$$

From this, we find the equivalence $I[Y_{1:\infty}; Y_0] = I[Y_{1:\infty}; X_0, Y_0]$, since X_0 is independent of all inputs, due to it being nonanticipatory. Thus, zero entropy production is equivalent to predictive ratchets for pattern extractors.

2. Retrodictive generators

An analogous argument can be made to show the relationship between retrodiction and zero entropy production for pattern generators, which are essentially time-reversed extractors.

Efficient pattern generators must satisfy

$$I[Y'_{0:N}; X_N] = I[Y'_{0:N}; X_{N+1}Y'_N].$$

The ratchet being retrodictive means that the ratchet state X_N shields the past $Y'_{0:N}$ from the future $Y'_{N:\infty}$ and that the future shields the ratchet from the past:

$$I[Y'_{0:N}; Y'_{N:\infty}|X_N] = 0$$
$$I[Y'_{0:N}; X_N|Y'_{N:\infty}] = 0.$$

Note that generators necessarily shield past from future $I[Y'_{0:N}; Y'_{N:\infty}|X_N] = 0$, since all temporal correlations must be stored in the generator's states. Thus, for a generator, being retrodictive is equivalent to

$$I[Y'_{0:N}; X_N] = I[Y'_{0:N}; Y'_{N:\infty}].$$

This can be seen by subtracting $I[Y'_{0:N}; X_N; Y'_{N:\infty}]$ from both sides, much as was done with the extractor.

First, to show that being retrodictive implies zero minimal entropy production, it is sufficient to show that

$$I[Y'_{0:N}; X_{N+1}, Y'_N] = I[Y'_{0:N}; Y'_{N:\infty}],$$

since we know that $I[Y'_{0:N}; X_N] = I[Y'_{0:N}; Y'_{N:\infty}]$. To do this, consider the information diagram in Fig. 9. There, we see that the difference between the two mutual pieces of information of interest reduces to the difference between the two information atoms:

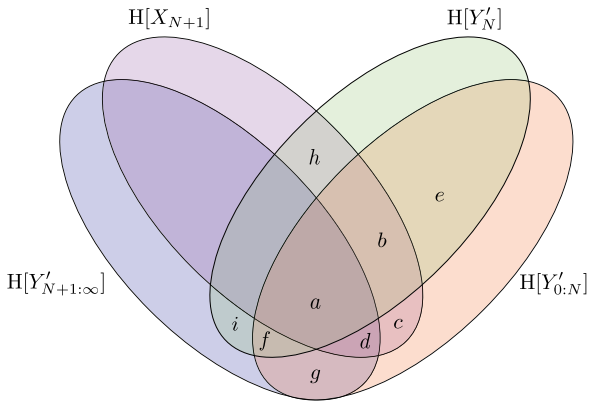

FIG. 9. Information shared between the output past $Y'_{0:N}$, next output Y'_N, next ratchet state X_{N+1}, and output future $Y'_{N+1:\infty}$, excluding the next input. Key information atoms are labeled.

$$I[Y'_{0:N}; X_{N+1}Y'_N] - I[Y'_{0:N}; Y'_{N:\infty}] = c - g.$$

The fact that the ratchet state X_{N+1} shields the past $Y'_{0:N+1}$ from the future $Y'_{N+1:\infty}$ and the future shields the ratchet from the past gives us the following four relations:

$$I[Y'_{0:N}Y'_N; Y'_{N+1:\infty}|X_{N+1}] = i + f + g = 0$$
$$I[Y'_N; Y'_{N+1:\infty}|X_{N+1}] = i + f = 0$$
$$I[Y'_{0:N}Y'_N; X_{N+1}|Y'_{N+1:\infty}] = h + b + c = 0$$
$$I[Y'_N; X_{N+1}|Y'_{N+1:\infty}] = h + b = 0.$$

These equations show that that $c = g = 0$ and, thus,

$$\frac{\langle \Sigma_N^{\text{gen}} \rangle_{\text{min}}}{k_B T \ln 2} = 0.$$

Going the other direction—zero entropy production implies retrodiction—requires that we use $I[Y'_{0:N}; X_N] = I[Y'_{0:N}; X_{N+1}, Y'_N]$ to show $I[Y'_{0:N}; X_N] = I[Y'_{0:N}; Y'_{N:\infty}]$. If X_{N+1} is retrodictive, then we can show that X_N must be as well. However, this makes the base case of the recursion difficult, since there is not yet a reason to conclude that X_∞ is retrodictive. While we conjecture the equivalence of optimally retrodictive generators and efficient generators, at this point, we can only conclusively say that retrodictive generators are also efficient.

APPENDIX C: ZERO COUPLING IN LOCAL CONTROL

The assumptions that control is limited to the interacting subsystem Z^i and that the whole system Z is an information reservoir in its default states imply that there is zero energetic coupling between the interacting subsystem Z^i and the stationary subsystem Z^s. Since information reservoir states are energetically mute, the energy of all states is

the same $E_z = E$, for all $z \in \mathcal{Z}$. We establish this using the quantum mechanical framework appearing elsewhere on thermodynamics of control restrictions [9,10]. There, a changing Hamiltonian $\mathcal{H}(t)$ is broken into the default Hamiltonian $\mathcal{H}^0$ over Z and the externally controlled Hamiltonian $H^{\text{ext}}(t)$, which only affects the local interacting subsystem:

$$\mathcal{H}(t) = \mathcal{H}^{\text{ext}}(t) + \mathcal{H}^0. \tag{C1}$$

Note that, while this operator formalism applies to quantum systems, it applies readily to classical systems [67]. In the present case, it is a particularly direct connection, as we limit our discussion to the basis of states of the information reservoir $\{|z\rangle : z \in \mathcal{Z}\}$.

Local control means that our Hamiltonian control over the joint system $\mathcal{H}^{\text{ext}}(t)$ is limited to the interacting subsystem and, thus, commutes with the stationary subsystem. However, since the joint system is an information reservoir at the beginning $(t = t_0)$ and end $(t = t_0 + \tau)$ of the computation, the result of the Hamiltonian is constant:

$$\mathcal{H}(t_0)|z\rangle = E_z|z\rangle \tag{C2}$$

$$= E|z\rangle. \tag{C3}$$

Thus, the Hamiltonian can be expressed in terms of the identity operator $\hat{I}$:

$$\mathcal{H}(t_0) = E\hat{I}. \tag{C4}$$

Also, this means that the default Hamiltonian is given by

$$\mathcal{H}^0 = E\hat{I} - \mathcal{H}^{\text{ext}}(t_0). \tag{C5}$$

Both $\hat{I}$ and $\mathcal{H}^{\text{ext}}(t)$ commute with the stationary subsystem, so the default Hamiltonian $\mathcal{H}^0$ does as well. Then, by extension, the full Hamiltonian $\mathcal{H}(t) = \mathcal{H}^{\text{ext}}(t) + \mathcal{H}^0$ commutes with the stationary subsystem. Thus, there is no energetic coupling between interacting subsystem and stationary subsystem. One concludes that the interacting system is effectively isolated from the stationary system, allowing us to consider its behavior using only its marginal distribution and local estimates of entropy production.

[1] R. Landauer, *Irreversibility, and Heat Generation in the Computing Process*, IBM J. Res. Dev. **5**, 183 (1961).

[2] S. Manipatruni, D. E. Nikonov, and I. A. Young, *Beyond CMOS Computing with Spin and Polarization*, Nat. Phys. **14**, 338 (2018).

[3] E. Fredkin and T. Toffoli, *Conservative Logic*, Int. J. Theor. Phys. **21**, 219 (1982).

[4] F. Rieke, D. Warland, R. de Ruyter van Steveninck, and W. Bialek, *Spikes: Exploring the Neural Code* (Bradford Book, New York, 1999).

[5] U. Alon, *An Introduction to Systems Biology: Design Principles of Biological Circuits* (Chapman and Hall/CRC, Boca Raton, Louisiana, 2007).

[6] K. Ulrich, *Fundamentals of Modularity*, in *Management of Design*, edited by S. Dasu and C. Eastman (Springer, Amsterdam, Netherlands, 1994), pp. 219–231.

[7] C.-C. Chen and N. Crilly, *From Modularity to Emergence: A Primer on Design and Science of Complex Systems*, Department of Engineering, University of Cambridge, Cambridge, United Kingdom, Technical Report CUED/C-EDC/TR.155, 2016.

[8] J. Maynard-Smith and E. Szathmary, *The Major Transitions in Evolution* (Oxford University Press, Oxford, 1998), reprint edition.

[9] J. Lekscha, H. Wilming, J. Eisert, and R. Gallego, *Quantum Thermodynamics with Local Control*, Phys. Rev. E **97**, 022142 (2018).

[10] H. Wilming, R. Gallego, and J. Eisert, *Second Law of Thermodynamics under Control Restrictions*, Phys. Rev. E **93**, 042126 (2016).

[11] S. Lloyd, *Use of Mutual Information to Decrease Entropy: Implications for the Second Law of Thermodynamics*, Phys. Rev. A **39**, 5378 (1989).

[12] T. Sagawa and M. Ueda, *Fluctuation Theorem with Information Exchange: Role of Correlations in Stochastic Thermodynamics*, Phys. Rev. Lett. **109**, 180602 (2012).

[13] H. Touchette and S. Lloyd, *Information-Theoretic Limits of Control*, Phys. Rev. Lett. **84**, 1156 (2000).

[14] H. Touchette and S. Lloyd, *Information-Theoretic Approach to the Study of Control Systems*, Physica (Amsterdam) **331A**, 140 (2004).

[15] J. M. R. Parrondo, J. M. Horowitz, and T. Sagawa, *Thermodynamics of Information*, Nat. Phys. **11**, 131 (2015).

[16] R. Storn and K. Price, *Differential Evolution: A Simple and Efficient Heuristic for Global Optimization over Continuous Space*, J. Global Optim. **11**, 341 (1997).

[17] J. D. Lohn, G. S. Hornby, and D. S. Linden, *An Evolved Antenna for Deployment on NASA's Space Technology 5 Mission*, in *Genetic Programming Theory and Practice II*, edited by U.-M. O'Reilly, T. Yu, R. Riolo, and B. Worzel (Springer US, New York, 2005), pp. 301–315.

[18] J. Koza, *Genetic Programming: On the Programming of Computers by Means of Natural Selection* (Bradford Book, New York, 1992).

[19] S. Deffner and C. Jarzynski, *Information Processing and the Second Law of Thermodynamics: An Inclusive, Hamiltonian Approach*, Phys. Rev. X **3**, 041003 (2013).

[20] D. Mandal and C. Jarzynski, *Work and Information Processing in a Solvable Model of Maxwell's Demon*, Proc. Natl. Acad. Sci. U.S.A. **109**, 11641 (2012).

[21] A. B. Boyd, D. Mandal, and J. P. Crutchfield, *Identifying Functional Thermodynamics in Autonomous Maxwellian Ratchets*, New J. Phys. **18**, 023049 (2016).

[22] N. Merhav, *Sequence Complexity and Work Extraction*, J. Stat. Mech. (2015) P06037.

[23] A. B. Boyd, D. Mandal, and J. P. Crutchfield, *Leveraging Environmental Correlations: The Thermodynamics of Requisite Variety*, J. Stat. Phys. **167**, 1555 (2017).

[24] A. B. Boyd, D. Mandal, and J. P. Crutchfield, *Correlation-Powered Information Engines and the Thermodynamics of Self-Correction*, Phys. Rev. E **95**, 012152 (2017).

[25] N. Barnett and J. P. Crutchfield, *Computational Mechanics of Input-Output Processes: Structured Transformations and the ϵ-Transducer*, J. Stat. Phys. **161**, 404 (2015).

[26] J. G. Brookshear, *Theory of Computation: Formal Languages, Automata, and Complexity* (Benjamin/Cummings, Redwood City, California, 1989).

[27] A. J. P. Garner, J. Thompson, V. Vedral, and M. Gu, *Thermodynamics of Complexity and Pattern Manipulation*, Phys. Rev. E **95**, 042140 (2017).

[28] M. Esposito and C. van den Broeck, *Second Law and Landauer Principle Far from Equilibrium*, Europhys. Lett. **95**, 40004 (2011).

[29] R. Motwani and P. Raghavan, *Randomized Algorithms* (Cambridge University Press, New York, 1995).

[30] M. Mitzenmacher and E. Upfal, *Probability and Computing: Randomized Algorithms and Probabilistic Analysis* (Cambridge University Press, New York, 2005).

[31] C. Aghamohammadi and J. P. Crutchfield, *Thermodynamics of Random Number Generation*, Phys. Rev. E **95**, 062139 (2017).

[32] J. P. Crutchfield, *The Calculi of Emergence: Computation, Dynamics, and Induction*, Physica (Amsterdam) **75D**, 11 (1994).

[33] P. R. Zulkowski and M. R. DeWeese, *Optimal Finite-Time Erasure of a Classical Bit*, Phys. Rev. E **89**, 052140 (2014).

[34] M. Esposito, *Stochastic Thermodynamics under Coarse Graining*, Phys. Rev. E **85**, 041125 (2012).

[35] F. M. Reza, *An Introduction to Information Theory* (Courier Corporation, North Chelmsford, MA, USA, 1961).

[36] C. J. Ellison, J. R. Mahoney, R. G. James, J. P. Crutchfield, and J. Reichardt, *Information Symmetries in Irreversible Processes*, Chaos **21**, 037107 (2011).

[37] J. P. Crutchfield, C. J. Ellison, and J. R. Mahoney, *Time's Barbed Arrow: Irreversibility, Crypticity, and Stored Information*, Phys. Rev. Lett. **103**, 094101 (2009).

[38] F. N. Fahn, *Maxwell's Demon and the Entropy Cost of Information*, Found. Phys. **26**, 71 (1996).

[39] T. E. Ouldridge, C. C. Govern, and P. Rein ten Wolde, *Thermodynamics of Computational Copying in Biochemical Systems*, Phys. Rev. X **7**, 021004 (2017).

[40] A. B. Boyd and J. P. Crutchfield, *Maxwell Demon Dynamics: Deterministic Chaos, the Szilard Map, and the Intelligence of Thermodynamic Systems*, Phys. Rev. Lett. **116**, 190601 (2016).

[41] T. Simula, M. J. Davis, and K. Helmerson, *Emergence of Order from Turbulence in an Isolated Planar Superfluid*, Phys. Rev. Lett. **113**, 165302 (2014).

[42] I. Kozinsky, H. W. Ch. Postma, O. Kogan, A. Husain, and M. L. Roukes, *Basins of Attraction of a Nonlinear Nanomechanical Resonator*, Phys. Rev. Lett. **99**, 207201 (2007).

[43] C. M. Quintana *et al.*, *Observation of Classical Quantum Crossover of 1/f Flux Noise and Its Paramagnetic Temperature Dependence*, Phys. Rev. Lett. **118**, 057702 (2017).

[44] F. Yan, S. Gustavsson, A. Kamal, J. Birenbaum, A. P. Sears, D. Hover, D. Rosenberg, G. Samach, T. J. Gudmundsen, J. L. Yoder, T. P. Orlando, J. Clarke, A. J. Kerman, and W. D. Oliver, *The Flux Qubit Revisited to Enhance Coherence and Reproducibility*, Nat. Commun. **7**, 12964 (2016).

[45] Y. Jun, M. Gavrilov, and J. Bechhoefer, *High-Precision Test of Landauer's Principle in a Feedback Trap*, Phys. Rev. Lett. **113**, 190601 (2014).

[46] S. Still, D. A. Sivak, A. J. Bell, and G. E. Crooks, *Thermodynamics of Prediction*, Phys. Rev. Lett. **109**, 120604 (2012).

[47] S. Still, *Thermodynamic Cost and Benefit of Data Representations*, arXiv:1705.00612.

[48] S. Toyabe, T. Sagawa, M. Ueda, E. Muneyuki, and M. Sano, *Experimental Demonstration of Information-to-Energy Conversion and Validation of the Generalized Jarzynski Equality*, Nat. Phys. **6**, 988 (2010).

[49] T. Sagawa and M. Ueda, *Generalized Jarzynski Equality under Nonequilibrium Feedback Control*, Phys. Rev. Lett. **104**, 090602 (2010).

[50] J. M. Horowitz, T. Sagawa, and J. M. R. Parrondo, *Imitating Chemical Motors with Optimal Information Motors*, Phys. Rev. Lett. **111**, 010602 (2013).

[51] T. McGrath, N. S. Jones, P. Rein ten Wolde, and T. E. Ouldridge, *Biochemical Machines for the Interconversion of Mutual Information and Work*, Phys. Rev. Lett. **118**, 028101 (2017).

[52] L. H. Hartwell, J. J. Hopfield, S. Leibler, and A. W. Murray, *From Molecular to Modular Cell Biology*, Nature (London) **402**, C47 (1999).

[53] A. B. Boyd, D. Mandal, P. M. Riechers, and J. P. Crutchfield, *Transient Dissipation and Structural Costs of Physical Information Transduction*, Phys. Rev. Lett. **118**, 220602 (2017).

[54] Z. Lu, D. Mandal, and C. Jarzynski, *Engineering Maxwell's Demon*, Phys. Today **67**, No. 8, 60 (2014).

[55] D. Mandal, H. T. Quan, and C. Jarzynski, *Maxwell's Refrigerator: An Exactly Solvable Model*, Phys. Rev. Lett. **111**, 030602 (2013).

[56] A. Turing, *On Computable Numbers, with an Application to the Entschiedungsproblem*, Proc. London Math. Soc. **s2-42**, 230 (1937).

[57] Space limitations here do not allow a full digression on possible implementations. Suffice it to say that for unidirectional tape reading, the ratchet state requires a storage register or an auxiliary internal working tape as portrayed in Fig. 3 of Ref. [32].

[58] C. E. Shannon, *A Universal Turing Machine with Two Internal States*, in *Automata Studies*, Annals of Mathematical Studies Vol. 34, edited by C. E. Shannon and J. McCarthy (Princeton University Press, Princeton, New Jersey, 1956), pp. 157–165.

[59] M. Minsky, *Computation: Finite and Infinite Machines* (Prentice-Hall, Englewood Cliffs, New Jersey, 1967).

[60] P. Strasberg, J. Cerrillo, G. Schaller, and T. Brandes, *Thermodynamics of Stochastic Turing Machines*, Phys. Rev. E **92**, 042104 (2015).

[61] N. Merhav, *Relations between Work and Entropy Production for General Information-Driven, Finite-State Engines*, J. Stat. Mech. (2017) P023207.

[62] J. P. Crutchfield and D. P. Feldman, *Regularities Unseen, Randomness Observed: Levels of Entropy Convergence*, Chaos **13**, 25 (2003).

[63] P. Strasberg, *Thermodynamics and Information Processing at the Nanoscale*, Ph.D. thesis, Technische Universität Berlin, 2016.

[64] A. Chapman and A. Miyake, *How Can an Autonomous Quantum Maxwell Demon Harness Correlated Information?*, Phys. Rev. E **92**, 062125 (2015).

[65] J. P. Crutchfield and K. Young, *Inferring Statistical Complexity*, Phys. Rev. Lett. **63**, 105 (1989).

[66] D. H. Wolpert, A. Kolchinsky, and J. A. Owen, *The Minimal Hidden Computer Needed to Implement a Visible Computation*, arXiv:1708.08494.

[67] P. M. Riechers and J. P. Crutchfield, *Fluctuations When Driving between Nonequilibrium Steady States*, J. Stat. Phys. **168**, 873 (2017).