Harnessing fluctuations in thermodynamic computing via time-reversal symmetries

Gregory Wimsatt,[1,*] Olli-Pentti Saira,[2,†] Alexander B. Boyd,[1,‡] Matthew H. Matheny,[2,§] Siyuan Han,[3,∥]
Michael L. Roukes,[2,¶] and James P. Crutchfield[1,2,#]

[1]*Complexity Sciences Center and Physics Department, University of California at Davis, One Shields Avenue, Davis, California 95616, USA*
[2]*Condensed Matter Physics and Kavli Nanoscience Institute, California Institute of Technology, Pasadena, California 91125, USA*
[3]*Department of Physics and Astronomy, University of Kansas, Lawrence, Kansas 66045, USA*

We experimentally demonstrate that highly structured distributions of work emerge during even the simple task of erasing a single bit. These are signatures of a refined suite of time-reversal symmetries in distinct functional classes of microscopic trajectories. As a consequence, we introduce a broad family of conditional fluctuation theorems that the component work distributions must satisfy. Since they identify entropy production, the component work distributions encode the frequency of various mechanisms of both success and failure during computing, as well giving improved estimates of the total irreversibly dissipated heat. This new diagnostic tool provides strong evidence that thermodynamic computing at the nanoscale can be constructively harnessed. We experimentally verify this functional decomposition and the new class of fluctuation theorems by measuring transitions between flux states in a superconducting circuit.

I. INTRODUCTION

Physics dictates that all computing is subject to spontaneous error. These days, this truism repeatedly reveals itself: despite the once-predictable miniaturization of nanoscale electronics, computing performance increases have dramatically slowed in the last decade or so. In large measure, this is due to the concomitant rapid decrease in the number of information-bearing physical degrees of freedom, rendering information storage and processing increasingly susceptible to corruption by thermal fluctuations. Said simply, all physical computing is thermodynamic.

Controlling the production of fluctuations and removing heat pose key technological challenges to further progress. Practically, the challenge remains of how to probe and diagnose information processing in overtly noisy systems. The following introduces trajectory class fluctuation theorems to do this by identifying the thermodynamic signature of successful and failed information processing. It then experimentally demonstrates how this is practically implemented in a new microscale platform for thermodynamic computing.

———————

*gwwimsatt@ucdavis.edu
†osaira@caltech.edu
‡abboyd@ucdavis.edu
§matheny@caltech.edu
∥han@ku.edu
¶roukes@caltech.edu
#Corresponding author: chaos@ucdavis.edu

Only recently have tools appeared that precisely describe what tradeoffs exist between thermodynamic resources and useful information processing—these are highly reminiscent of the centuries-old puzzle of how Maxwell's "very observant and neat-fingered" demon uses its "intelligence" to convert disorganized heat energy to useful work [1]. In our modern era, his demon has led to the realization that information itself is physical [2–4]—or, most constructively, that information is a thermodynamic resource (see [5] and references therein). This opened up the new paradigm of *thermodynamic computing* [6] in which fluctuations play a positive role in efficient information processing on the nanoscale. We now conceptualize this via *information engines*: physical systems that are driven by, manipulate, store, and dissipate energy, but simultaneously generate, store, lose, communicate, and transform information. In short, information engines combine traditional engines comprised of heat, work, and other familiar reservoirs with what we now call *information reservoirs* [7,8].

Reliable thermodynamic computing requires detecting and controlling fluctuations in informational and energetic resources and in engine functioning. To do so requires a new generation of diagnostic tools. For these, we appeal to fluctuation theorems that capture exact time-reversal symmetries and predict entropy production leading to irreversible dissipation [9–15]. As the following demonstrates, these place us on the doorstep of the very far-from-equilibrium thermodynamics needed to understand the physics of computing. And, in turn, the physical principles of how nature processes information in the service of biological functioning and survival have begun to emerge.

Proof-of-concept experimental tests have been carried out in several substrates: probing biomolecule free energies [16–18], work expended during elementary computing (bit erasure) [19–24], and Maxwellian demons [25]. That said, the
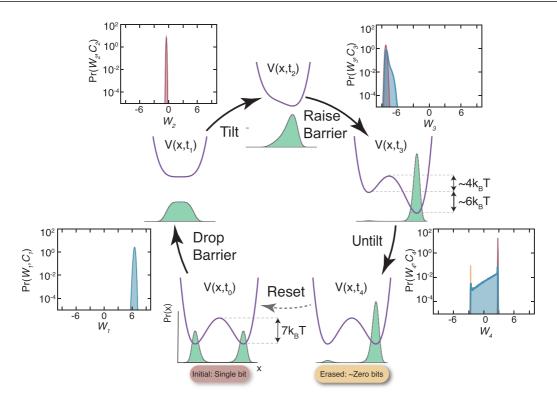
FIG. 1. Inner plot sequence: Erasure protocol (Table I) evolution of position distribution $\Pr(x)$ from simulation. Potential $V(x, t_s)$ at substage boundary times t_s, $s = 0, 1, 2, 3, 4$. Starting at $t = t_0$, the potential evolves clockwise, ending at $t = t_4$ in the same configuration as it starts: $V(x, t_0) = V(x, t_4)$. However, the final position distribution $\Pr(x)$ predominantly indicates the R state. The original one bit of information in the distribution at time $t = t_0$ has been erased. Outer plot sequence: Substage work distributions from simulation $\Pr(W_s, C_s)$ during substages s: (1) Barrier Drop, (2) Tilt, (3) Barrier Raise, (4) Untilt. During each substage s, distributions are given for up to three substage trajectory classes C_s: red consists of trajectories always in the R state, orange trajectories always in the L state, and blue the rest, spending some time in each state.

suite of theoretical predictions and contemporary principles (Appendix A) far outstrips experimental validation to date.

To close the gap, we show how to diagnose thermodynamic computing on the nanoscale by explaining the signature structures in work distributions generated during information processing. Previous efforts explored features in work and heat distributions that track the mesoscale evolution of a system's *informational states*; see Refs. [26,27] and Appendix A. Here, we show that functional and nonfunctional informational-state evolutions can be identified by appropriate conditioning, and that their thermodynamics obey a suite of trajectory-class fluctuation theorems. As such, the latter give accurate bounds on work, entropy production, and dissipation for computing subprocesses. The result is a practical tool that employs mesoscopic (work) measurements to diagnose microscopic thermodynamic computing. For simplicity, and to make direct contact with previous efforts, we demonstrate the tools on Landauer erasure [2] of a bit of information in a superconducting flux qubit.

II. MODEL SYSTEM

As a reference, we first explore the thermodynamics of bit erasure in a simple model: a particle with position and momentum in a double-well potential $V(x, t)$ and in contact with a heat reservoir at temperature T. (Refer to Fig. 1.) An external controller adds or removes energy from a work

reservoir to change the form of the potential $V(\cdot, t)$ via a predetermined *erasure protocol* $\{(\beta(t), \delta(t)) : 0 \leqslant t \leqslant \tau\}$. $\beta(t)$ and $\delta(t)$ change one at a time piecewise linearly through four protocol substages: (i) *drop barrier*, (ii) *tilt*, (iii) *raise barrier*, and (iv) *untilt*. (See Appendix B.) The system starts at time $t = 0$ in the equilibrium distribution for a double-well $V(x, 0)$ at temperature T.

We use *underdamped* Langevin dynamics to simulate this model:

$$dx = v\,dt,$$

$$m\,dv = \sqrt{2k_{\mathrm{B}}T\gamma}\,r(t)\sqrt{dt} - \left(\frac{\partial}{\partial x}V(x, t) + \gamma v\right)dt, \quad (1)$$

where k_{B} is Boltzmann's constant, γ is the coupling between the heat reservoir and the system, m is the particle's mass, and

TABLE I. Erasure protocol.

Stage		Drop Barrier		Tilt		Raise Barrier		Untilt	
t_s	t_0			t_1		t_2		t_3	t_4
$\beta(t)$	|	$\frac{t_1-t}{t_1-t_0}$	|	0	|	$\frac{t-t_2}{t_3-t_2}$	|	1	|
$\delta(t)$	|	0	|	$\frac{t-t_1}{t_2-t_1}$	|	1	|	$\frac{t_4-t}{t_4-t_3}$	|

$r(t)$ is a memoryless Gaussian random variable with $\langle r(t) \rangle = 0$ and $\langle r(t)r(t') \rangle = \delta(t - t')$.

The default potential, $V(\cdot, 0) = V(\cdot, \tau)$, has two symmetric wells separated by a barrier. Following common practice we call the two wells, from negative to positive position, the *left* (*L*) and *right* (*R*) *informational states*, respectively. Initially being equiprobable, the informational states associated with each of the two wells thus contain 1 bit of information [28].

The erasure protocol is designed so that the particle ends in the R state with high probability, regardless of its initial state. Simulating the protocol 3.5×10^6 times, 96.2% of the particles were successfully erased into the R state. Thus, as measured by the Shannon entropy, the initial 1 bit of information was reduced to 0.231 bits. We intentionally designed the protocol to fail frequently at erasure to better illustrate our main results on diagnosing success *and* failure. But, crucially, the results we present hold for arbitrarily successful erasure protocols.

At all other times t, $V(\cdot, t)$ has either one or two local minima, naturally defining metastable regions for a particle to be constrained and gradually evolve toward local equilibrium. We therefore define the informational states at time $0 \leqslant t \leqslant \tau$ to be the metastable regions, labeling them R and, if two exist, L—from most positive to negative in position.

Since the protocol is composed of four simple substages, we coarse-grain the system's response by its activity during each substage at the level of its informational state. Specifically, for each substage, we assign one of three *substage trajectory classes*: the system (i) was always in the R state, (ii) was always in the L state, or (iii) spent time in each. Sometimes there is only one informational state, and so the latter two classes are not achievable for all substages.

III. WORK CHARACTERIZATION

We then focus on a single mesoscopic observable—the thermodynamic work expended during erasure. An individual realization generates a trajectory of system microstates, with $W(t, t')$ being the work done on the system between times $0 \leqslant t < t' \leqslant \tau$; see Appendix H. Let $W_s = W(t_{s-1}, t_s)$ denote the work generated during substage s, and C_s denotes the substage trajectory class. Figure 1 (outer plot sequence) shows the corresponding substage work distributions $\Pr(W_s, C_s)$ obtained from our simulations. (See Appendix I.)

The drop-barrier and tilt substage work distributions are rather simple, being narrow and unimodal. The raise-barrier distributions have some asymmetry, but they are also similarly simple. However, the untilt work distributions (farthest right in Fig. 1) exhibit unusual features that are significant for understanding the intricacies of erasure. Trajectories that spend all of the untilt substage in either the R state or the L state form peaks at the most positive (red) and negative (orange) work values, respectively. This is because the R-state well is always increasing in potential energy while the L-state well is always decreasing during untilt. In contrast, the other trajectories contribute a log-linear ramp of work values (blue) dependent on the time spent in each. The ramp's positive slope signifies that more time is typically spent in the R state in this last set of trajectories.
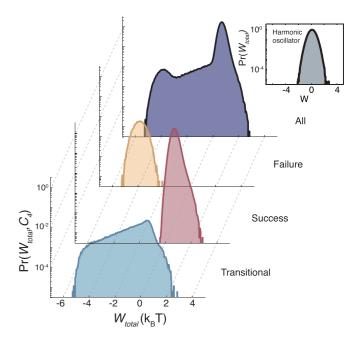

FIG. 2. Rear (purple): total work distribution of all trajectories $\Pr(W_{\text{total}})$ during erasure simulation: a histogram generated from 3.5×10^6 trials for $W_{\text{total}} \in [-6, 4]$ over 201 bins. Inset (gray): typical unimodal work distribution illustrated for a spatially translated thermally driven simple harmonic oscillator. Three front plots: work distributions $\Pr(W_{\text{total}}, C_4)$ for the trajectory classes C_4 determined by the untilt trajectory partition in simulation: The red work distribution (middle) is that of Success trajectories, the orange (rear) is that of Fail trajectories, and the blue (front) is that of the remaining, Transitional trajectories.

Looking at the total work $W_{\text{total}} = W(0, \tau)$ generated for each trajectory over the course of the entire erasure protocol, we observe the strikingly complex and structured distribution $\Pr(W_{\text{total}})$ shown in Fig. 2 (rear). There are two clear peaks at the most positive and negative work values separated by a ramp. This highly structured work distribution, generated by bit erasure, contrasts sharply with the unimodal work distributions common in previous studies of fluctuation theorems; see, for example, Fig. 2 (inset) for the work distribution generated by a thermodynamically driven simple harmonic oscillator translated in space or Fig. 2 in Ref. [13]. The total average work was $0.634k_BT$, satisfying Landauer's bound by being greater than the informational-state Shannon entropy decrease of $0.769 \times \ln 2 k_BT = 0.533k_BT$.

We can understand the mechanisms behind this structure when decomposing Fig. 2 (rear)'s total work distribution under the untilt substage trajectory classes C_4. We label trajectories that spend all of the untilting substage in the R state as *Success* since, via the previous substages, they reach the intended R state by the untilting substage and remain there until the protocol's end. Similarly, trajectories that spend all of the untilt substage in the L state are labeled *Fail*. The remaining trajectories are labeled *Transitional*, since they transition between the two informational states during untilt, potentially succeeding or failing to end in the R state.

Figure 2's three front plots show the work distribution for each of these three trajectory classes. Together they recover

the total work distribution over all trajectories shown in Fig. 2 (rear). However, now the thermodynamic contributions to the total from the functionally distinct component trajectories are made apparent.

IV. TRAJECTORY-CLASS FLUCTUATION THEOREM

Exploring the mesoscale dynamics of erasure revealed signatures of a "thermodynamics" for each trajectory that is uniquely associated with successful or failed information processing. We now introduce the underlying fluctuation theory from which the trajectory thermodynamics follow. Key to this is comparing system behaviors in both forward and reverse time [9–15]. (See Appendixes C and F.)

This suite of trajectory-class fluctuation theorems (TCFTs) applies to arbitrary classes of system microstate trajectories obtainable during a thermodynamic transformation. Importantly, they interpolate between Jarzynki's equality [11] and Crooks' detailed fluctuation theorem [13] as the trajectory class varies from the entire ensemble of trajectories to a single particular trajectory, respectively. Accordingly, they unify a wide range of other previously established fluctuation theorems. (See Appendix C.)

One TCFT presents a lower bound on the average work $\langle W \rangle_C$ over any measurable subset C of the ensemble of system microstate trajectories $\vec{\mathcal{Z}}$, where W is the total work for a trajectory:

$$\langle W \rangle_C \geqslant \Delta F + k_B T \ln \frac{\mathcal{P}(C)}{\mathcal{R}(C^R)}$$
$$= \langle W \rangle_C^{\min}, \tag{2}$$

with ΔF the change in equilibrium free energy over the protocol, $\mathcal{P}(C)$ the probability of realizing the class C during the protocol, and $\mathcal{R}(C^R)$ the probability of obtaining the time reverse of class C under the time-reverse protocol. As detailed in Appendix G, this allows accurate estimation of the work generated for trajectory classes with narrow work distributions, such as the Success and Fail classes of erasure, even with limited knowledge (low sampling) of system response under the protocol and its time reverse.

The TCFTs lead to additional consequences. First, they more strongly bound the average work over all trajectories compared to the equilibrium free-energy change ΔF. Second, they provide a new expression for obtaining equilibrium free-energy changes:

$$\Delta F = -k_B T \ln \left(\frac{\mathcal{P}(C)}{\mathcal{R}(C^R)} \langle e^{-W/k_B T} \rangle_C \right). \tag{3}$$

Remarkably, this only requires statistics for any particular class C and its reverse C^R to produce the system's free-energy change. Since rare microstate trajectories may generate sufficiently negative works that dominate the average exponential work, this leads to a substantial statistical advantage over direct use of Jaryznski's equality $\Delta F = -k_B T \ln \langle e^{-W/k_B T} \rangle_{\vec{\mathcal{Z}}}$ for estimating free energies [29]. (See Appendix C.)

The erasure protocol's 96.2% success rate is reflected in the Success class's dominance in the work distributions of Fig. 2. We can adjust the protocol to exhibit an arbitrarily higher success rate while still maintaining high efficiency, i.e.,

keeping the total average work close to Landauer's bound. When done, we still observe the same qualitative features of the work distribution: two peaks separated by a log-linear ramp, each associated with the Success, Fail, and Transitional classes. Though, of course, the probabilities of the Fail and Transitional classes become arbitrarily small.

That said, the contributions of the Fail and Transitional classes to various fluctuation theorems—such as, Jarzynski's equality to mention one—remain significant since the work generated by those classes compensates by becoming increasingly negative. In fact, the contribution to the exponential average work of the Success class only approaches the value 1/2 out of the required value of 1 when averaging over all trajectories. Thus, while the probabilities of the Fail and Transitional classes can become arbitrarily small by considering Erasure protocols with higher success rates, we cannot ignore the existence of the rare events due to Transitional and Fail trajectories unless we employ particular fluctuation theorems, in particular, a TCFT. (Again, see Appendix C.) In this way, one sees that the TCFT provides a detailed diagnosis of successful and failed information processing and of the associated energetics.

V. REALIZING THERMODYNAMIC COMPUTING

To explore these predictions, we selected a superconducting flux qubit composed of paired Josephson junctions [Fig. 3(a)], resulting in a double-well nonlinear potential that supports information storage and processing [Fig. 3(b)]. Appendix J 1 explains the physics underlying their nonlinear equations of motion, comparing the similarities and differences with our model's idealized Langevin dynamics.

Despite control protocols for double-well potentials that perform accurate and efficient bit erasure [30], we run the flux qubit in a mode that yields imperfect erasure [Fig. 3(c)]. As with the simulations, our intention is to illustrate how trajectory classes and the TCFT can be used to diagnose and interpret success and failure in microscopic information processing using only mesoscopic measurements of work, which is done more clearly by increasing the probabilities of rare events.

Interplay between the linear geometric magnetic and nonlinear Josephson inductances gives rise to a potential landscape that can be controlled with external bias fluxes. It is natural to call the ϕ_x and ϕ_{xdc} fluxes, threading the differential mode and the small SQUID loop, respectively, the *tilt* and *barrier controls*. [See the caption to Fig. 3(a).] Appendix J presents a derivation of the flux qubit potential and details its calibration. All experiments presented here were carried out at a temperature of 500 mK so that the effect of quantum fluctuations is negligible.

To execute an erasure protocol, we first choose an information-storage state with a tall barrier and two equal-depth wells. The two-dimensional potential for this at the calibrated device parameters is depicted in Fig. 3(b). We implement the bit erasure protocol as a time-domain deformation imposed by the two control fluxes that starts and ends at the storage configuration. In contrast to the simulation, the flux qubit maintains two metastable regions and, hence, two informational states L and R at all times, though they are
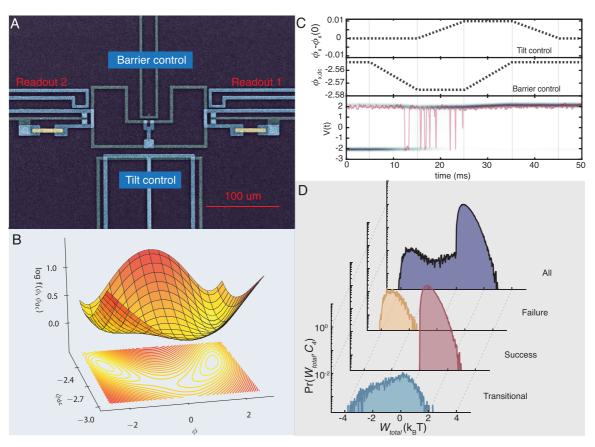
FIG. 3. Superconducting implementation of metastable memory and bit erasure driven by thermal fluctuations: (a) Optical micrograph of a gradiometric flux qubit with control lines and local magnetometers for state readout. The flux ϕ_x, threading the large U-shaped differential-mode loop, controls the potential's tilt, and the flux ϕ_{xdc}, threading the small SQUID loop, controls the potential barrier height. Currents in the *barrier control* and *tilt control* lines modulate those fluxes. (b) Calculated potential energy landscape at the beginning of the erasure protocol; see Eqs. (J1) and (J2). (c, top) Sequence of tilt and barrier control waveforms implementing bit erasure, and (c, bottom) sample of resulting magnetometer traces tracking the system's internal state. Note that for this experiment, in contrast to the simulation, two possible informational states were present at all times during the protocol. So, though the barrier was reduced sufficiently to allow transitions to the target R state, the trace is attracted to either a positive or negative value at all times. (d) Work distributions $\Pr(W_{\text{total}}|C_4)$ over trajectories conditioning on the Success, Fail, and Transitional classes. Experimental distributions obtained from 10^5 protocol repetitions.

shallow enough to allow transitions as the barrier drops. The amplitudes of the control waveforms in reduced units are small; see Fig. 3(c). Due to this, the microscopic energetics change linearly as a function of the control fluxes.

We use a local dc-SQUID magnetometer to continuously monitor the trapped flux state in the device—*Readout* 1 in Fig. 3(a). The digitized signal has a rise time of 100 μs, after which the two logical states are discriminated virtually without error. A typical magnetometer trace $V(t)$ acquired during the execution of the erasure protocol is shown in Fig. 3(c). We operate the magnetometer with a low-amplitude AC current bias at 10 MHz to avoid an increase in the effective temperature during continuous readout of the flux state due to wide-band electromagnetic interference.

To collect work statistics, we repeat the erasure protocol 10^5 times. We identify the logical-state transitions from the magnetometer traces as zero-crossings, recording the direction δ_i—sign convention: $+1$ (-1) for an L-to-R (R-to-L) transition—and the time t_i relative to the start of the protocol. We evaluate a single-shot work estimate $W = \sum_i \delta_i U_{LR}(t_i)$, where $U_{LR}(t) = U_R(t) - U_L(t)$ is the biasing of the potential minima at time t_i. Making use of the linearity of the

system energetics and the choice of offsets and compensation coefficients, we find $U_{LR}(t) = A[\phi_x(t) - \phi_x(0)]$, with the coefficient $A = 210\,\text{K} \times k_B$ evaluated from the calibrated potential. The above work estimate based on the logical-state transitions is an accurate estimate of the true microscopic work assuming that the timescales for the state transitions and for changes in the control parameters are much slower than the intrawell equilibration. (See Appendix H.)

The total work distribution estimated from the flux qubit experiments is shown as the rearmost distribution in Fig. 3(d). Using the previous microstate trajectory partitioning into the *Success, Fail,* and *Transitional* trajectory classes reveals a decomposition of the total work distribution given by Fig. 3(d) (three front panels). The close similarity with our simulations (Fig. 2) is notable, especially given the rather substantial differences between the simulated system (idealized double-well potential, thermal noise, exactly one-dimensional system, etc.) and the experimental system (complex potential in two dimensions, nonideal fluctuations, etc.). *A priori* it is not clear that the TCFT predictions should apply so directly and immediately to the real-world qubit, that is until one recalls that trajectory-class membership is a topological

property and that trajectories carry their probabilities and so their thermodynamics.

In point of fact, these differences serve to emphasize the descriptive power and robustness of the mesoscopic-work TCFT: Despite substantial differences in system detail, they successfully diagnose the information-processing classes of microscopic trajectories.

Indeed, looking to thermodynamic transformations beyond bit erasure, the essential requirement of our analysis is for the protocol to be slow enough compared to the timescales of oscillations due to the potential and of the thermal fluctuations. With this, the system is always near metastable equilibrium—what we call *metastable quasistationarity*. This ensures that in an amount of time, small on the timescale of the protocol, the system visits every point in the potential in proportion to the metastable distribution for the metastable region it occupies. Since the work rate for a particle is determined by the rate of change in potential at its location, the work rate must be that of the metastable distribution's. We then need only describe which metastable region a particle is in as a function of time to characterize its total work. The specific shape and dimensionality of these metastable regions are then insignificant for determining the shape and qualitative features of the total work distribution. Under these conditions, there is substantial robustness. This will be especially helpful when using the TCFT to monitor thermodynamic computing in biological systems where, in many cases, information-bearing degrees-of-freedom cannot be precisely modeled.

VI. CONCLUSION

We experimentally demonstrated that work fluctuations generated by information engines are highly structured. Nonetheless, they strictly obeyed a suite of time-reversal symmetries—the trajectory-class fluctuation theorems introduced here. The latter are direct signatures of how a system's informational states evolve, and they identify functional and nonfunctional microscopic trajectory bundles. We showed that the trajectory-class fluctuation theorems naturally interpolate between Jarzynski's integral and Crooks' detailed fluctuation theorems, providing a unified diagnostic probe of nonequilibrium thermodynamic transformations that support information processing.

The trajectory-class fluctuation theorems gave a detailed thermodynamic analysis of the now-common example of erasing a bit of information as an external protocol manipulated a stochastic particle in a double-well potential (simulation) and the stochastic state of a flux qubit (experiment). To give insight into the new level of mechanistic analysis possible, we briefly discussed the untilt trajectory-class partitioning. Though ignoring other protocol stages, this was sufficient to capture the basic trajectory classes that generate the overall work distribution's features. Partitioning on informational-state occupation times during barrier raising and untilting—an alternative used in follow-on studies—yields an even more incisive decomposition of the work distributions and diagnosis of informational functioning. Practically, the corresponding bounds on thermodynamic resources obtained via the TCFT also improve on current estimation methods. The net result is that trajectory-class fluctuation analysis can be readily applied

to debug thermodynamic computing by engineered or biological systems.

The data that support the findings of this study are available from the corresponding author upon reasonable request.

APPENDIX A: PRINCIPLES OF THERMODYNAMIC COMPUTING: A RECENT SYNOPSIS

A number of closely related thermodynamic costs of computing have been identified, above and beyond the *housekeeping heat* that maintains a system's overall nonequilibrium dynamical state. First, there is the *information-processing Second Law* [31] that extends Landauer's original bound on erasure [2] to dissipation in general computing and properly highlights the central role of information generation measured via the physical substrate's dynamical Kolmogorov-Sinai entropy. It specifies the minimum amount of energy that must be supplied to drive a given amount of computation forward. Second, when coupling thermodynamic systems together, even a single system and a complex environment, there are transient costs as the system synchronizes to, predicts, and then adapts to errors in its environment [32–34]. Third, the very modularity of a system's organization imposes thermodynamic costs [35]. Fourth, since computing is necessarily far out of equilibrium and nonsteady state, there are costs due to driving transitions between information-storage states [36]. Fifth, there are costs to generating randomness [37], which is itself a widely useful resource. Finally, by way of harnessing these principles, new strategies for optimally controlling nonequilibrium transformations have been introduced [30,38–40].

APPENDIX B: MICROSCOPIC STOCHASTIC THERMODYNAMICAL SYSTEM

For concreteness, we concentrate on a one-dimensional system: a particle with position and momentum in an external potential $V(x, t)$ and in contact with a heat reservoir at temperature T. An external controller adds or removes energy from a work reservoir to change the form of the potential $V(\cdot, t)$ via

a predetermined *erasure protocol* $\{(\beta(t), \delta(t)) : 0 \leqslant t \leqslant \tau\}$. (See Appendix H for details on the alternative definitions of work.) The potential takes the form

$$V(x, t) = ax^4 - b_0\beta(t)x^2 - c_0\delta(t)x,$$

with constants $a, b_0, c_0 > 0$. During the erasure protocol, $\beta(t)$ and $\delta(t)$ change one at a time piecewise-linearly through four protocol substages: (i) *drop barrier*, (ii) *tilt*, (iii) *raise barrier*, and (iv) *untilt*, as shown in Table I. The system starts at time $t = 0$ in the equilibrium distribution for a double-well $V(x, 0)$ at temperature T. Being equiprobable, the informational states associated with each of the two wells thus contain 1 bit of information [28]. The effect of the control protocol on the system potential and system response is graphically displayed in Fig. 1.

We model the erasure physical information processing with *underdamped* Langevin dynamics:

$$dx = v\, dt,$$

$$m\, dv = \sqrt{2k_B T\gamma}\, r(t)\sqrt{dt} - \left(\frac{\partial}{\partial x}V(x, t) + \gamma v\right)dt,$$

where k_B is Boltzmann's constant, γ is the coupling between the heat reservoir and the system, m is the particle's mass, and $r(t)$ is a memoryless Gaussian random variable with $\langle r(t)\rangle = 0$ and $\langle r(t)r(t')\rangle = \delta(t - t')$.

For comparison to experiment, we simulated erasure with the following parameters, sufficient to fully specify the dynamics: $\gamma\tau/m = 500$, $2k_B T\tau^2 a/(mb_0) = 2.5 \times 10^5$, $b_0^2/(4ak_B T) = 7$, and $\sqrt{8a/b_0^3}c_0 = 0.4$. The resulting potential, snapshot at times during the erasure substages, is shown in Fig. 1 (inner plot sequence).

Reliable information processing dictates that we set timescales so that the system temporarily, but stably, stores information. To support metastable-quasistatic behavior at all times, the relaxation rates of the informational states are much faster than the rate of change of the potential, keeping the system near metastable equilibrium throughout. The entropy production for such protocols tends to be minimized.

APPENDIX C: TRAJECTORY-CLASS FLUCTUATION THEOREM AND INTERPRETATION

Here, we describe the trajectory-class fluctuation theorems, explaining several of their possible implications and exploring their application to both the simulations and flux qubit experiment. Their derivations are given in Appendix F.

First, we treat each system trajectory $\overrightarrow{z}$ as a function from time between 0 and τ (the time interval of a control protocol) to the set of possible system microstates. Then consider a forward process distribution $\mathcal{P}$, defined by the probabilities of the system microstate trajectories $\overrightarrow{\mathcal{Z}}$ due to an initial equilibrium microstate distribution evolving forward in time under the control protocol. Then, the reverse process distribution $\mathcal{R}$ is determined by preparing the system in equilibrium in the final protocol configuration and running the reverse protocol. The reverse protocol is the original protocol conducted in reverse order but also with objects that are odd under time reversal, like magnetic fields, negated. The time-reversal of a trajectory

is $\overrightarrow{z}^R(t) = [\overrightarrow{z}(\tau - t)]^R$, where for a microstate z the time-reverse z^R is simply z but with time-odd components (e.g., momentum or spin) negated. In other words, time-reversing a trajectory runs the trajectory backwards while also negating all time-odd components of the microstates. For a measurable subset of trajectories $C \subset \overrightarrow{\mathcal{Z}}$, called a *trajectory class*, let the class average $\langle\cdot\rangle_C$ denote an average over the ensemble of forward process trajectories conditioned on the trajectories being in the class C. Let $\mathcal{P}(C)$ and $\mathcal{R}(C^R)$ denote the probabilities of observing the class C in the forward process and the reverse class $C^R = \{\overrightarrow{z}^R | \overrightarrow{z} \in C\}$ in the reverse process, respectively.

We first introduce a *trajectory-class fluctuation theorem* (TCFT) for the *class-averaged exponential work* $\langle e^{-W/k_B T}\rangle_C$:

$$\langle e^{-W/k_B T}\rangle_C = \frac{\mathcal{R}(C^R)}{\mathcal{P}(C)}e^{-\Delta F/k_B T}, \tag{C1}$$

with ΔF the system equilibrium free-energy change. We also introduce a *class-averaged work* TCFT:

$$\langle W\rangle_C = \Delta F + k_B T\bigg(D_{KL}[\mathcal{P}(\overrightarrow{\mathcal{Z}}|C)||\mathcal{R}(\overrightarrow{\mathcal{Z}}^R|C^R)]$$

$$+ \ln\frac{\mathcal{P}(C)}{\mathcal{R}(C^R)}\bigg). \tag{C2}$$

This employs the Kullback-Liebler divergence $D_{KL}[\cdot]$ taken between forward and reverse process distributions over all class trajectories $\overrightarrow{z} \in C$, conditioned on the forward class C and reverse class C^R, respectively. If we disregard this divergence, which is non-negative and can generally be difficult to obtain experimentally, we then find the lower bound $\langle W\rangle_C^{\min}$ on the class-averaged work of Eq. (2).

If we choose the class C to consist of only a single trajectory, we recover detailed fluctuation theorems. For example, Eq. (C1) then simplifies to Crooks' detailed fluctuation theorem [13]:

$$e^{-W(\overrightarrow{z})/k_B T} = \frac{\mathcal{R}(\overrightarrow{z}^R)}{\mathcal{P}(\overrightarrow{z})}e^{-\Delta F/k_B T}. \tag{C3}$$

If, however, we take C to be the entire set of trajectories $\overrightarrow{\mathcal{Z}}$, we recover integral fluctuation theorems. In this case, Eq. (C1) simplifies to Jarzynski's equality [11]:

$$\langle e^{-W/k_B T}\rangle_{\overrightarrow{\mathcal{Z}}} = e^{-\Delta F/k_B T}, \tag{C4}$$

exploiting the fact that $\overrightarrow{\mathcal{Z}}^R = \overrightarrow{\mathcal{Z}}$ and $\mathcal{P}(\overrightarrow{\mathcal{Z}}) = \mathcal{R}(\overrightarrow{\mathcal{Z}}) = 1$.

Furthermore, many other fluctuation theorems can be seen as special cases of the TCFT. In particular, Eq. (9) of Ref. [41] is closely related to Eq. (C1). Having a nearly identical form, the former is a special case of the latter in that the corresponding classes consist of trajectories defined by restrictions on the visited microstates up to only a finite number of times. Similarly, Eqs. (6) and (7) of Ref. [42] derive from Eqs. (C1) and (2) by considering a system with negligible contact with the thermal bath during the protocol and coarse-graining on features of the visited microstate at a single time. Along the same lines, Eqs. (7) and (8) of Ref. [20] are obtained from Eq. (C1) by considering a bit erasure process and trajectory classes corresponding to ending in the target well and in the opposite well, respectively. And, letting the trajectory class be all trajectories that yield a particular value of obtained work

during the forward process, Eq. (C1) reduces to Crooks' work fluctuation theorem [13]:

$$\frac{\mathcal{P}(W)}{\mathcal{R}(-W)} = e^{(W-\Delta F)/k_B T}, \tag{C5}$$

where $\mathcal{P}(W)$ and $\mathcal{R}(-W)$ are the probabilities of obtaining values W and $-W$ for the work when running the forward and reverse process, respectively. Finally, yet other fluctuation theorems arise directly from the TCFT by particular class choices [16,27,43,44].

In this way, one sees that the TCFT is a suite that spans the space of fluctuation theorems between the extreme of the detailed theorems, which require very precise information about an individual trajectory, and the integral theorems, which describe the system's entire trajectory ensemble. It thus unifies a wide range of existing (and future) fluctuation theorems. Appendix F below provides proofs.

APPENDIX D: EMPIRICAL USE IN STATISTICAL ESTIMATION

Beyond the synthesis of distinct fluctuation theorems, the TCFT is empirically useful in greatly improving sampling and errors in statistical estimation. And, this is its primary role here—a diagnostic tool for thermodynamic computing. We can rearrange Eq. (C1) to obtain Eq. (3)—an expression for estimating equilibrium free-energy changes:

$$\Delta F = -k_B T \ln \left(\frac{\mathcal{P}(C)}{\mathcal{R}(C^R)} \langle e^{-W/k_B T} \rangle_C \right). \tag{D1}$$

Thus, to estimate free energy one sees that statistics are needed for only one particular class and its reverse. Generally, this gives a substantial statistical advantage over direct use of Jaryznski's equality:

$$\Delta F = -k_B T \ln \langle e^{-W/k_B T} \rangle_{\overrightarrow{\mathcal{Z}}},$$

since rare microstate trajectories may generate negative work values that dominate the average exponential work [29]. The problem is clear in the case of erasure. Recall from Fig. 2 (three front panels) that Fail trajectories generate the most negative work values. In the limit of higher success-rate protocols that maintain low entropy production, failures generate more and more negative works, leading them to dominate when estimating average exponential works despite becoming negligible in probability.

In contrast, to efficiently determine the change in equilibrium free energy from Eq. (3), its form indicates that one should choose a class that (i) is common in the forward process, (ii) has a reverse class that is common in the reverse process, and (iii) generates a narrow work distribution. This maximizes the accuracy of statistical estimates for the three factors on the right-hand side. For example, while the equilibrium free-energy change in the case of our erasure protocol is theoretically simple (zero), the Success class fits the criteria.

We can then monitor the class-averaged work in excess of its bound:

$$E_C = \langle W \rangle_C - \langle W \rangle_C^{\min}$$

$$= k_B T D_{KL} \left[\mathcal{P}(\overrightarrow{\mathcal{Z}} | C) || \mathcal{R}(\overrightarrow{\mathcal{Z}}^R | C^R) \right] \geqslant 0.$$
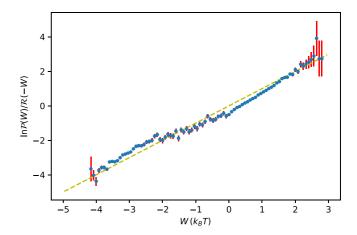

FIG. 4. Flux qubit experiment work fluctuations: Crooks' work fluctuation theorem prediction (green dashed line), measured values (blue), and 1-σ statistical errors (red).

The inequality in Eq. (2) is a refinement of the equilibrium Second Law and therefore the bound $\langle W \rangle_C^{\min}$ generally provides a more accurate estimate of the average work of trajectories in a class compared to the equilibrium free-energy change ΔF. More precisely, as we will see below, an average of the excess E_C over all classes C in a partition of trajectories must be smaller than the dissipated work $\langle W \rangle - \Delta F$. For trajectory classes with narrow work distributions, this can be a significant improvement. We can see this by Taylor expanding the left-hand side of Eq. (C1) about the mean dimensionless work $\langle W/k_B T \rangle_C$. This shows that Eq. (2) becomes an equality when the variance and higher moments vanish. Appendix G below delves more into moment approximations. In any case, trajectory classes with narrow work distributions have small excess works E_C.

APPENDIX E: FLUCTUATIONS IN THERMODYNAMIC COMPUTING: THE CASE OF ERASURE

Before applying the TCFT to analyze thermodynamic fluctuations during erasure, we first explore both Jarzynski's equality Eq. (C4) and Crooks' work fluctuation theorem Eq. (C5).

Since the erasure protocol is cyclic, the change in equilibrium free energy ΔF vanishes. Jarzynski's equality then predicts that the average exponential work $\langle e^{-W/k_B T} \rangle_{\overrightarrow{\mathcal{Z}}}$ must be 1. From simulation, we obtain a value of $1.0025 \pm 5 \times 10^{-5}$, which is very close to the predicted value. From experiment, we obtain a value of $0.89 \pm 5 \times 10^{-5}$—within 10% of the prediction, but it falls somewhat outside the expected error.

Crooks' work fluctuation theorem predicts that the quantity $\ln(\mathcal{P}(W)/\mathcal{R}(-W))$ must equal βW at each W. We verify this experimentally by building probability histograms for $\mathcal{P}(W)$ and $\mathcal{R}(-W)$, taking their log ratios, and plotting against their binned work values expressed in units of $k_B T$. Figure 4 shows that the experiment follows the theoretical prediction quite closely. However, as for the case of Jarzynski's equality, the experimental results are not all within expected errors. The discrepancies appear to arise in the statistical errors in

TABLE II. Class-average works and bounds for different trajectory classes during erasure: *All* trajectories $\overrightarrow{\mathcal{Z}}$, *Success* trajectories, *Fail* trajectories, and *Transitional* trajectories. These are identified in Fig. 2 (four front panels). From left to right, columns give the estimated class-average work $\langle W \rangle_C$, TCFT lower bound $\langle W \rangle_C^{\min}$, and their difference E_C. 3.5×10^6 simulations were run for each of the forward and reverse processes, with 96.2% trajectories successfully ending in the R informational state under the forward process.

Class C	$\langle W \rangle_C$	$\langle W \rangle_C^{\min}$	E_C
All $\overrightarrow{\mathcal{Z}}$	0.634	0.0	0.634
Success	0.713	0.683	0.030
Fail	-3.885	-3.951	0.066
Transitional	-0.546	-1.650	1.170

estimating the obtained work values for each trajectory. Due to the complicated nature of the approximations (see Appendix H 3), estimating the corresponding error is challenging and will be left for future detailed investigation.

We now turn to analyze work fluctuations in various trajectory classes during the erasure operation, demonstrating that the TCFT allows analyzing the thermodynamics of trajectories falling between the Jarzynski and Crooks extreme classes. The main point here is that the classes between these extremes consist of "functionally" interpretable trajectories— e.g., successful and failed erasure. In this way, one can diagnose the energetics and general thermodynamics of this functioning in the physical computing device.

To estimate $\mathcal{R}(C^R)$ for the three chosen classes $C \in$ {Success, Failure, Transitional} of our simulated Erasure process, we ran 3.5×10^6 simulations of the reverse process. Table II shows that the Success and Fail classes have small excesses, and, as seen in Fig. 2 (three front panels), these classes indeed have narrow work distributions. Elsewhere, we explore these and additional partition schemes, finding that the Transitional trajectories can be further partitioned to yield narrow work distributions so that all trajectory classes have small excesses E_C. In short, this demonstrates how well-formulated trajectory classes allow accurate estimates of the works for all trajectories.

To measure the efficacy of a given partition Q of trajectories into classes, we ask what the ensemble-average of class-average excess works is:

$$
\begin{aligned}
E_Q &= \sum_{C \in Q} \mathcal{P}(C) E_C \\
&= \langle W \rangle_{\overrightarrow{\mathcal{Z}}} - \sum_{C \in Q} \mathcal{P}(C) \langle W \rangle_C^{\min} \\
&= \langle W \rangle_{\overrightarrow{\mathcal{Z}}} - \langle W \rangle_Q^{\min},
\end{aligned}
$$

with $\langle W \rangle_Q^{\min} = \sum_{C \in Q} \mathcal{P}(C) \langle W \rangle_C^{\min}$.

From Eq. (2), we see that $\langle W \rangle_Q^{\min}$ is the coarse-grained lower bound on ensemble-average dissipation from Ref. [45]:

$$
\langle W \rangle_Q^{\min} = \Delta F + k_B T D_{KL}[\mathcal{P}(Q) || \mathcal{R}(Q^R)],
$$

where $D_{KL}[\cdot]$ is the Kullback-Liebler divergence between forward and reverse process distributions over the trajec-

TABLE III. Ensemble-average work and bounds due to different partitions: *Trivial* partition; *Untilt-Centric I* partition, composed of Success, Fail, and Transitional; and *Untilt-Centric II* partition, described in follow-on work. From left to right, columns give the estimated ensemble-average work, the partition bound $\langle W \rangle_Q^{\min}$, and their difference E_Q. All values in units of $k_B T$.

Partition Q	$\langle W \rangle_{\overrightarrow{\mathcal{Z}}}$	$\langle W \rangle_Q^{\min}$	E_Q
Trivial $\{\overrightarrow{\mathcal{Z}}\}$	0.634	0.0	0.634
Untilt-Centric I	0.634	0.560	0.074
Untilt-Centric II	0.634	0.601	0.032

tory classes $C \in Q$. Since Kullback-Liebler divergences are nonnegative, such a bound always provides an improvement over the equilibrium Second Law. Table III shows both $\langle W \rangle_Q^{\min}$ and E_Q for the trivial partition $\{\overrightarrow{\mathcal{Z}}\}$, where $\langle W \rangle_Q^{\min} = \Delta F$, our three-class partition, labeled *Untilt-Centric I*, and the improved partition described in follow-on work, labeled *Untilt-Centric II*. Application of the first partition simply implies the equilibrium Second Law. In this case, the latter two also improve on the nonequilibrium Second Law that, calculated by assuming metastable starting and ending distributions, provides a lower bound on the average work equal to 0.533, the change in nonequilibrium free energy ΔF.

We can appeal to Landauer's erasure bound—$k_B T \ln 2 \approx 0.693 \, k_B T$—to calibrate the excesses E_C and E_Q. We see for the simulation data that our three-class partition Untilt-Centric I provides class-average work bounds that, on average, are only about 11% of $k_B T \ln 2$ from the actual class-average works. The more refined Untilt-Centric II partition reduces this excess to about 5% while the trivial partition fails by about 91% of $k_B T \ln 2$.

We also recover the equality of Ref. [45] for the ensemble-average work by averaging Eq. (C2) over each class:

$$
\begin{aligned}
\langle W \rangle &= \sum_C \mathcal{P}(C) \langle W \rangle_C \\
&= \Delta F + k_B T \bigg(\sum_{C \in Q} \mathcal{P}(C) D_{KL}[\mathcal{P}(\overrightarrow{\mathcal{Z}} | C) || \mathcal{R}(\overrightarrow{\mathcal{Z}}^R | C^R)] \\
&\quad + D_{KL}[\mathcal{P}(Q) || \mathcal{R}(Q^R)] \bigg),
\end{aligned}
$$

which of course is lower bounded by $\langle W \rangle_Q^{\min}$.

These results suggest the criterion for optimal trajectory partitions: Select a partition sufficiently refined to yield tight bounds on class-average works, but no finer. Machine learning methods for model-order selection provide a basis for a natural classification scheme for trajectories that captures all relevant thermodynamics and information processing. Moreover, by changing the forward and reverse processes $\mathcal{P}$ and $\mathcal{R}$ to begin in system microstate distributions other than equilibrium, a still broader class of TCFTs emerge. We can then find analogous results for heats and comparisons with works and nonequilibrium free-energy changes. We explore these extensions in depth elsewhere.

APPENDIX F: TCFT DERIVATIONS

We now present derivations for the two TCFTs introduced in Eqs. (C1) and (C2).

Assume that the system dynamics is described by a Hamiltonian specified in part by an external control protocol, as well as by a weak coupling to a thermal environment that induces steady relaxation to canonical equilibrium.

Start the system in equilibrium distribution π_0 for Hamiltonian $\mathcal{H}_0$ and run a protocol until time τ, causing the system Hamiltonian to evolve to $\mathcal{H}_\tau$. If we then hold the Hamiltonian at $\mathcal{H}_\tau$ for a long time, the system relaxes into the equilibrium distribution π_τ. The system's ensemble entropy change from $t = 0$ to $t = \infty$ is then

$$\Delta S_{\text{sys}} = \sum_z [-\pi_\tau(z) \ln \pi_\tau(z) + \pi_0(z) \ln \pi_0(z)].$$

The trajectorywise system entropy difference is defined to be

$$\Delta s_{\text{sys}}(\vec{z}) = \ln \frac{\pi_0(z_0)}{\pi_\tau(z_\tau)},$$

where z_0 and z_τ are the initial and final microstates of system microstate trajectory $\vec{z}$, respectively. Averaged over all trajectories $\vec{z} \in \vec{\mathcal{Z}}$, this then becomes the ensemble entropy change.

Let $p(\vec{z} | z_0)$ denote the probability of obtaining system microstate trajectory $\vec{z}$ via the protocol conditioned on starting the system in state $z_0 = \vec{z}(0)$.

Now, start the system Hamiltonian at $\mathcal{H}_\tau$ and run the reverse protocol, ending the Hamiltonian at $\mathcal{H}_0$. We then obtain the trajectory $\vec{z}$ with a different conditional probability: $r(\vec{z} | z_0)$.

Assuming microscopic reversibility and given a system trajectory $\vec{z}$, the change in the heat bath's entropy is

$$\Delta S_{\text{res}}(\vec{z}) = -\beta Q(\vec{z}) \tag{F1}$$

$$= \ln \frac{p(\vec{z} | z_0)}{r(\vec{z}^{\text{R}} | z_\tau^{\text{R}})}, \tag{F2}$$

where $\beta = 1/k_B T$, $Q(\vec{z})$ is the net energy that flows out of the heat bath into the system given the trajectory $\vec{z}$, and $(\cdot)^{\text{R}}$ denotes time-reversal. This holds for systems with strictly finite energies and Markov dynamics that induce the equilibrium distribution when control parameters are held fixed [46]. Both our simulated Duffing potential system and flux qubit obey these requirements at sufficiently short timescales. Then we can express the total trajectorywise change in entropy production due to a trajectory $\vec{z}$ as the sum of system and heat reservoir entropy changes:

$$\Delta S_{\text{tot}}(\vec{z}) = \Delta s_{\text{sys}}(\vec{z}) + \Delta S_{\text{res}}(\vec{z}).$$

Since $\pi_t(z) = e^{-\beta(\mathcal{H}_t(z) - F_t)}$, with F_t the system's equilibrium free energy at time t, we can write

$$\Delta S_{\text{tot}}(\vec{z}) = -\ln \pi_\tau(z_\tau) + \ln \pi_0(z_0) - \beta Q(\vec{z})$$
$$= \beta(\mathcal{H}_\tau(z_\tau) - F_\tau) - \beta(\mathcal{H}_0(z_0) - F_0) - \beta Q(\vec{z})$$
$$= \beta(\Delta \mathcal{H}(\vec{z}) - Q(\vec{z}) - \Delta F)$$
$$= \beta(W(\vec{z}) - \Delta F).$$

Using Eq. (F1), we also have

$$\Delta S_{\text{tot}}(\vec{z}) = \Delta s_{\text{sys}}(\vec{z}) + \Delta S_{\text{res}}(\vec{z})$$
$$= \ln \frac{\pi_0(z_0)}{\pi_\tau(z_\tau)} \frac{p(\vec{z} | z_0)}{r(\vec{z}^{\text{R}} | z_\tau^{\text{R}})}$$
$$= \ln \frac{\mathcal{P}(\vec{z})}{\mathcal{R}(\vec{z}^{\text{R}})}$$

with $\mathcal{P}(\vec{z}) = \pi_0(z_0) p(\vec{z} | z_0)$ and
$$\mathcal{R}(\vec{z}^{\text{R}}) = \pi_\tau(z_\tau) r(\vec{z}^{\text{R}} | z_\tau^{\text{R}}).$$

Combining, we obtain Crooks' detailed fluctuation theorem [13]:

$$\mathcal{R}(\vec{z}^{\text{R}}) = \mathcal{P}(\vec{z}) e^{-\beta(W(\vec{z}) - \Delta F)}. \tag{F3}$$

From here, we derive our first TCFT by integrating each side of Eq. (F3) over all trajectories $\vec{z}$ in a measurable set $C \subset \vec{\mathcal{Z}}$. Starting with the left-hand side and recalling the Iverson bracket $[\cdot]$, which is 1 when the interior expression is true and 0 when false, we have

$$\int d\vec{z} [\vec{z} \in C] \mathcal{R}(\vec{z}^{\text{R}}) = \int d\vec{z}^{\text{R}} [\vec{z} \in C] \mathcal{R}(\vec{z}^{\text{R}})$$
$$= \int d\vec{z}^{\text{R}} [\vec{z}^{\text{R}} \in C^{\text{R}}] \mathcal{R}(\vec{z}^{\text{R}})$$
$$= \int d\vec{z} [\vec{z} \in C^{\text{R}}] \mathcal{R}(\vec{z})$$
$$= \mathcal{R}(C^{\text{R}}).$$

The first three steps used the unity of the Jacobian in reversing a microstate, the definition $C^{\text{R}} = \{\vec{z}^{\text{R}} | \vec{z} \in C\}$, and swapping all instances of $\vec{z}^{\text{R}}$ with $\vec{z}$, respectively. Integrating the right-hand side of Eq. (F3) then gives

$$\int d\vec{z} [\vec{z} \in C] \mathcal{P}(\vec{z}) e^{-\beta(W(\vec{z}) - \Delta F)}$$
$$= e^{\beta \Delta F} \int d\vec{z} \mathcal{P}(\vec{z}, C) e^{-\beta W(\vec{z})}$$
$$= \mathcal{P}(C) e^{\beta \Delta F} \int d\vec{z} \mathcal{P}(\vec{z} | C) e^{-\beta W(\vec{z})}$$
$$= \mathcal{P}(C) e^{\beta \Delta F} \langle e^{-\beta W} \rangle_C.$$

Combining, we have our first TCFT, Eq. (C1).

To obtain the second TCFT, we first change the form of Eq. (F3):

$$W(\vec{z}) = \Delta F + \beta^{-1} \ln \frac{\mathcal{P}(\vec{z})}{\mathcal{R}(\vec{z}^{\text{R}})}.$$

Then we calculate the class-average. The equilibrium free-energy change is unaffected while the rightmost term becomes

$$
\begin{aligned}
\beta^{-1}\left\langle \ln \frac{\mathcal{P}(\vec{z})}{\mathcal{R}(\vec{z}^{\,\mathrm{R}})} \right\rangle_C &= \beta^{-1} \int_C d\vec{z}\, \mathcal{P}(\vec{z}\,|C) \ln \frac{\mathcal{P}(\vec{z})}{\mathcal{R}(\vec{z}^{\,\mathrm{R}})} \\
&= \beta^{-1} \int_C d\vec{z}\, \mathcal{P}(\vec{z}\,|C) \ln \frac{\mathcal{P}(\vec{z}\,|C)\mathcal{P}(C)}{\mathcal{R}(\vec{z}^{\,\mathrm{R}}|C^{\mathrm{R}})\mathcal{R}(C^{\mathrm{R}})} \\
&= \beta^{-1}\left(\int_C d\vec{z}\, \mathcal{P}(\vec{z}\,|C) \ln \frac{\mathcal{P}(\vec{z}\,|C)}{\mathcal{R}(\vec{z}^{\,\mathrm{R}}|C^{\mathrm{R}})} + \ln \frac{\mathcal{P}(C)}{\mathcal{R}(C^{\mathrm{R}})} \right) \\
&= \beta^{-1}\left(D_{\mathrm{KL}}\big[\mathcal{P}(\vec{z}\,|C)||\mathcal{R}(\vec{z}^{\,\mathrm{R}}|C^{\mathrm{R}})\big] + \ln \frac{\mathcal{P}(C)}{\mathcal{R}(C^{\mathrm{R}})} \right),
\end{aligned}
$$

which gives Eq. (C2)'s TCFT.

APPENDIX G: CLASS-AVERAGED WORK APPROXIMATION FOR NARROW DISTRIBUTIONS

Here, we demonstrate that the class-averaged work $\langle W \rangle_C$ approaches its bound $\langle W \rangle_C^{\min}$ when the variance and higher moments of the class' distribution of works vanish. One concludes that $\langle W \rangle_C^{\min}$ is a good approximation for $\langle W \rangle_C$ when the class' work distribution is narrow.

We first express the left-hand side of Eq. (C1) in terms of the unitless distance of work from its class-average:

$$
\langle e^{-\beta W} \rangle_C = \langle e^{-x} \rangle_C\, e^{-\beta \langle W \rangle_C},
$$

with $x = \beta(W - \langle W \rangle_C)$. Then, we Taylor-expand the exponential inside the class-average:

$$
\begin{aligned}
\langle e^{-x} \rangle_C &= \sum_{n=0}^{\infty} \frac{(-1)^n}{n!} \langle x^n \rangle_C \\
&= 1 + a,
\end{aligned}
$$

with $a = \sum_{n=2}^{\infty} \frac{(-1)^n}{n!} \langle x^n \rangle_C$. Equation (C1) then gives

$$
(1+a)e^{-\beta \langle W \rangle_C} = \frac{\mathcal{R}(C^{\mathrm{R}})}{\mathcal{P}(C)} e^{-\beta \Delta F}.
$$

Since e^{-x} is convex,

$$
(1+a) = \langle e^{-x} \rangle_C \geqslant e^{-\langle x \rangle_C} = 1,
$$

so $a \geqslant 0$. Then

$$
\begin{aligned}
\langle W \rangle_C &= \Delta F + \beta^{-1} \ln \frac{\mathcal{P}(C)}{\mathcal{R}(C^{\mathrm{R}})} + \beta^{-1} \ln(1+a) \\
&\geqslant \Delta F + \beta^{-1} \ln \frac{\mathcal{P}(C)}{\mathcal{R}(C^{\mathrm{R}})} \\
&= \langle W \rangle_C^{\min}.
\end{aligned}
$$

The second line becomes an equality when a goes to zero, which occurs as the variance and higher moments vanish.

APPENDIX H: WORK DEFINITIONS AND EXPERIMENTAL ESTIMATION

Properly estimating the required works and devolved heats from experimental devices undergoing cyclic control protocols requires explicitly and consistently accounting for energy and information flows between the system, its environment,

and the controlling laboratory apparatus. To this end, we construct a model Hamiltonian universe for common processes involving small systems interacting with laboratory apparatus and a thermal environment. After deriving key equalities for two definitions of work, the inclusive and exclusive works, we define a method of approximating them in appropriate cyclic protocols.

1. The model universe and Hamiltonian

To study a small system that exchanges energy with its environment in the forms of heat and work, we introduce a model universe: a *system of interest*, a *heat bath*, and a *lab* (laboratory apparatus) that controls the system and derives any needed energy from a *work reservoir*. The system directly interacts with both the heat bath and the laboratory, but the heat bath and lab are not directly coupled.

We assume that a Hamiltonian $\mathcal{H}$ describes the universe's evolution and that there is a set of generalized coordinates that can be sensibly partitioned into those for the system, heat bath, and laboratory. Then, we decompose the universe Hamiltonian into the following form:

$$
\mathcal{H}(s, b, l) = H_{\mathrm{B}}(b) + h_{\mathrm{S,B}}(s, b) + H_{\mathrm{S}}(s) + h_{\mathrm{S,L}}(s, l) + H_{\mathrm{L}}(l),
$$

where s, b, and l denote both the generalized coordinates and conjugate momenta for the system, bath, and laboratory, respectively. For any universe Hamiltonian $\mathcal{H}$, there can be many choices for this decomposition.

We also define the system Hamiltonian $\mathcal{H}'$ as the three components that depend on the system coordinates:

$$
\mathcal{H}'(s; b, l) = h_{\mathrm{S,B}}(s, b) + H_{\mathrm{S}}(s) + h_{\mathrm{S,L}}(s, l).
$$

First, consider the subset of laboratory coordinates l for which $h_{\mathrm{S,L}}$ has nontrivial dependence. These so-called *protocol parameters* λ are often simple and much fewer than the entire set of l. We often assume that we have total control of their evolution. More precisely, under an appropriate preparation for the laboratory at time $t = 0$, a specific trajectory for the protocol parameters $\{\lambda(t)\}_t$ for $0 \leqslant t \leqslant \tau$ is guaranteed for all preparations of the heat bath and system coordinates. We refer to the parameter trajectory as the *protocol*.

Suppose the heat-bath degrees of freedom that interact with the system change much faster than the system's. We can assume that the system response to the bath resembles Brownian motion. On the timescale of changes in the system coordinates, then, we ignore the system-bath interaction term

$h_{S,B}$ in writing the system Hamiltonian $\mathcal{H}'$:

$$\mathcal{H}'(s;\lambda) = H_S(s) + h_{S,L}(s,\lambda)$$
$$= T(s) + V(s,\lambda).$$

The latter decomposition into kinetic energy T and potential energy V can be used to write Langevin equations of motion for the system. Furthermore, if the heat bath has a relaxation time sufficiently short that it is roughly in equilibrium at all times with fixed temperature, then its influence on the system will be memoryless.

2. Inclusive and exclusive works and heats

The basic scenario for executing a protocol is as follows. The universe coordinates begin according to a given initial distribution $\Pr(s)$ at time $t = 0$, and they evolve in isolation until $t = \tau$. As above, we assume that a well-defined protocol $\{\lambda_t\}_t$ emerges due to our preparation of the lab coordinates.

We label all energy exchanged between the system and laboratory as *work* and all energy exchanged between the system and heat bath as *heat*. Since the laboratory is directly coupled only to the system, the work they exchange is given by the change in energy of the laboratory's work reservoir. Similarly, since the heat bath is directly coupled only to the system, the heat exchanged is given by the change in the heat bath's energy.

Note that this requires choices as to what constitute the energies of the three universe subsystems. While H_B, H_S, and H_L define energies for the heat bath, system, and work reservoir, respectively, what of $h_{S,L}$ and $h_{S,B}$? If all subsystems were macroscopic, these interaction terms would be negligible. While it may be desirable to assume that the system is only weakly coupled to the heat bath—so that $h_{S,B}$ can be ignored—$h_{S,L}$ can be significant in many important small systems.

And so, in general, we define the system energy to be H_S plus any portions of $h_{S,L}$ and $h_{S,B}$. Then the work reservoir energy is H_L plus the rest of $h_{S,L}$, while the heat bath energy is H_B plus the rest of $h_{S,B}$. To make these distinctions clear, we label two types of works, each corresponding to the two extremes for allocation of $h_{S,L}$ between the system and work reservoir: the *inclusive work* W and the *exclusive work* W_0 [47]. Specifically,

$$\frac{dW}{dt} = -\frac{d}{dt}(H_L) = \frac{d}{dt}(H_B + h_{S,B} + H_S + h_{S,L}),$$
$$\frac{dW_0}{dt} = -\frac{d}{dt}(h_{S,L} + H_L) = \frac{d}{dt}(H_B + h_{S,B} + H_S).$$

We can similarly define the *inclusive heat* Q and *exclusive heat* Q_0 depending on how we allocate $h_{S,B}$ between the system and heat bath:

$$\frac{dQ}{dt} = -\frac{d}{dt}(H_B) = \frac{d}{dt}(h_{S,B} + H_S + h_{S,L} + H_L),$$
$$\frac{dQ_0}{dt} = -\frac{d}{dt}(H_B + h_{S,B}) = \frac{d}{dt}(H_S + h_{S,L} + H_L).$$

The inclusive work corresponds to fully including $h_{S,L}$ in the system energy, while the exclusive work corresponds to excluding it. Inclusive and exclusive heat correspond similarly with respect to $h_{S,B}$.

There is a key relation between the inclusive and exclusive works:

$$\frac{dW}{dt} = \frac{dW_0}{dt} + \frac{dh_{S,L}}{dt}. \tag{H1}$$

That is, the inclusive work for an interval of time equals the sum of the exclusive work and the change in the system-lab interaction term $h_{S,L}$.

In the above expressions, calculating the rate of change of a work or heat requires the time derivative of one or more of H_L and H_B. This can be problematic. Fortunately, there are alternate forms that are amenable. One can show that the inclusive work rate is given by

$$\frac{dW}{dt} = -\frac{dH_L}{dt}$$
$$= \frac{\partial h_{S,L}}{\partial \lambda}\frac{d\lambda}{dt}. \tag{H2}$$

This is a more common definition for the work rate in small-system nonequilibrium thermodynamics. And, it allows the work to be calculated as

$$W(t,t') = \int_t^{t'} dt'' \frac{d\lambda}{dt''} \frac{\partial h_{S,L}(s,\lambda)}{\partial \lambda}\Big|_{s=s(t''),\lambda=\lambda(t'')}. \tag{H3}$$

The exclusive work W_0 has a corresponding form:

$$\frac{dW_0}{dt} = -\frac{d(h_{S,L} + H_L)}{dt}$$
$$= -\frac{\partial h_{S,L}}{\partial s}\frac{ds}{dt}. \tag{H4}$$

For the case in which $h_{S,L}$ is a scalar potential for s, this is the product of the corresponding force with velocity. This makes the exclusive work equal to a familiar mechanics definition of work as the integral of the dot product of force and displacement:

$$W_0(t,t') = -\int_t^{t'} dt'' \frac{ds}{dt''} \frac{\partial h_{S,L}(s,\lambda)}{\partial s}\Big|_{s=s(t''),\lambda=\lambda(t'')}.$$

In this way, we write the inclusive and exclusive work rates in terms of the rates of change of the system and work-reservoir interaction term $h_{S,L}$ with respect to either the system or work reservoir coordinates.

3. Approximating inclusive work experimentally

For the flux qubit experimental system investigated here, we assume the following:

$$H_S(s) + h_{S,L}(s,\lambda) = T(s) + V(s,\lambda).$$

That is, as far as the flux qubit and work reservoir are concerned, the only relevant energies at least partially ascribable to the flux qubit are its kinetic energy and the potential energy with the work reservoir. $h_{S,L}$ must then capture the change in the potential V due to changes in the protocol parameters. We could simply define $h_{S,L}(s,\lambda) = V(s,\lambda)$ so that $H_S(s) = T(s)$. However, it is more useful to allocate the initial potential energy to H_S. That is,

$$H_S(s) = T(s) + V(s,\lambda_0) \text{ and}$$
$$h_{S,L}(s,\lambda) = V(s,\lambda) - V(s,\lambda_0).$$

For cyclic protocols where $V(\cdot, \lambda_0) = V(\cdot, \lambda_\tau)$, such as in our erasure operation, $h_{\mathrm{S,L}}(s(t), \lambda(t))$ vanishes for all trajectories at $t = 0, \tau$. By Eq. (H1) we then have the useful equality $W = W_0$ between inclusive and exclusive works taken over the entire protocol.

Estimating W for a system trajectory is then equivalent to estimating W_0 for the cyclic protocols we consider. In the flux qubit, the form of $h_{\mathrm{S,L}}$ is known and the specific protocol $\{\lambda_t\}_{t \in [0,\tau]}$ is known. Unfortunately, we lack sufficient information about its instantaneous state s at all times to estimate the total inclusive work from Eq. (H3), since the device's physics precludes precise measurements of system flux ϕ—the relevant part of s for determining the potential $h_{\mathrm{S,L}}$. Instead, we do have reliable measurement of large and stable changes in the flux ϕ. This specifically monitors when the system moves between wells in a double-well potential $V(\cdot, \lambda(t))$, if the rate of transition between wells is sufficiently slow.

And so, we can use information about the flux ϕ to approximate the exclusive work contribution at each moment in time. Then, adding up these contributions yields an approximation of the total exclusive work W_0 over the entire protocol and therefore of the inclusive work W over the entire protocol. Note that the protocols used here maintain two wells at all times for the system flux ϕ. We develop the approximation in two steps.

a. First-order approximation

We first partition the potential in flux space into three segments. Two segments constitute the wells for the flux in which that state spends all its time, except for very brief transitions between wells. Then, the third segment connects the two wells, capturing the dynamics arising from crossing the barrier that separates them.

We require that the partitioning allows the following two approximations. First, the particle spends negligible total duration in between the two wells. Second, the wells do not change shape over the protocol, but instead simply raise or lower in potential at different times, if they change at all. This means that the shape of the system-lab interaction term $h_{\mathrm{S,L}}(\cdot, \lambda(t))$ at any time t is very simple in the two wells—flat.

The result is that the exclusive work over any time duration is easily calculable from the experimental data. During times when the flux remains in a well, the exclusive work rate must be zero, since $h_{\mathrm{S,L}}$ does not change with s. During a transition,

the shape of $h_{\mathrm{S,L}}$ does not change significantly due to the first approximation. Then, the exclusive work during a transition is the difference in heights of the two wells as measured by $h_{\mathrm{S,L}}$:

$$
\begin{aligned}
\Delta W_0^{\mathrm{trans}} &= \int d\phi \left(-\frac{\partial h_{\mathrm{S,L}}}{\partial \phi} \right) \\
&\approx h_{\mathrm{S,L}}(w_0, \lambda(t)) - h_{\mathrm{S,L}}(w_1, \lambda(t)),
\end{aligned} \tag{H5}
$$

where $\lambda(t)$ is the protocol parameter setting at any time during the transition, and w_0 and w_1 are arbitrary flux values in the starting and ending wells, respectively.

Thus, the total inclusive work W over the protocol for a trajectory is simply the sum of the jump contributions above for each transition.

b. Second-order approximation

In point of fact, the potential wells do change shape. Fortunately, our method for calculating the inclusive work over the protocol remains valid under weaker constraints on the protocol.

We first require the protocol to maintain two metastable regions, the *informational states*, at all times, each possessing a unique local potential minimum continuously in time. We denote the flux value at the potential minima of informational state i at time t as ϕ_i^t. The protocol must also evolve slowly enough so that the potential landscape changes slowly compared to the system's relaxation rate in each metastable region. Both of these criteria are met by our erasure protocol.

Consider a short duration Δt during which the potential $V(\cdot, t)$ changes little but long compared to the relaxation rates of the informational states. Consider two cases: either the system crosses the barrier between the two informational states during this time, or it remains in one informational state.

First, suppose that the system transitions from one informational state, i, to the other, j. Denote the system flux at the beginning of the transition as ϕ^t and at the end as $\phi^{t+\Delta t}$. By Eq. (H1), the exclusive work contribution $\Delta W_0^{\mathrm{trans}}$ is the difference of the inclusive work contribution and the change in system-lab interaction term $\Delta h_{\mathrm{S,L}}$. Our protocol ensures that the total number of transitions is so small and the time durations Δt so narrow that we can ignore the total contributions of inclusive works $\Delta W^{\mathrm{trans}}$ due to these transition durations. The change $\Delta h_{\mathrm{S,L}}$ can itself be broken down into two terms, one for the difference in $h_{\mathrm{S,L}}$ between the informational-state minima, and the other for the change in $h_{\mathrm{S,L}}$ local to the respective minima. In other words,

$$
\begin{aligned}
\Delta h_{\mathrm{S,L}} &= h_{\mathrm{S,L}}(\phi^{t+\Delta t}) - h_{\mathrm{S,L}}(\phi^t) \\
&= \left[\left(h_{\mathrm{S,L}}(\phi^{t+\Delta t}) - h_{\mathrm{S,L}}\left(\phi_j^{t+\Delta t}\right) \right) + h_{\mathrm{S,L}}\left(\phi_j^{t+\Delta t}\right) \right] - \left[\left(h_{\mathrm{S,L}}(\phi^t) - h_{\mathrm{S,L}}\left(\phi_i^t\right) \right) + h_{\mathrm{S,L}}\left(\phi_i^t\right) \right] \\
&= \left[h_{\mathrm{S,L}}\left(\phi_j^{t+\Delta t}\right) - h_{\mathrm{S,L}}\left(\phi_i^t\right) \right] + \left[\left(h_{\mathrm{S,L}}(\phi^{t+\Delta t}) - h_{\mathrm{S,L}}\left(\phi_j^{t+\Delta t}\right) \right) - \left(h_{\mathrm{S,L}}(\phi^t) - h_{\mathrm{S,L}}\left(\phi_i^t\right) \right) \right] \tag{H6}
\end{aligned}
$$

$$
= \Delta m^t + \Delta l^t, \tag{H7}
$$

where Δm^t, Eq. (H6)'s first term, is the change in $h_{\mathrm{S,L}}$ at the informational-state minima, and Δl^t, Eq. (H6)'s second

term, is the change in $h_{\mathrm{S,L}}$ of the system with respect to the informational-state minima. We can therefore approximate the

exclusive work contribution during a transition as follows:

$$\Delta W_0^{\text{trans}} = -\Delta h_{\text{S,L}}$$
$$= -\Delta m^t - \Delta l^t.$$

Suppose now that the system remains in one informational state i during a time interval Δt. Since the relaxation rate is fast compared to the duration Δt, we assume that the system visits all microstates in the informational state roughly in proportion to the local equilibrium distribution. Then, the inclusive work contribution ΔW^{stay} is approximately independent of the specific system trajectory during this time, and instead is determined by the time duration and the informational state i. If during this time we had additionally simultaneously shifted the entire potential up or down by a given amount, we would have added an inclusive work contribution equal to the potential shift, but the system trajectory would have remained unchanged since adding a term constant over position to the potential does not affect dynamics. Thus, the actual inclusive work contribution is equal to an amount due to the change in the system-lab interaction term at the informational-state minimum plus an amount due solely to the change in potential shape at the informational state with respect to its minimum. That is,

$$\Delta W^{\text{stay}} = \Delta W_s + \Delta m^t, \tag{H8}$$

where ΔW_s is the inclusive work contribution due to the change in potential shape at the informational state, and Δm^t is defined as above. Equation (H7) applies equally well here in describing the change in system-lab interaction term. Thus, first using Eq. (H1) once more, the exclusive work contribution ΔW_0^{stay} for a time interval where the particle remains in one informational state is as follows:

$$\Delta W_0^{\text{stay}} = \Delta W^{\text{stay}} - \Delta h_{\text{S,L}} \tag{H9}$$

$$= \Delta W_s - \Delta l^t. \tag{H10}$$

The result is that we have exclusive work contributions for both durations when the system transitions between informational states and when it remains in one.

To find the total exclusive work over the protocol for a given trajectory, we add up these contributions. The sum of all local $h_{\text{S,L}}$ changes Δl^t over all durations is the net local change in $h_{\text{S,L}}$. Recall, though, that the minima of the informational states begin and end at the same values. And so, the total local change in $h_{\text{S,L}}$ reduces to the absolute change in $h_{\text{S,L}}$. However, since we chose $h_{\text{S,L}}(\cdot, 0) = h_{\text{S,L}}(\cdot, \lambda(t)) = 0$, this must vanish:

$$\sum_t \Delta l^t = 0.$$

We can now specify our final approximation: For any time interval Δt, the inclusive work contribution ΔW_s due to the change in potential shape is independent of the informational state. This is reasonable for our erasure protocol since the asymmetric contribution to the change in potential—the tilt—is slight. While it clearly breaks the symmetry of the double-well potential by changing the well heights, it has less of an effect on the well shapes and even less in making those shapes distinct.

Then, we can assume that the sum of ΔW_s for any trajectory is the same as that for a particle that stays in either informational state the entire time. But since the protocol is cyclic and very slow compared to the informational states' relaxation rates, a particle that stays in one informational state the entire time must receive approximately zero inclusive work W. Given that the sum over all ΔW_s must be equal to W for such a trajectory, the former must also be negligible.

Altogether, the total exclusive work is approximately given by the sum over all transitions between informational states of the difference in potential at the informational-state minima:

$$W_0(0, \tau) = -\sum_{\text{trans}} \Delta m^t. \tag{H11}$$

To reiterate, since $\Delta h_{\text{S,L}} = 0$, this is also the total inclusive work $W(0, \tau)$ for a trajectory over the entire protocol.

APPENDIX I: SUBSTAGE WORK DISTRIBUTIONS COMMENTARY

Here, we briefly interpret several features of the substage work distributions observed in Fig. 1 (outer plots).

The distributions for barrier dropping and tilting are narrow, symmetric peaks; see Fig. 1 (outer plots). Barrier raising also has a rather narrow peak, composed primarily of trajectories always in the R state, but it also exhibits a bulge toward positive work; see Fig. 1 (top right). Note that the L state is created midway through barrier raising, allowing for trajectories that spend some time in either informational state, but disallowing trajectories that spend all time in the L state. The former induce the positive work bulge toward less negative works, which, while notable, will not be further explored here.

The substage work distributions for untilting present the most striking picture; see Fig. 1 (bottom right). Always-R trajectories induce a large positive work peak (red), always-L trajectories induce a large negative work peak (orange), and all other trajectories induce a ramp between them (blue).

These features can be directly interpreted by following the locations of the potential minima over time and noting how the shifting potential adds or removes energy from a particle. During barrier dropping, to take one example, the protocol raises both minima by over $7k_\text{B}T$, resulting in a narrow, peaked work distribution with a mean near $7k_\text{B}T$.

Most interesting is the untilt substage. Since most particles start and then stay in the R state for this substage, a large positive work is probable, due to the rising R-state well. However, it is also possible for the system to start in and then get stuck in the L-state well, resulting in a large negative work. The final possibility is transitioning between states during untilting, resulting in an intermediate range of less likely work values. For a trajectory that does transition between states during untilting, it is more likely to spend more time in the R state, since it is energetically favored, resulting in the rising probability with increasing work in their work distribution—giving rise to the log-linear ramp in the work distribution.

Note that there are small peaks on each end of this third class' distributions that require a more nuanced explanation. When a particle crosses a barrier—due to random thermal

excitation—the surplus energy may quickly send the particle back to the previous well before it can be dissipated. Such particles then spend almost all of the substage in this first well, generating a work value accordingly. The statistics of the ramp proper are due to particles that have time to locally equilibrate before crossing any barriers.

Follow-on work develops the theory underlying this detailed mechanistic analysis and analyzes similar behavior in all metastable-quasistatic processes.

APPENDIX J: FLUX QUBIT DEVICE, CALIBRATION, AND MEASUREMENT

The benefits of flux qubit devices are severalfold. First, their physics provides a genuine two-degree-of-freedom dynamics, while other comparable experiments on Maxwellian demons and bit erasure are very high dimensional, only indirectly providing an effectively few-degree-of-freedom dynamics [20,22,48]. Second, they operate at very high frequency and so one readily captures the substantial amounts of data required to accurately estimate rare-fluctuation statistics. Third, they leverage recent advances in superconducting-device manufacturing technology led by efforts in quantum computing. Fourth, being constructed via modern integrated circuit technology, they form the basis of a technology that even today is ready to scale to large, multicomponent circuit devices for more sophisticated thermodynamic computing. And finally, in the near future flux qubits will facilitate experiments that probe the thermodynamics of the transition from classical to quantum information processing.

At the microscopic level, a fraction of the electrons in a superconducting metal form bosonic Cooper pairs—a quantum-coherent condensate. For designing superconducting electronic circuits, though, one can forgo the microscopic description and work with higher-level phenomena, such as flux quantization and the Josephson relations for weak links. Importantly, the circuit-level degrees of freedom are not coarse-grained quantities, but display a full range of quantum behavior, including quantized excitations, coherent superpositions, and entangled states in such circuits. For our purposes here, however, we run the device so that it exhibits only classical stochastic dynamics, reserving quantum information thermodynamic explorations for the future.

This Appendix lays out the basic physics of the flux qubit device and details of the experimental implementation. A fuller discussion of the platform and its calibration is found in Ref. [49].

1. Flux qubit physics

Our experimental information processor is a special type of superconducting quantum interference device (SQUID) with two degrees of freedom—a gradiometric flux qubit or the variable-I_c rf SQUID introduced by Ref. [50]. Notably, the energies associated with the motion perpendicular to and along the escape direction differ substantially by about a factor of 12. Practically, this asymmetry reduces the two-dimensional potential to one dimension. The net result is a device with an effective double-well potential with barriers as low as $\Delta U \sim k_B T$ that operates at frequencies in the GHz range.

The potential shape is controlled by fluxes that are readily controlled by currents. SQUID device parameters, used to determine the potential shape and energy scales, were all independently determined.

The variable-I_c rf SQUID replaces the single Josephson junction in a standard rf SQUID with a symmetric dc SQUID with small inductance $\beta_{dc} = 2\pi \ell I_{c0}/\Phi_0 \ll 1$, where 2ℓ is the loop inductance, $I_{c0} = i_{c1} + I_{c2}$ is the sum of critical currents of the two junctions, and Φ_0 is the flux quantum $h/2e$. This architecture gives a device whose parameters can be accurately measured and that can be selected to exhibit a range of phenomena including thermal activation, macroscopic quantum tunneling, incoherent relaxation, photon-induced transitions, and macroscopic quantum coherence. It also allows us to perform, as we demonstrate, nanoscale thermodynamic computing.

Its macroscopic dynamical variables are the magnetic flux Φ through the rf SQUID loop and Φ_{dc} through the dc SQUID loop. Based on the resistively-capacitively-shunted junction model of Josephson junctions, in the classical limit the variable-I_c rf SQUID's deterministic equations of motion are [51]

$$2C\ddot{\Phi} + \frac{\dot{\Phi}}{R/2} = -\frac{\partial U(\Phi, \Phi_{dc})}{\partial \Phi} \text{ and}$$

$$\frac{C}{2}\ddot{\Phi}_{dc} + \frac{\dot{\Phi}}{2R} = -\frac{\partial U(\Phi, \Phi_{dc})}{\partial \Phi_{dc}}. \quad \text{(J1)}$$

In units of $\Phi_0/2\pi$, the two-dimensional potential for the variable-I_c rf SQUID is $U(\phi, \phi_{dc}) = U_0 f(\phi, \phi_{dc})$ with

$$f(\phi, \phi_{dc}) = \frac{1}{2}(\phi - \phi_x)^2 + \frac{\gamma}{2}(\phi_{dc} - \phi_{xdc})^2$$
$$- \beta_0 \cos\frac{\phi_{dc}}{2}\cos\phi + \delta\beta \sin\frac{\phi_{dc}}{2}\sin\phi, \quad \text{(J2)}$$

where $U_0 = \Phi_0^2/(4\pi^2 L)$. Here, $\gamma = L/(2\ell)$ is the ratio of rf and dc SQUID inductances; ϕ_x (ϕ_{xdc}) is the external flux applied to the rf (dc) SQUID loop; ϕ (ϕ_{dc}) is the flux enclosed in the rf (dc) SQUID loop; $\beta_0 = 2\pi L I_{c0}/\Phi_0$; and $\delta\beta = 2\pi L(I_{c2} - I_{c1})/\Phi_0$.

For large-amplitude tuning of the external controls, the system response to ϕ_x (ϕ_{xdc}) is 2π (4π) periodic. We make use of the global features to accurately determine the coefficients of the potential.

In the experiment, cross-coupling between the barrier and tilt controls was canceled by an affine transformation $(\phi_x, \phi_{xdc}) \rightarrow (\phi_x + \alpha\phi_{xdc}, \phi_{xdc})$, with the coefficient α chosen such that the equilibrium population of the left and right wells was unaffected to first order by the barrier control ϕ_{xdc}.

Operating the magnetometer generates wide-band local electromagnetic interference that can affect the dynamics of the flux qubit. A careful study of the back-action indicates that low-amplitude operation of the magnetometer can induce transitions in a manner that corresponds to a shift in the effective tilt and flux controls. Importantly, the effective temperature under magnetometer operations was not elevated from 500 mK.

The dynamical variable ϕ describes the in-phase motion of the two junctions that results in a current circulating in the rf SQUID loop. The dynamical variable ϕ_{dc} describes the

out-of-phase motion, resulting in a current circulating in the dc SQUID loop. The shape of the effective potential is completely determined by the dimensionless function $f(\phi, \phi_{dc})$, and the energy scale of the potential is determined by U_0. With suitable device parameters and applied fluxes (ϕ_x and ϕ_{xdc}), one obtains a smooth family of double-well potentials. The barrier height ΔU separating the two wells is readily adjusted by varying ϕ_{xdc}. The effective potential is plotted in Fig. 3(b) with the following parameters: $\beta_0 = 6.2$, $\gamma = 12$, and $\delta_b = 0.2$.

2. Experimental implementation

The junctions were 1×1 μm^2 Nb/Al$_2$O$_3$/Nb tunnel junctions of very low subgap leakage, typically having a quality factor of $V_m \approx 70$ mV at 4.2 K.

We followed a standard procedure (see, e.g., Ref. [51]) for calibrating the flux qubit parameters. An outline of the steps is given below. A complete description of the measurements is presented in Ref. [49].

First, by executing wide-range sweeps of the coil currents I_{tilt} and $I_{barrier}$, parameter values corresponding to single-valued and bistable potential landscapes are recorded. A linear transformation from I_{tilt} and $I_{barrier}$ to (ϕ_x, ϕ_{xdc}) is established by matching the experimental periodicity with the theoretical one (2π, 4π). Linear cross-talk from $I_{barrier}$ to I_{tilt} is calibrated by orthogonalizing the global response. Cross-talk from I_{tilt} to $I_{barrier}$ can be assumed to be small due to the symmetry of the on-chip flux lines and is taken to be zero.

The parameter values $\beta_0 = 6.2$, $\gamma = 12$, and $\delta_b = 0.2$ are determined by equating the observed extent of hysteresis at $\phi_{xdc} = 0$ and the differential flux response $d\langle\phi\rangle/d\phi_x$ at $\phi_{xdc} = 2\pi$ to theoretical predictions. The prefactor $U_0 = 56.3$ K is determined by equating the observed escape energy for interwell transitions at high temperatures with $k_B T$. The plasma frequency $\omega_p = 1/\sqrt{LC} = 2\pi \times 13.7$ GHz is determined from the observed low-temperature crossover temperature $T_{cr} = 103$ mK to macroscopic quantum tunneling (MQT) dominated dynamics. We obtain an upper bound $Q = \omega_p RC < 130$ from the coupling to the passive shunt resistor of the magnetometer. Parameter calibration measurements are performed in such a way that the effect of magnetometer back-action is nulled through pulsing of the readout or otherwise minimized. The effective temperature under continuous magnetometer operation was determined by repeating the measurement for escape energy for interwell transitions and comparing the result to that obtained under pulsed magnetometer operation.

[1] C. G. Knott, *Life and Scientific Work of Peter Gurthrie Tait* (Cambridge University Press, Cambridge, UK, 1911), Letter from Maxwell to Tait, 11 December 1867, quoted herein on pp. 213 and 214.

[2] R. Landauer, Irreversibility and heat generation in the computing process, IBM J. Res. Dev. **5**, 183 (1961).

[3] C. H. Bennett, Thermodynamics of computation—A review, Int. J. Theor. Phys. **21**, 905 (1982).

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

[5] A. B. Boyd, D. Mandal, and J. P. Crutchfield, Leveraging environmental correlations: The thermodynamics of requisite variety, J. Stat. Phys. **167**, 1555 (2016).

[6] T. Conte *et al.*, Thermodynamic computing, arXiv:1911.01968.

[7] D. Mandal and C. Jarzynski, Work and information processing in a solvable model of Maxwell's demon, Proc. Natl. Acad. Sci. (USA) **109**, 11641 (2012).

[8] 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).

[9] G. N. Bochkov and Y. E. Kuzovlev, Nonlinear fluctuation-dissipation relations and stochastic models in nonequilibrium thermodynamics: I. Generalized fluctuation-dissipation theorem, Phys. A **106**, 443 (1981).

[10] D. J. Evans and D. J. Searles, Equilibrium microstates which generate second law violating steady states, Phys. Rev. E **50**, 1645 (1994).

[11] C. Jarzynski, Nonequilibrium Equality for Free Energy Differences, Phys. Rev. Lett. **78**, 2690 (1997).

[12] G. E. Crooks, Nonequilibrium measurements of free energy differences for microscopically reversible Markovian systems, J. Stat. Phys. **90**, 1481 (1998).

[13] G. E. Crooks, Entropy production fluctuation theorem and the nonequilibrium work relation for free energy differences, Phys. Rev. E **60**, 2721 (1999).

[14] U. Seifert, Stochastic thermodynamics, fluctuation theorems and molecular machines, Rep. Prog. Phys. **75**, 126001 (2012).

[15] *Nonequilibrium Statistical Physics of Small Systems: Fluctuation Relations and Beyond*, edited by R. Klages, W. Just, and C. Jarzynski (Wiley, New York, 2013).

[16] P. Maragakis, M. Spichty, and M. Karplus, A differential fluctuation theorem, J. Phys. Chem. B **112**, 6168 (2008).

[17] I. Junier, A. Mossa, M. Manosas, and F. Ritort, Recovery of Free Energy Branches in Single Molecule Experiments, Phys. Rev. Lett. **102**, 070602 (2009).

[18] A. Alemany, A. Mossa, I. Junier, and F. Ritort, Experimental free-energy measurements of kinetic molecular states using fluctuation theorems, Nat. Phys. **8**, 688 (2012).

[19] B. Lambson, D. Carlton, and J. Bokor, Exploring the Thermodynamic Limits of Computation in Integrated Systems: Magnetic Memory, Nanomagnetic Logic, and the Landauer Limit, Phys. Rev. Lett. **107**, 010604 (2011).

[20] A. Berut, A. Petrosyan, and S. Ciliberto, Detailed Jarzynski equality applied to a logically irreversible procedure, Europhys. Lett. **103**, 60002 (2013).

[21] M. Madami, M. d'YAquino, G. Gubbiotti, S. Tacchi, C. Serpico, and G. Carlotti, Micromagnetic study of minimum-energy dissipation during Landauer erasure of either isolated or coupled nanomagnetic switches, Phys. Rev. B **90**, 104405 (2014).

[22] Y. Jun, M. Gavrilov, and J. Bechhoefer, High-Precision Test of Landauer's principle, Phys. Rev. Lett. **113**, 190601 (2014).

[23] A. Bérut, A. Petrosyan, and S. Ciliberto, Information and thermodynamics: Experimental verification of Landauer's erasure principle, J. Stat. Mech.: Theor. Expt. (2015) P06015.

[24] J. Hong, B. Lambson, S. Dhuey, and J. Bokor, Experimental test of Landauer's principle in single-bit operations on nanomagnetic memory bits, Sci. Adv. **2**, e1501492 (2016).

[25] J. V. Koski, A. Kutvonen, I. M. Khaymovich, T. Ala-Nissila, and J. P. Pekola, On-Chip Maxwell's Demon as an Information-Powered Refrigerator, Phys. Rev. Lett. **115**, 260602 (2015).

[26] É. Roldán, I. A. Martinez, J. M. R. Parrondo, and D. Petrov, Universal features in the energetics of symmetry breaking, Nat. Phys. **10**, 457 (2014).

[27] M. Gavrilov, R. Chétrite, and J. Bechhoefer, Direct measurement of weakly nonequilibrium system entropy is consistent with Gibbs-Shannon form, Proc. Natl. Acad. Sci. (USA) **114**, 11097 (2017).

[28] T. M. Cover and J. A. Thomas, *Elements of Information Theory*, 2nd ed. (Wiley-Interscience, New York, 2006).

[29] C. Jarzynski, Rare events and the convergence of exponentially averaged work values, Phys. Rev. E **73**, 046105 (2006).

[30] A. B. Boyd, A. Patra, C. Jarzynski, and J. P. Crutchfield, Shortcuts to thermodynamic computing: The cost of fast and faithful erasure, arXiv:1812.11241.

[31] A. B. Boyd, D. Mandal, and J. P. Crutchfield, Identifying functional thermodynamics in autonomous Maxwellian ratchets, New J. Phys. **18**, 023049 (2016).

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

[33] 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).

[34] 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).

[35] A. B. Boyd, D. Mandal, and J. P. Crutchfield, Thermodynamics of Modularity: Structural Costs Beyond the Landauer Bound, Phys. Rev. X **8**, 031036 (2018).

[36] P. M. Riechers and J. P. Crutchfield, Fluctuations when driving between nonequilibrium steady states, J. Stat. Phys. **168**, 873 (2017).

[37] C. Aghamohammdi and J. P. Crutchfield, Thermodynamics of random number generation, Phys. Rev. E **95**, 062139 (2017).

[38] P. R. Zulkowski and M. R. DeWeese, Optimal control of over-damped systems, Phys. Rev. E **92**, 032117 (2015).

[39] T. R. Gingrich, G. M. Rotskoff, G. E. Crooks, and P. L. Geissler, Near optimal protocols in complex nonequilibrium transformations, Proc. Natl. Acad. Sci. (USA) **113**, 10263 (2016).

[40] A. Patra and C. Jarzynski, Classical and quantum shortcuts to adiabaticity in a tilted piston, J. Phys. Chem. B **121**, 3403 (2017).

[41] S. Asban and S. Rahav, Nonequilibrium free-energy estimation conditioned on measurement outcomes, Phys. Rev. E **96**, 022155 (2017).

[42] R. Kawai, J. M. R. Parrondo, and C. Van den Broeck, Dissipation: The Phase-Space Perspective, Phys. Rev. Lett. **98**, 080602 (2007).

[43] S. Vaikuntanathan and C. Jarzynski, Dissipation and lag in irreversible processes, Europhys. Lett. **87**, 60005 (2009).

[44] T. M. Hoang, R. Pan, J. Ahn, J. Bang, H. T. Quan, and T. Li, Experimental Test of the Differential Fluctuation Theorem and a Generalized Jarzynski Equality for Arbitrary Initial States, Phys. Rev. Lett. **120**, 080602 (2018).

[45] A. Gomez-Marin, J. M. R. Parrondo, and C. Van den Broeck, Lower bounds on dissipation upon coarse graining, Phys. Rev. E **78**, 011107 (2008).

[46] G. E. Crooks, Excursions in statistical dynamics, Ph.D. thesis, University of California, Berkeley, 1999, https://threeplusone.com/pubs/gecthesis/.

[47] C. Jarzynski, Comparison of far-from-equilibrium work relations, C. R. Phys. **8**, 495 (2007).

[48] A. Berut, A. Arakelyan, A. Petrosyan, S. Ciliberto, R. Dillenschneider, and E. Lutz, Experimental verification of Landauer's principle linking information and thermodynamics, Nature **483**, 187 (2012).

[49] O.-P. Saira, M. H. Matheny, R. Katti, W. Fon, G. Wimsatt, S. Han, J. P. Crutchfield, and M. L. Roukes, Nonequilibrium thermodynamics of erasure with superconducting flux logic, Phys. Rev. Res. **2**, 013249 (2020).

[50] S. Han, J. Lapointe, and J. E. Lukens, Thermal Activation in a Two-Dimensional Potential, Phys. Rev. Lett. **63**, 1712 (1989).

[51] S. Han, J. Lapointe, and J. E. Lukens, Effect of a two-dimensional potential on the rate of thermally induced escape over the potential barrier, Phys. Rev. B **46**, 6338 (1992).