

# Harnessing Fluctuations in Thermodynamic Computing via Time-Reversal Symmetries

One Sentence Summary:

**Distinct distributions of thermodynamic work identify signatures of successful and failed information processing in a microscale flux qubit**

Gregory Wimsatt,<sup>1,\*</sup> Olli-Pentti Saira,<sup>2,†</sup> Alexander B. Boyd,<sup>1,‡</sup> Matthew H. Matheny,<sup>2,§</sup> Siyuan Han,<sup>3,¶</sup> Michael L. Roukes,<sup>2,\*\*</sup> and James P. Crutchfield<sup>1,2,††</sup>

<sup>1</sup>*Complexity Sciences Center and Physics Department,  
University of California at Davis, One Shields Avenue, Davis, CA 95616*

<sup>2</sup>*Condensed Matter Physics and Kavli Nanoscience Institute,  
California Institute of Technology, Pasadena, CA 91125*

<sup>3</sup>*Department of Physics and Astronomy, University of Kansas, Lawrence, KS 66045*

(Dated: July 1, 2019)

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 both the frequency of various mechanisms of 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.

Keywords: Jarzynski integral fluctuation theorem, Crooks detailed fluctuation theorem, Landauer’s Principle, thermodynamics of computation, information thermodynamics

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. All computing is thermodynamic. Controlling the production of fluctuations and removing heat pose key technological challenges to further progress. One comparable setting that gives some optimism, though, is the overtly functional behavior exhibited by biological cells—presumably functional information processing by

small numbers of molecules subject to substantial thermal fluctuations. Computing technologies are very far away from this kind of robust information processing.

Only recently have tools appeared that precisely describe what trade-offs 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 [5]. This has opened up the new paradigm of *thermodynamic computing* 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* [6, 7].

Reliable thermodynamic computing requires detecting

---

\* [gwwimsatt@ucdavis.edu](mailto:gwwimsatt@ucdavis.edu)

† [osaira@caltech.edu](mailto:osaira@caltech.edu)

‡ [aboyd@ucdavis.edu](mailto:aboyd@ucdavis.edu)

§ [matheny@caltech.edu](mailto:matheny@caltech.edu)

¶ [han@ku.edu](mailto:han@ku.edu)

\*\* [roukes@caltech.edu](mailto:roukes@caltech.edu)

†† [chaos@ucdavis.edu](mailto:chaos@ucdavis.edu); Corresponding author

and controlling fluctuations in informational and energetic resources and in engine functioning. For this, one appeals to fluctuation theorems that capture exact time-reversal symmetries and predict entropy production leading to irreversible dissipation [8–14]. We are now on the doorstep of the very far-from-equilibrium thermodynamics needed to understand the physics of computing. And, in turn, this has started to reveal the physical principles of how nature processes information in the service of biological functioning and survival.

Proof-of-concept experimental tests have been carried out in several substrates: probing biomolecule free energies [15–17], work expended during elementary computing (bit erasure) [18–23], and Maxwellian demons [24]. That said, the suite of contemporary principles (Supplementary Materials (SM) I) 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. These structures track the mesoscale evolution of a system’s *informational states* and reveal classes of functional and nonfunctional microscopic trajectories. We show that the informational-state evolutions are identified by appropriate conditioning and that they obey a suite of trajectory-class fluctuation theorems, which give accurate bounds on work, entropy production, and dissipation. The result is a new tool that employs mesoscopic measurements to diagnose nanoscale thermodynamic computing. For simplicity and to make direct contact with previous efforts, we demonstrate the tools on Landauer erasure of a bit of information in a superconducting flux qubit.

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*  $\{(\beta(t), \delta(t)) : 0 \leq t \leq \tau\}$ .  $\beta(t)$  and  $\delta(t)$  change one at a time piecewise-linearly through four protocol substages: (1) *drop barrier*, (2) *tilt*, (3) *raise barrier*, and (4) *untilt*. (See SM VI.) 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 [25].

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.

The erasure protocol is designed so that the particle ends in the *R* state with high probability, regardless of its initial state. Conducting our simulation  $3.5 \times 10^6$  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. Note that we chose the protocol to give partially inaccurate erasure in order to illustrate our main results on diagnosing success and failure.

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 towards local equilibrium. We therefore define the informational states at time  $0 \leq t \leq \tau$  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.

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  $0 \leq t < t' \leq \tau$ ; see SM VI. Let  $W_s = W(t_{s-1}, t_s)$  denote the work generated during substage  $s$  and  $C_s$  the substage trajectory class. Figure 1 (Outer plot sequence) shows the corresponding substage work distributions  $\text{Pr}(W_s, C_s)$  obtained from our simulations. (See SM VII.)

The drop-barrier and tilt substage work distributions are rather simple, being narrow and unimodal. The raise-barrier distributions have some asymmetry, but 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 *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.

Looking at the total work  $W_{\text{total}} = W(0, \tau)$  generated for each trajectory over the course of the entire erasure

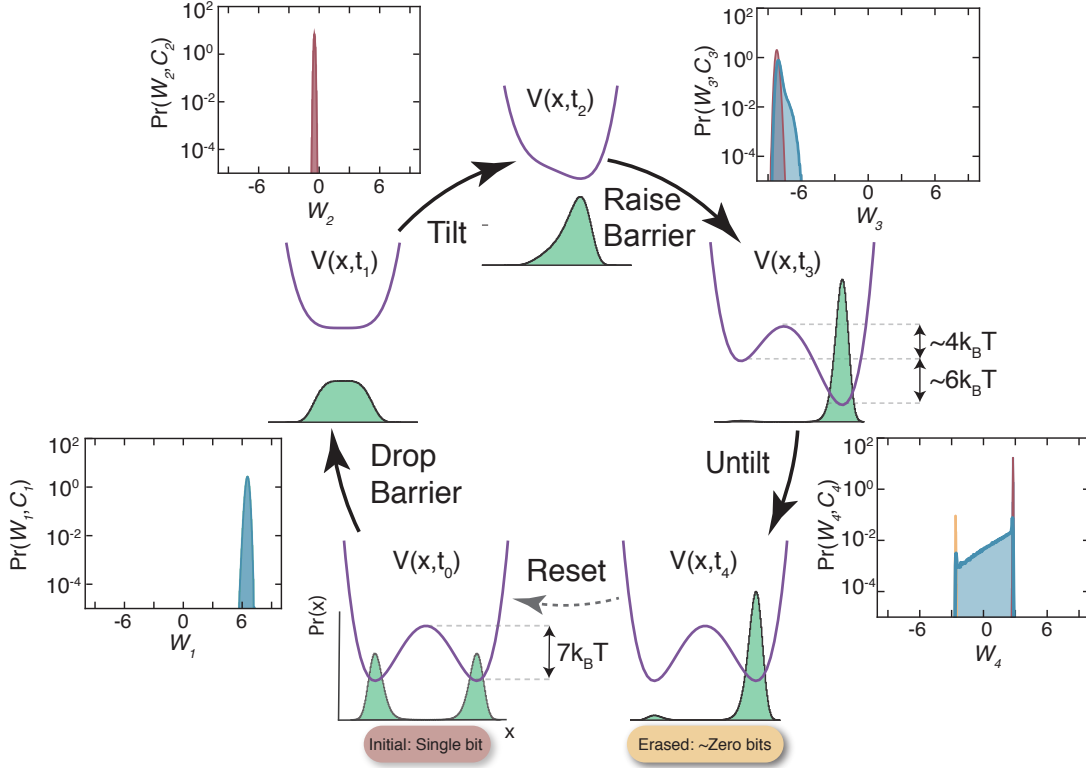


FIG. 1. Inner plot sequence: Erasure protocol (Table S1) evolution of position distribution  $\text{Pr}(x)$ . Potential  $V(x, t_s)$  at substage boundary times  $t_s$ ,  $s = 0, 1, 2, 3, 4$ . Starting at  $t = t_0$ , the potential evolves clockwise, ending at  $t = t_4$  in the same configuration as it starts:  $V(x, t_0) = V(x, t_4)$ . However, the final position distribution  $\text{Pr}(x)$  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  $\text{Pr}(W_s, C_s)$  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 are of trajectories always in the  $R$  state, orange are of trajectories always in the  $L$  state, and blue are of the rest, spending some time in each state.

protocol, we observe the strikingly complex and structured distribution  $\text{Pr}(W_{\text{total}})$  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; 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. [12].

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 *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 (Three front plots) shows 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). Though, now the thermodynamic contributions to the total from the functionally-distinct component trajectories are made apparent.

Exploring the mesoscale dynamics of erasure revealed signatures of a “thermodynamics” for each trajectory that is closely 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 [8–14]. (See SM III and IV.)

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 Jarzynski’s equality [10] and Crooks’ detailed fluctuation theorem [12], as the trajectory class varies. This lower bounds the average work  $\langle W \rangle_C$  over any measurable subset  $C$  of the

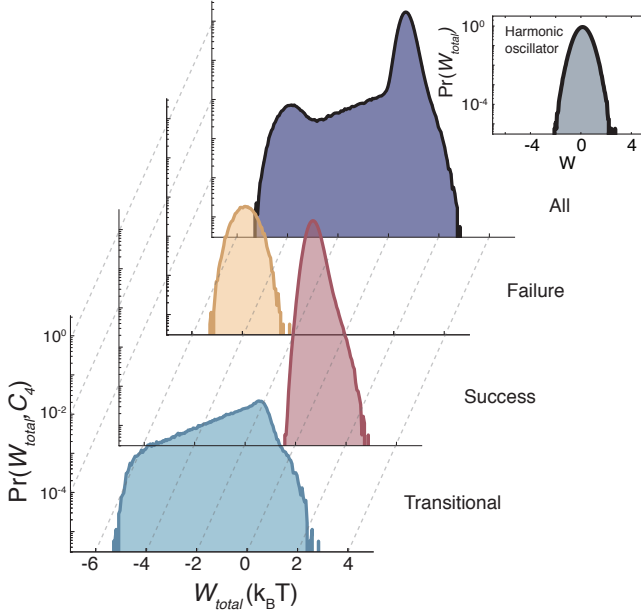


FIG. 2. (Rear, purple) Total work distribution of all trajectories  $\text{Pr}(W_{\text{total}})$  during erasure: A histogram generated from  $3.5 \times 10^6$  trials for  $W_{\text{total}} \in [-6, 4]$  over 201 bins. (Inset, gray) Typical unimodal work distribution illustrated for spatially-translated thermally-driven simple harmonic oscillator. (Three front plots) Work distributions  $\text{Pr}(W_{\text{total}}, C_4)$  for the trajectory classes  $C_4$  determined by the untilt trajectory partition: 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.

ensemble of system microstate trajectories  $\vec{Z}$ , where  $W$  is the total work for a trajectory:

$$\begin{aligned} \langle W \rangle_C &\geq \Delta F + k_B T \ln \frac{\mathcal{P}(C)}{\mathcal{R}(C^R)} \\ &= \langle W \rangle_C^{\min}, \end{aligned} \quad (1)$$

with  $\Delta F$  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.  $k_B$  is Boltzmann's constant.

The TCFTs lead to several consequences. First, well-formulated trajectory classes allow accurate estimates of the works for their trajectories, even with limited knowledge of system response under the protocol and its time reverse. Second, they strictly and more strongly bound the average work over all trajectories compared to the equilibrium free energy change  $\Delta F$ . Third, they provide a new expression for obtaining equilibrium free-energy

changes:

$$\Delta F = -k_B T \ln \left( \frac{\mathcal{P}(C)}{\mathcal{R}(C^R)} \langle e^{-W/k_B T} \rangle_C \right). \quad (2)$$

Remarkably, this only requires statistics for a 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  $\Delta F = -k_B T \ln \langle e^{-W/k_B T} \rangle_{\vec{Z}}$  for estimating free energies [26].

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)). SM VIII A 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 [27], 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 TCFTs can be used to diagnose and interpret success and failure in microscopic information processing using only mesoscopic measurements of work.

Interplay between the geometric, linear magnetic inductances and the nonlinear Josephson inductances gives rise to a potential landscape that can be controlled with external bias fluxes. It is natural to call the  $\phi_x$  and  $\phi_{\text{xdc}}$  fluxes, threading the differential mode and the small SQUID loop, respectively, the *tilt* and *barrier controls*. (See (Fig. 3(A) caption.) SM VIII 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.

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. The amplitudes of the control waveforms in reduced units are small; see Fig. 3(C). Hence, 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  $\mu\text{s}$ , after which the two logical states are discriminated



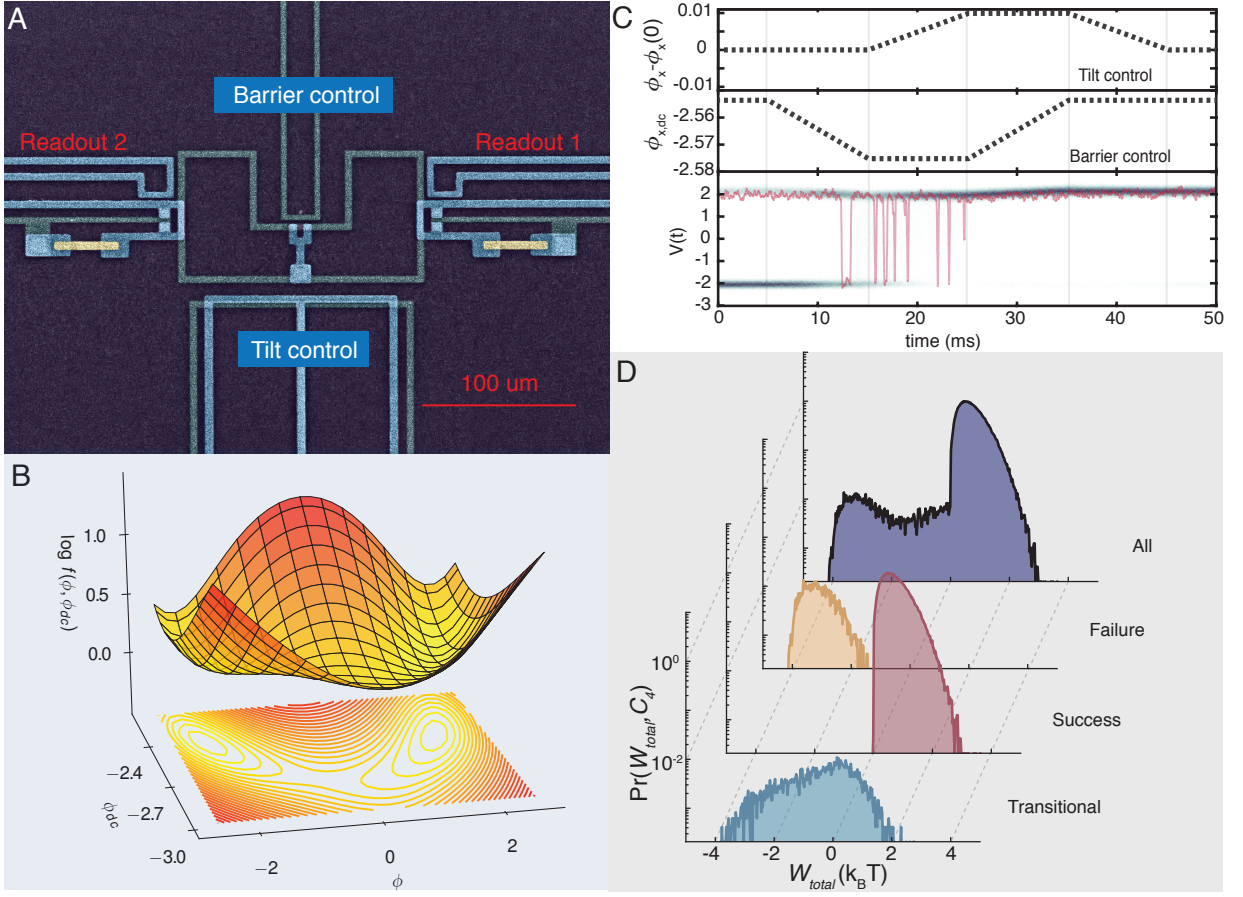


FIG. 3. Superconducting implementation of metastable memory and bit erasure driven by thermal fluctuations: (A) Optical micrograph of a gradiometric flux qubit with control lines and local magnetometers for state readout. The flux  $\phi_x$ , threading the large U-shaped differential-mode loop, controls the potential’s tilt and flux  $\phi_{x,dc}$ , 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. (S18) and (S19). (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. (D) Work distributions  $\text{Pr}(W_{\text{total}}|C_4)$  over trajectories conditioning on the Success, Fail, and Transitional classes. Experimental distributions obtained from  $10^5$  protocol repetitions.

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 wideband 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  $\delta_i$ —sign convention: +1 (−1) for a 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 intra-well equilibration. (See SM VI.)

The total work distribution estimated from the flux qubit experiments is shown as the rear-most 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 so, given the rather substantial differences between the simulated system (idealized double-well potential and thermal noise, exactly one-dimension system,

...) and the experimental system (complex potential in two dimensions, nonideal fluctuations, ...). *A priori* it is not clear that the theoretical predictions of the informational classes should apply so directly and immediately to the real-world qubit. In point of fact, these differences serve to emphasize the descriptive power of the mesoscopic work fluctuation theorems: despite substantial differences in system detail they successfully diagnose the information-processing classes of microscopic trajectories. This robustness will be especially helpful in monitoring thermodynamic computing in biological systems, where, in many cases, information-bearing degrees-of-freedom cannot be precisely modeled.

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.

Using them we gave a detailed mechanistic analysis of the thermodynamics 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. The corresponding bounds on thermodynamic resources obtained via the TCFTs also improve on current estimation methods. The net result is that trajectory-class analysis can be readily applied to debug thermodynamic computing by engineered or biological systems.

- 
- [1] C. G. Knott. *Life and Scientific Work of Peter Gurthrie Tait*. Cambridge University press, Cambridge, United Kingdom, 1911. Letter from Maxwell to Tait, 11 December 1867, quoted herein pp. 213–214.
  - [2] R. Landauer. Irreversibility and heat generation in the computing process. *IBM J. Res. Develop.*, 5(3):183–191, 1961.
  - [3] C. H. Bennett. Thermodynamics of computation—A review. *Intl. J. Theo. Phys.*, 21:905, 1982.
  - [4] J. M. R. Parrondo, J. M. Horowitz, and T. Sagawa. Thermodynamics of information. *Nature Physics*, 11(2):131–139, 2015.
  - [5] A. B. Boyd, D. Mandal, and J. P. Crutchfield. Leveraging environmental correlations: The thermodynamics of requisite variety. *J. Stat. Phys.*, 167(6):1555–1585, 2016.
  - [6] D. Mandal and C. Jarzynski. Work and information processing in a solvable model of Maxwell’s demon. *Proc. Natl. Acad. Sci. USA*, 109(29):11641–11645, 2012.
  - [7] 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.
  - [8] G. N. Bochkov and Y. E. Kuzovlev. Nonlinear fluctuation-dissipation relations and stochastic models in nonequilibrium thermodynamics: I. generalized fluctuation-dissipation theorem. *Physica A: Stat. Mech. App.*, 106(3):443–479, 1981.
  - [9] D. J. Evans and D. J. Searles. Equilibrium microstates which generate second law violating steady states. *Phys. Rev. E*, 50(2):1645–1648, 1994.
  - [10] C. Jarzynski. Nonequilibrium equality for free energy differences. *Phys. Rev. Lett.*, 78(14):2690–2693, 1997.
  - [11] G. E. Crooks. Nonequilibrium measurements of free energy differences for microscopically reversible Markovian systems. *J. Stat. Phys.*, 90(5/6):1481–1487, 1998.
  - [12] G. E. Crooks. Entropy production fluctuation theorem and the nonequilibrium work relation for free energy differences. *Phys. Rev. E*, 60:2721, 1999.
  - [13] U. Seifert. Stochastic thermodynamics, fluctuation theorems and molecular machines. *Rep. Prog. Phys.*, 75:126001, 2012.
  - [14] R. Klages, W. Just, and C. Jarzynski, editors. *Nonequilibrium Statistical Physics of Small Systems: Fluctuation Relations and Beyond*. Wiley, New York, 2013.
  - [15] P. Maragakis, M. Spichty, and M. Karplus. A differential fluctuation theorem. *J. Phys. Chem. B*, 112(19):6168–6174, 2008.
  - [16] I. Junier, A. Mossa, M. Manosas, and F. Ritort. Recovery of free energy branches in single molecule experiments. *Phys. Rev. Lett.*, 102(7):070602, 2009.
  - [17] A. Alemany, A. Mossa, I. Junier, and F. Ritort. Experimental free-energy measurements of kinetic molecular states using fluctuation theorems. *Nature Physics*, 8:688–694, 2012.
  - [18] 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.

- [19] A. Berut, A. Petrosyan, and S. Ciliberto. Detailed Jarzynski equality applied to a logically irreversible procedure. *Euro. Phys. Lett.*, 103:60002, 2013.
  - [20] 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.
  - [21] Y. Jun, M. Gavrilov, and J. Bechhoefer. High-precision test of Landauer’s principle. *Phys. Rev. Lett.*, 113:190601, 2014.
  - [22] A. Berut, A. Petrosyan, and S. Ciliberto. Information and thermodynamics: Experimental verification of Landauer’s erasure principle. *J. Stat Mech: Theory and Experiment*, 2015(6):P06015, 2015.
  - [23] 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.
  - [24] 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.
  - [25] T. M. Cover and J. A. Thomas. *Elements of Information Theory*. Wiley-Interscience, New York, second edition, 2006.
  - [26] C. Jarzynski. Rare events and the convergence of exponentially averaged work values. *Phys. Rev. E*, 73(4):046105, 2006.
  - [27] 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.
  - [28] A. B. Boyd, D. Mandal, and J. P. Crutchfield. Identifying functional thermodynamics in autonomous Maxwellian ratchets. *New J. Physics*, 18:023049, 2016.
  - [29] S. Still, D. A. Sivak, A. J. Bell, and G. E. Crooks. Thermodynamics of prediction. *Phys. Rev. Lett.*, 109:120604, 2012.
  - [30] 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.
  - [31] A. B. Boyd, D. Mandal, and J. P. Crutchfield. Correlation-powered information engines and the thermodynamics of self-correction. *Phys. Rev. E*, 95(1):012152, 2017.
  - [32] A. B. Boyd, D. Mandal, and J. P. Crutchfield. Thermodynamics of modularity: Structural costs beyond the Landauer bound. *Phys. Rev. X*, 8(3):031036, 2018.
  - [33] P. M. Riechers and J. P. Crutchfield. Fluctuations when driving between nonequilibrium steady states. *J. Stat. Phys.*, 168(4):873–918, 2017.
  - [34] C. Aghamohammadi and J. P. Crutchfield. Thermodynamics of random number generation. *Phys. Rev. E*, 95(6):062139, 2017.
  - [35] P. R. Zulkowski and M. R. DeWeese. Optimal control of overdamped systems. *Phys. Rev. E*, 92(5):032117, 2015.
  - [36] T. R. Gingrich, G. M. Rotskoff, G. E. Crooks, and P. L. Geissler. Near optimal protocols in complex nonequilibrium transformations. *Proc. Natl. Acad. Sci. U.S.A.*, 113(37):10263–10268, 2016.
  - [37] A. Patra and C. Jarzynski. Classical and quantum shortcuts to adiabaticity in a tilted piston. *J. Phys. Chem. B*, 121:3403–3411, 2017.
  - [38] A. Gomez-Marin, J. M. R. Parrondo, and C. Van den Broeck. Lower bounds on dissipation upon coarse graining. *Phys. Rev. E*, 78(1):011107, 2008.
  - [39] G.E. Crooks. *Excursions in statistical dynamics*. PhD thesis, University of California, Berkeley, 1999.
  - [40] C. Jarzynski. Comparison of far-from-equilibrium work relations. *Comptes Rendus Physique*, 8(5-6):495–495, 2007.
  - [41] A. Berut, A. Arakelyan, A. Petrosyan, S. Ciliberto, R. Dillenschneider, and E. Lutz. Experimental verification of Landauer’s principle linking information and thermodynamics. *Nature*, 483:187–190, 2012.
  - [42] 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.
- Acknowledgments:** We thank C. Jarzynski, D. Mandal, and P. Riechers for helpful discussions. As an External Faculty member, JPC thanks the Santa Fe Institute and the Telluride Science Research Center for their hospitality during visits.
- Funding:** This material is based upon work supported by, or in part by, the U. S. Army Research Laboratory and the U. S. Army Research Office under contracts W911NF-13-1-0390 and W911NF-18-1-0028.
- Author Contributions Statement:** JPC, OPS, MLR, and GW conceived of the project. ABB, GW, and JPC developed the theory. SH provided the flux qubit and the experimental and analytical methods for its calibration. MHM, MLR, OPS, designed, implemented, and carried out the experiments. GW and JPC performed the calculations. JPC, OPS, MLR, and GW drafted the manuscript. JPC and MLR supervised the project.
- Competing Financial Interests Statement:** The authors declare that they have no competing financial interests.
- Data Availability:** The data that support the findings of this study are available from the corresponding author on reasonable request.

## Supplementary Materials

Materials and Methods: Derivation of trajectory-class fluctuation theorems, further discussion and interpretation, and experiment implementation, calibration, and work estimation methods.

## Supplementary Materials

### *Harnessing Fluctuations in Thermodynamic Computing via Time-Reversal Symmetries*

*One Sentence Summary:*

*Distinct distributions of thermodynamic work identify signatures of successful and failed information processing in a microscale flux qubit*

Gregory Wimsatt, Olli-Pentti Saira, Alexander B. Boyd, Matthew H. Matheny,  
Siyuan Han, Michael L. Roukes, and James P. Crutchfield

## Materials and Methods

The following presents details, derivations, and explanations for the theoretical claims and experimental results. First, we give a synopsis of recently developed principles of thermodynamic computing. Then, we introduce the model microscopic stochastic thermodynamical system, including its equations of motion and its physical calibration. The following sections explain how to use and interpret the trajectory-class fluctuation theorems and provide their derivation. Their practical application to work estimation for trajectories and classes, in light of the alternative kinds of work (inclusive and exclusive) and methods for experimental estimation, is laid out. A brief commentary on the substage work distributions then follows. Finally, we turn to describe the flux qubit, its equations of motion, implementation, calibration, and measurement.

### I. PRINCIPLES OF THERMODYNAMIC COMPUTING: A RECENT SYNOPSIS

A number of closely-related thermodynamic costs of computing have been identified, above and beyond the *house-keeping heat* that maintains a system's overall nonequilibrium dynamical state. First, there is the *information-processing Second Law* [28] 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 [29–31]. Third, the very modularity of a system's organization imposes thermodynamic costs [32]. Fourth, since computing is necessarily far out of equilibrium and nonsteady state, there are costs due to driving transitions between information-storage states [33]. Fifth, there are costs to generating randomness [34], which is itself a widely useful resource. Finally, by way of harnessing these principles, new strategies for optimally controlling nonequilibrium transformations have been introduced [27, 35–37].

### II. 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*  $\{(\beta(t), \delta(t)) : 0 \leq t \leq \tau\}$ . (See Supplementary Materials (SM) VI 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: (1) *drop barrier*, (2) *tilt*, (3) *raise barrier*, and (4) *untilt*, as shown in Table S1. 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 [25]. The effect of the control protocol on the system potential and system response is graphically displayed in Fig. 1.

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	$\frac{t_4-t}{t_4-t_3}$	

TABLE S1. Erasure protocol.

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

$$\begin{aligned} dx &= v dt \\ dv &= \sqrt{2k_B T \gamma / m} r(t) \sqrt{dt} - \left( \frac{\partial}{\partial x} V(x, t) + \gamma v \right) dt, \end{aligned} \quad (\text{S1})$$

where  $k_B$  is Boltzmann's constant,  $\gamma$  is the coupling between the heat reservoir and 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, snapshotted at times during the erasure substages, is shown in Fig. 1 (Inner plot sequence).

Reliable information processing dictates that we set time scales 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.

### III. TRAJECTORY-CLASS FLUCTUATION THEOREMS: USE 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 the section following.

First, consider a forward process distribution  $\mathcal{P}$ , defined by the probabilities of the system microstate trajectories  $\vec{z}$  due to an initial equilibrium microstate distribution evolving forward in time under a 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  $\vec{z} = (z_0, z_1, \dots, z_\ell)$  is  $\vec{z}^R = (z_\ell^R, \dots, z_1^R, z_0^R)$ , where  $z_i^R = -z_i$  if  $z_i$  is odd under time-reversal (e.g., momentum or spin), otherwise  $z_i^R = z_i$ . For a measurable subset of trajectories  $C \subset \vec{z}$ , let  $\langle \cdot \rangle_C$  denote an average over the ensemble of forward process trajectories conditioned on the trajectory 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  $C^R = \{\vec{z}^R | \vec{z} \in C\}$  in the reverse process, respectively. We first introduce a *trajectory-class fluctuation theorem* (TCFT) for the class-averaged exponential work  $\langle e^{-W/k_B T} \rangle_C$ :

$$\langle e^{-W/k_B T} \rangle_C = \frac{\mathcal{R}(C^R)}{\mathcal{P}(C)} e^{-\Delta F/k_B T}, \quad (\text{S2})$$

with  $\Delta F$  the system equilibrium free energy change. We also introduce a *class-averaged work* TCFT:

$$\langle W \rangle_C = \Delta F + k_B T \left( D_{\text{KL}} \left[ \mathcal{P}(\vec{z}|C) || \mathcal{R}(\vec{z}^R|C^R) \right] + \ln \frac{\mathcal{P}(C)}{\mathcal{R}(C^R)} \right). \quad (\text{S3})$$

This employs the Kullback-Liebler divergence  $D_{\text{KL}}[\cdot]$  taken between forward and reverse process distributions over all class trajectories  $\vec{z} \in C$ , conditioned on the forward class  $C$  and reverse class  $C^{\text{R}}$ , respectively. If we disregard this divergence, which is nonnegative and would generally be difficult to obtain experimentally, we then find the lower bound  $\langle W \rangle_C^{\text{min}}$  on the class-averaged work of Eq. (1).

In the limit of class  $C$  possessing only a single trajectory, we recover detailed fluctuation theorems as in Ref. [12]. If, however, we take  $C$  to be the entire set of trajectories  $\vec{Z}$ , we recover integral fluctuation theorems as in Jarzynski's equality [10]. Thus, the TCFTs are a suite that spans the space of fluctuation theorems between the extreme of the detailed theorems, that require very precise information about an individual trajectory, and the integral theorems, that describe the system's entire trajectory ensemble. SM IV below provides proofs for both TCFTs.

We can rearrange Eq. (S2) to obtain Eq. (2)—an expression for estimating equilibrium free energy changes:

$$\Delta F = -k_{\text{B}}T \ln \left( \frac{\mathcal{P}(C)}{\mathcal{R}(C^{\text{R}})} \langle e^{-W/k_{\text{B}}T} \rangle_C \right). \quad (\text{S4})$$

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 Jarzynski's equality:

$$\Delta F = -k_{\text{B}}T \ln \langle e^{-W/k_{\text{B}}T} \rangle_{\vec{Z}},$$

since rare microstate trajectories may generate negative work values that dominate the average exponential work [26]. 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.

In contrast, to efficiently determine the change in equilibrium free energy from Eq. (2), 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 RHS. 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:

$$\begin{aligned} E_C &= \langle W \rangle_C - \langle W \rangle_C^{\text{min}} \\ &= k_{\text{B}}T D_{\text{KL}} \left[ \mathcal{P}(\vec{Z}|C) \parallel \mathcal{R}(\vec{Z}^{\text{R}}|C^{\text{R}}) \right] \\ &\geq 0. \end{aligned}$$

The inequality in Eq. (1) is a refinement of the equilibrium Second Law and therefore the bound  $\langle W \rangle_C^{\text{min}}$  generally provides a more accurate estimate of the average work of trajectories in a class compared to the equilibrium free energy change  $\Delta F$ . 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 LHS of Eq. (S2) about the mean dimensionless work  $\langle W/k_{\text{B}}T \rangle_C$ . This shows that Eq. (1) becomes an equality when the variance and higher moments vanish. SM V below delves more into moment approximations. In any case, trajectory classes with narrow work distributions have small excess works  $E_C$ .

To estimate  $\mathcal{R}(C^{\text{R}})$ , we ran  $3.5 \times 10^6$  simulations of the reverse process. Table S2 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 on 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

## Simulation

Class $C$	$\langle W \rangle_C$	$\langle W \rangle_C^{\min}$	$E_C$	Partition $Q$	$\langle W \rangle_{\vec{Z}}$	$\langle W \rangle_Q^{\min}$	$E_Q$
All $\vec{Z}$	0.634	0.0	0.634	Trivial $\{\vec{Z}\}$	0.634	0.0	0.634
Success	0.713	0.683	0.030	Untilt-Centric I	0.634	0.560	0.074
Fail	-3.885	-3.951	0.066	Untilt-Centric II	0.634	0.601	0.032
Transitional	-0.546	-1.650	1.170				

## Experiment

Class $C$	$\langle W \rangle_C$	$\langle W \rangle_C^{\min}$	$E_C$	Partition $Q$	$\langle W \rangle_{\vec{Z}}$	$\langle W \rangle_Q^{\min}$	$E_Q$
All $\vec{Z}$	0.668	0.0	0.668	Trivial $\{\vec{Z}\}$	0.668	0.0	0.668
Success	0.742	0.643	0.099	Untilt-Centric I	0.668	0.554	0.115
Fail	-3.132	-3.475	0.343				
Transitional	-0.443	-1.215	0.772				

TABLE S2. (Top Left) Comparison of simulated class-average works and bounds for different trajectory classes: *All* trajectories  $\vec{Z}$ , *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  $E_C$ .  $3.5 \times 10^6$  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. (Top Right) Comparison of 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  $k_B T$ . (Bottom) Parallel results from the flux qubit experiment. *All* trajectories  $\vec{Z}$ , *Success*, *Fail*, and *Transitional* trajectories identified in Fig. 3(D). Data from 195,050 trajectories from the forward protocol and 250,000 trajectories from the reverse protocol.

class-average excess works is:

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

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

From Eq. (1), we see that  $\langle W \rangle_Q^{\min}$  is the coarse-grained lower bound on ensemble-average dissipation from Ref. [38]:

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

where  $D_{\text{KL}} [\cdot]$  is the Kullback-Liebler divergence between forward and reverse process distributions over the trajectory classes  $C \in Q$ . Since Kullback-Liebler divergences are nonnegative, such a bound always provides an improvement over the equilibrium Second Law. Table S2 shows both  $\langle W \rangle_Q^{\min}$  and  $E_Q$  for the trivial partition  $\{\vec{Z}\}$ , our three-class partition, labeled *Untilt-Centric I*, and the improved partition described in follow-on work, labeled *Untilt-Centric II*. In this case, the latter two also provide an improvement on the nonequilibrium Second Law which, assuming metastable starting and ending distributions, provides a lower bound on the average work equal to 0.533, the change in nonequilibrium free energy.

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  $k_B T \ln 2$ .

The experimental data matches these results, with the largest discrepancy occurring for the class-average excess for the Fail class. This is not wholly coincidence, since we determined the parameters of the experimental protocol by adjusting the parameters of a simple two-state Markov simulation to obtain a work distribution similar to that obtained by our Langevin simulations of the Duffing potential system described in the main text. However, it is interesting that this was sufficient to provide matches in both the clean decomposition of the total work distribution by trajectory class and the quantities of Table S2.

We also recover the equality of Ref. [38] for the ensemble-average work by averaging Eq. (S3) over each class:

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

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 will provide a basis for a natural classification scheme for trajectories that captures all relevant thermodynamics and information processing.

By changing our forward and reverse processes  $\mathcal{P}$  and  $\mathcal{R}$  to begin in system microstate distributions other than equilibrium, a yet-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 in depth elsewhere.

#### IV. TCFT DERIVATIONS

We now present derivations for the two TCFTs introduced in Eqs. (S2) and (S3).

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  $\pi_0$  for Hamiltonian  $\mathcal{H}_0$  and run a protocol until time  $\tau$ , 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  $\pi_\tau$ . 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 trajectory-wise system entropy difference is defined to be:

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

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

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

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  $\vec{\mathcal{Z}}$  with a different conditional probability:  $r(\vec{\mathcal{Z}}|z_0)$ .

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

$$\Delta S_{\text{res}}(\vec{\mathcal{Z}}) = -\beta Q(\vec{\mathcal{Z}}) = \ln \frac{p(\vec{\mathcal{Z}}|z_0)}{r(\vec{\mathcal{Z}}^R|z_\tau^R)} , \quad (\text{S5})$$

where  $\beta = 1/k_B T$ ,  $Q(\vec{\mathcal{Z}})$  is the net energy that flows out of the heat bath into the system given the trajectory  $\vec{\mathcal{Z}}$ , and  $(\cdot)^R$  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 [39]. Both our simulated Duffing potential system and flux qubit obey these requirements at sufficiently short time scales. Then we can express the total trajectory-wise 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:

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

Using Eq. (S5), we also have:

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

with:

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

Combining, we obtain a detailed fluctuation theorem:

$$\mathcal{R}(\vec{z}^{\text{R}}) = \mathcal{P}(\vec{z}) e^{-\beta(W(\vec{z}) - \Delta F)} . \quad (\text{S6})$$

From here, we derive our first TCFT by integrating each side of Eq. (S6) over all trajectories  $\vec{z}$  in a measurable set  $C \subset \vec{\mathcal{Z}}$ . Starting with the LHS and recalling the Iverson bracket  $[\cdot]$ , which is 1 when the interior expression is true and 0 when false, we have:

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

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  $\vec{z}$ . Integrating the RHS of Eq. (S6) then gives:

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



Combining, we have our first TCFT, Eq. (S2).

To obtain the second TCFT, we first change the form of Eq. (S6):

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

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

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

which gives Eq. (S3)'s TCFT.

## V. 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 LHS of Eq. (S2) 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 (S2) then gives:

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

Since  $e^{-x}$  is convex,

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

so  $a \geq 0$ . Then:

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

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

## VI. 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.

### A. 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 lab, 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 which can be sensibly partitioned into those for the system, heat bath, and lab. Then, we decompose the universe Hamiltonian into the following form:

$$\mathcal{H}(s, b, l) = H_B(b) + h_{S,B}(s, b) + H_S(s) + h_{S,L}(s, l) + H_L(l) ,$$

where  $s$ ,  $b$ , and  $l$  denote both the generalized coordinates and conjugate momenta for the system, bath, and lab, 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:

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

First, consider the subset of lab coordinates  $l$  for which  $h_{S,L}$  has nontrivial dependence. These so-called *protocol parameters*  $\lambda$  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 lab at time  $t = 0$ , a specific trajectory for the protocol parameters  $\{\lambda(t)\}_t$  for  $0 \leq t \leq \tau$  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 time scale of changes in the system coordinates, then, we ignore the system-bath interaction term  $h_{S,B}$  in writing the system Hamiltonian  $\mathcal{H}'$ :

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

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.

### B. 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  $\text{Pr}(s)$  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 lab as *work* and all energy exchanged between the system and heat bath as *heat*. Since the lab is directly coupled only to the system, the work they exchange is given by the change in energy of the lab'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*  $W_0$  [40]. Specifically:

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

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:

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

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:

$$\frac{dW}{dt} = \frac{dW_0}{dt} + \frac{dh_{S,L}}{dt} . \quad (S7)$$

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:

$$\begin{aligned}\frac{dW}{dt} &= -\frac{dH_L}{dt} \\ &= \frac{\partial h_{S,L}}{\partial \lambda} \frac{d\lambda}{dt} .\end{aligned} \quad (S8)$$

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

$$W(t, t') = \int_t^{t'} dt'' \frac{d\lambda}{dt''} \frac{\partial h_{S,L}(s, \lambda)}{\partial \lambda} \Big|_{\lambda=\lambda(t'')} . \quad (S9)$$

The exclusive work  $W_0$  has a corresponding form:

$$\begin{aligned}\frac{dW_0}{dt} &= -\frac{d(h_{S,L} + H_L)}{dt} \\ &= -\frac{\partial h_{S,L}}{\partial s} \frac{ds}{dt} ,\end{aligned} \quad (S10)$$

For the case where  $h_{S,L}$  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:

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

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.

### C. 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:

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

For cyclic protocols where  $V(\cdot, \lambda_0) = V(\cdot, \lambda_\tau)$ , such as in our erasure operation,  $h_{S,L}(s(t), \lambda(t))$  vanishes for all trajectories at  $t = 0, \tau$ . By Eq. (S7) 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  $h_{S,L}$  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, since the device's physics precludes precise measurements of system flux  $\phi$ —the relevant part of  $s$  for determining the potential  $h_{S,L}$ . Instead, we do have reliable measurement of large and stable changes in the flux  $\phi$ . 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  $\phi$  to approximate the exclusive work contribution at each moment in time. Then, adding up these contributions yields an approximation to 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  $\phi$ . We develop the approximation in two steps.

#### 1. 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  $h_{S,L}(\cdot, \lambda(t))$  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 must be zero, since  $h_{S,L}$  does not change with  $s$ . During a

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

$$\begin{aligned}\Delta W_0^{\text{trans}} &= \int d\phi \left( -\frac{\partial h_{S,L}}{\partial \phi} \right) \\ &\approx h_{S,L}(w_0, \lambda(t)) - h_{S,L}(w_1, \lambda(t)) ,\end{aligned}\tag{S11}$$

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.

## 2. 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  $\phi_i^t$ . 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  $\Delta t$  during which the potential  $V(\cdot, t)$  changes little but long enough 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  $\phi^t$  and at the end as  $\phi^{t+\Delta t}$ . By Eq. (S7), the exclusive work contribution  $\Delta W_0^{\text{trans}}$  is the difference of the inclusive work contribution and the change in system-lab interaction term  $\Delta h_{S,L}$ . The change  $\Delta h_{S,L}$  can itself be broken down into two terms, one for the difference in  $h_{S,L}$  between the informational-state minima and the other for the change in  $h_{S,L}$  local to the respective minima. In other words:

$$\begin{aligned}\Delta h_{S,L} &= h_{S,L}(\phi^{t+\Delta t}) - h_{S,L}(\phi^t) \\ &= [(h_{S,L}(\phi^{t+\Delta t}) - h_{S,L}(\phi_j^{t+\Delta t})) + h_{S,L}(\phi_j^{t+\Delta t})] - [(h_{S,L}(\phi^t) - h_{S,L}(\phi_i^t)) + h_{S,L}(\phi_i^t)] \\ &= [h_{S,L}(\phi_j^{t+\Delta t}) - h_{S,L}(\phi_i^t)] + [(h_{S,L}(\phi^{t+\Delta t}) - h_{S,L}(\phi_j^{t+\Delta t})) - (h_{S,L}(\phi^t) - h_{S,L}(\phi_i^t))] \\ &= \Delta m^t + \Delta l^t ,\end{aligned}\tag{S12}$$

where  $\Delta m^t$ , Eq. (S12)'s first term, is the change in  $h_{S,L}$  at the informational-state minima and  $\Delta l^t$ , Eq. (S12)'s second term is the change in  $h_{S,L}$  of the system with respect to the informational-state minima. Our protocol ensures that the total number of transitions is so small and the time durations so narrow that we can ignore the total contributions of inclusive works  $\Delta W^{\text{trans}}$  due to these transition durations. Then, we approximate the exclusive work contribution during a transition via:

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

Suppose, now, that the system remains in one informational state  $i$  during a time interval  $\Delta t$ . Since the relaxation rate is fast compared to the duration  $\Delta t$ , 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  $\Delta W^{\text{stay}}$  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 simultaneously shift the entire potential up by a given amount, we add an inclusive work contribution equal to the potential shift but the system trajectory is unchanged. 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:

$$\Delta W^{\text{stay}} = \Delta W_s + \Delta m^t, \quad (\text{S14})$$

where  $\Delta W_s$  is the inclusive work contribution due to the change in potential shape at the informational state. Equation (S13) applies equally well here in describing the change in system-lab interaction term. Thus, the exclusive work contribution  $\Delta W_0^{\text{stay}}$  for this time interval is:

$$\Delta W_0^{\text{stay}} = \Delta W^{\text{stay}} - \Delta h_{\text{S,L}} \quad (\text{S15})$$

$$= \Delta W_s - \Delta l^t. \quad (\text{S16})$$

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 the contributions. The sum of all local  $h_{\text{S,L}}$  changes  $\Delta l^t$  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: At any time  $t$ , the inclusive work contribution  $\Delta W_s$  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 effect on the well shapes and even less in making those shapes distinct.

Then, we can assume that the sum of  $\Delta W_s$  for a trajectory is the same as that staying in one informational state the entire time. Since the protocol is very slow and cyclic, though, a particle that stays in one informational state the entire time must receive approximately zero inclusive work  $W$ . Given that the sum over all  $\Delta W_s$  must be equal to  $W$  for such a trajectory, it 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:

$$W_0(0, \tau) = - \sum_{\text{trans}} \Delta m^t. \quad (\text{S17})$$

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.

## VII. SUBSTAGE WORK DISTRIBUTIONS COMMENTARY

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

The distributions for barrier dropping and tilting are narrow, symmetric peaks; see Fig. 1 (Outer left plots). Barrier raising also has a rather narrow peak, composed primarily of trajectories always in the  $R$  state, but also exhibits a bulge toward positive work; see Fig. 1 (Top right). Note that the  $L$  state is created mid-way 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 presents 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  $7 k_B T$ , resulting in a narrow, peaked work distribution with a mean near  $7 k_B T$ . 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 trajectories that do 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. 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.

## VIII. FLUX QUBIT DEVICE, CALIBRATION, AND MEASUREMENT

The benefits of the flux qubit device are several-fold. First, their physics provide 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 [19, 21, 41]. 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 manufacturing technology led by efforts in quantum computing. Fourth, being constructed via modern integrated circuit technology they form the basis of a technology that will 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 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 section lays out the basic physics of the flux qubit device and details of the experimental implementation.

### A. 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. [42]. 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  $2\ell$  is the loop inductance,  $I_{c0} = i_{c1} + I_{c2}$  is the sum of critical

currents of the two junctions, and  $\Phi_0$  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  $\Phi$  through the rf SQUID loop and  $\Phi_{\text{dc}}$  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 [42]:

$$\begin{aligned} 2C\ddot{\Phi} + \frac{\dot{\Phi}}{R/2} &= -\frac{\partial U(\Phi, \Phi_{\text{dc}})}{\partial \Phi} \quad \text{and} \\ \frac{C}{2}\ddot{\Phi}_{\text{dc}} + \frac{\dot{\Phi}_{\text{dc}}}{2R} &= -\frac{\partial U(\Phi, \Phi_{\text{dc}})}{\partial \Phi_{\text{dc}}} . \end{aligned} \quad (\text{S18})$$

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

$$f(\phi, \phi_{\text{dc}}) = \frac{1}{2}(\phi - \phi_x)^2 + \frac{\gamma}{2}(\phi_{\text{dc}} - \phi_{\text{xdc}})^2 - \beta_0 \cos \frac{\phi_{\text{dc}}}{2} \cos \phi + \delta\beta \sin \frac{\phi_{\text{dc}}}{2} \sin \phi , \quad (\text{S19})$$

where  $U_0 = \Phi_0^2/(4\pi^2 L)$ . Here,  $\gamma = L/(2\ell)$  is the ratio of rf and dc SQUID inductances;  $\phi_x$  ( $\phi_{\text{xdc}}$ ) is the external flux applied to the rf (dc) SQUID loop;  $\phi$  ( $\phi_{\text{dc}}$ ) 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  $\phi_x$  ( $\phi_{\text{xdc}}$ ) is  $2\pi$  ( $4\pi$ ) 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  $(\phi_x, \phi_{\text{xdc}}) \rightarrow (\phi_x + \alpha\phi_{\text{xdc}}, \phi_{\text{xdc}})$ , with the coefficient  $\alpha$  chosen such that the equilibrium population of the left and right wells was unaffected to first order by the barrier control  $\phi_{\text{xdc}}$ .

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  $\phi$  describes the in-phase motion of the two junctions that results in a current circulating in the rf SQUID loop. The dynamical variable  $\phi_{\text{dc}}$  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  $f(\phi, \phi_{\text{dc}})$  and the energy scale of the potential is determined by  $U_0$ . With suitable device parameters and applied fluxes ( $\phi_x$  and  $\phi_{\text{xdc}}$ ) one obtains a double-well potential. The barrier height  $\Delta U$  separating the two wells is readily adjusted by varying  $\phi_{\text{xdc}}$ . The effective potential is plotted in Fig. 3(B) with parameters:  $\beta_0 = 6.2$ ,  $\gamma = 12$  and  $\delta_b = 0.2$ .

## B. Experimental implementation

The junctions were  $1 \times 1 \mu\text{m}^2 \text{Nb}/\text{Al}_2\text{O}_3/\text{Nb}$  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. [42]) for calibrating the flux qubit parameters. An outline of the steps is given below. A complete description of the measurements is presented elsewhere.

First, by executing wide-range sweeps of the coil currents  $I_{\text{tilt}}$  and  $I_{\text{barrier}}$ , parameter values corresponding to single-valued and bistable potential landscapes are recorded. A linear transformation from  $I_{\text{tilt}}$  and  $I_{\text{barrier}}$  to  $(\phi_x, \phi_{\text{xdc}})$  is established by matching the experimental periodicity with the theoretical one ( $2\pi, 4\pi$ ). Linear cross-talk from  $I_{\text{barrier}}$  to  $I_{\text{tilt}}$  is calibrated by orthogonalizing the global response. Cross-talk from  $I_{\text{tilt}}$  to  $I_{\text{barrier}}$  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  $\phi_{\text{xdc}} = 0$  and the differential flux response  $d\langle\phi\rangle/d\phi_x$  at  $\phi_{\text{xdc}} = 2\pi$  to theoretical predictions. The prefactor  $U_0 = 56.3$  K is determined by equating the observed escape energy for inter-well 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 cross-over 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.