Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Learning to extract work optimally from unknown quantum states with exponentially reduced dissipation

Josep Lumbreras

joseplumbreraszarapico@gmail.com

Nanyang Technological University https://orcid.org/0009-0002-3541-6148

Ruo Cheng Huang

Nanyang Technological University

Yanglin Hu

National University of Singapore https://orcid.org/0009-0000-1105-589X

Mile Gu

Nanyang Technological University https://orcid.org/0000-0002-5459-4313

Marco Tomamichel

National University of Singapore https://orcid.org/0000-0001-5410-3329

Article

Keywords:

Posted Date: May 11th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-9431975/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: There is **NO** Competing Interest.

Learning to extract work optimally from unknown quantum states with exponentially reduced dissipation

Josep Lumbreras,[1,2,*] Ruo Cheng Huang,[1,†] Yanglin Hu (胡杨林),[3,4,‡] Mile Gu,[1,2,5,§] and Marco Tomamichel[3,5,6,¶]

[1] *Nanyang Quantum Hub, School of Physical and Mathematical Sciences, Nanyang Technological University, Singapore*
[2] *Centre for Quantum Technologies, Nanyang Technological University, Singapore*
[3] *Centre for Quantum Technologies, National University of Singapore, 3 Science Drive 2, Singapore*
[4] *QICI Quantum Information and Computation Initiative, School of Computing and Data Science, The University of Hong Kong, Pokfulam Road, Hong Kong SAR, China*
[5] *MajuLab, CNRS-UNS-NUS-NTU International Joint Research Unit, UMI 3654, 117543, Singapore*
[6] *Department of Electrical and Computer Engineering, National University of Singapore*

(Dated: April 14, 2026)

A pure qubit is a source of non-equilibrium free energy, but this energy cannot be fully extracted when the qubit state is unknown. Here we consider an agent tasked with harvesting work from a sequence of N identically prepared pure qubits, without quantum memory to store them for joint processing. The agent must balance two competing objectives: extracting as much work as possible from each qubit given its current knowledge, while learning more about the state to improve extraction in subsequent rounds. We show that this learning cost can be reduced exponentially relative to existing approaches. By leveraging the exploration-exploitation trade-off from reinforcement learning, we construct fully adaptive work-extraction protocols whose cumulative dissipation scales logarithmically in N, compared to the square-root scaling of existing strategies. Our results identify a concrete setting in which the thermodynamic cost of learning an unknown quantum state can be sharply reduced, and reveal a precise connection between reinforcement learning and quantum thermodynamics.

The Szilard engine illustrates a central thermodynamic principle: the work that can be extracted from a system depends fundamentally on what we know about that system. Consider a particle trapped in one of two partitions of a box (Fig. 1) [25]. To extract the available free energy, we must correctly identify which side the particle occupies. Guess incorrectly, and useful work is lost through irreversible entropy production. This principle extends to quantum systems. A pure qubit in state $|\psi\rangle$ at temperature T, carries non-equilibrium free energy $k_B T \ln 2$. However, extracting this work requires tailoring the extraction procedure to $|\psi\rangle$. Any mismatch between the procedure and the true state leads to a loss of useful work, manifested as dissipation [17, 19, 23].

Yet in many natural settings, we do not begin with knowledge of the free-energy sources available to us. Consider an organism–an agent–placed in a foreign environment where free energy comes from sequential encounters with qubits, each identically prepared in some state $|\psi\rangle$. The agent does not initially know $|\psi\rangle$ and cannot store quantum memory. It must therefore decide how best to exploit each qubit without the capacity for joint operations. This setting arises naturally in the context of a simple organism foraging for free energy with limited physical reach [19]. Without learning something about $|\psi\rangle$, the agent is unable to extract useful work. Yet learning itself carries a direct thermodynamic cost: every inaccurate interaction dissipates part of the free energy that would have been accessible under perfect knowledge. How, then, should the agent minimize such dissipation?

One immediate–and currently best-known–approach is "tomography-first". The idea is to sacrifice a fraction of the available qubits to estimate $|\psi\rangle$, and then apply stan-

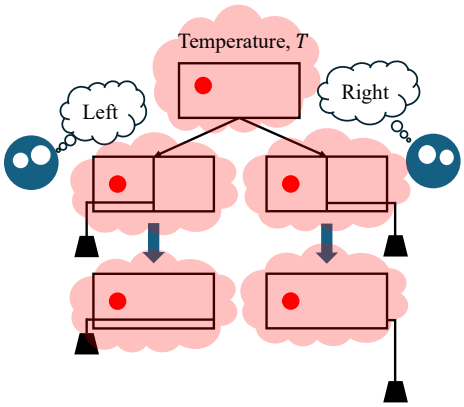


Figure 1. **Illustration of Szilard's engine.** A particle moves freely in a box at temperature T. The agent inserts a frictionless partition and must attach a weight to the side containing the particle in order to extract work. In the left column, the agent attaches the weight to the correct side, lifts it, and extracts free energy. In the right column, the agent attaches the weight to the wrong side, causing the weight to fall and indicating a loss of free energy. This illustrates that the agent can transfer the particle's free energy into the potential energy of the weight only when the coupling is correctly matched to the particle's state.

* josep.lumbreras@u.nus.edu
† ruocheng001@e.ntu.edu.sg
‡ yanglin.hu@u.nus.edu
§ mgu@quantumcomplexity.org
¶ marco.tomamichel@nus.edu.sg

dard state-aware work-extraction methods to the remaining qubits [2, 4, 8, 11, 24]. Such strategies are effective, in that they eventually learn $|\psi\rangle$ and enable an asymptotic extraction rate of $k_BT\ln 2$. However, this learning incurs a significant cost. Given a sequence of N qubits, the optimal such strategy sacrifices $\Theta(\sqrt{N})$ qubits in this way [16, 28, 29]. As a result, the cumulative dissipative cost of learning–the gap between the total free energy available in the N qubits and the total work actually extracted–scales as $O(\sqrt{N})$.

Here, we show how to reduce this learning cost by an exponential factor. Instead of committing each qubit exclusively to either learning or harvesting, we design an agent that leverages ideas from reinforcement learning to balance exploration–choosing procedures that reveal more about $|\psi\rangle$–against exploitation–choosing procedures that extract as much energy as possible given its current knowledge of $|\psi\rangle$. The resulting agent extracts nearly all available free energy, incurring a total cumulative dissipation that scales only as $O(\text{polylog}(N))$. In the limit of large N, the agent fully learns $|\psi\rangle$, and the dissipation rate therefore vanishes, in alignment with recently proposed state-agnostic work-extraction protocols [29]. Because the protocol is sequential, it requires neither collective measurements nor coherent interactions between qubits. More broadly, our results reveal a surprising link to reinforcement learning, showing that total *regret*–a key cost function in that domain–has a precise physical interpretation: it is the cumulative thermodynamic dissipation necessitated by a learning agent.

RESULTS

Overarching Approach

We consider a setting in which an agent is tasked with extracting energy from a source S that emits a qubit A_k at each time step k. The agent knows that every emitted qubit is identically prepared in the same unknown pure state $|\psi\rangle$, with corresponding density operator $\psi := |\psi\rangle\langle\psi|$. However, they have no knowledge of what $|\psi\rangle$ is. We also assume the agent has no quantum memory, such that qubits received at earlier times cannot be stored and jointly processed with later emissions. How much energy can such an agent extract over N time steps?

For conceptual clarity, we first assume that H_A, the Hamiltonian governing each qubit, is degenerate. The general case is treated in Supplementary Material [1, Section IIID]. In this setting, each qubit carries nonequilibrium free energy $k_BT\ln 2$. However, harvesting this energy as useful work requires knowledge of $|\psi\rangle$. More precisely, any extraction protocol must be tailored to some target input state σ. If σ does not exactly match ψ, the resulting dissipation is lower bounded by the quantum relative entropy $k_BTD(\psi\|\sigma)$, where $D(\rho\|\sigma) = \text{tr}[\rho(\ln\rho - \ln\sigma)]$, as shown in Ref. [23]. Thus, the agent's ability to extract useful work from the source depends critically on how well it has learned $|\psi\rangle$.

The best currently known strategy is the "tomography-first" approach. The idea is to devote an initial fraction α of the N qubits entirely to learning $|\psi\rangle$. This immediately sacrifices $\alpha N k_BT\ln 2$ of free energy, yielding an estimate whose relative-entropy error is bounded above by $\zeta = \Omega\big(\frac{1}{\alpha N}\big)$ [9, 12, 13], even if the agent is allowed to interact coherently with all αN training qubits. The resulting estimate is then used to extract work from the remaining $(1-\alpha)N$ qubits. The total dissipation is therefore lower bounded by

$$W_{\text{diss}}(N) = \min_{\alpha}\ \Omega\left(\alpha N + \frac{1-\alpha}{\alpha}\right) = \Omega(\sqrt{N})\ . \qquad (1)$$

This bound is tight: it can be matched by an upper bound $W_{\text{diss}}(N) = O(\sqrt{N})$ using an optimal learning algorithm with respect to relative entropy [12]. At first sight, this scaling appears to be a fundamental barrier imposed by the sample complexity of state tomography.

Our adaptive protocol bypasses this barrier by asking: *can we simultaneously extract useful energy while learning about $|\psi\rangle$?* Our key observation is that there is no need to allocate each qubit exclusively to either tomography or work extraction; it is possible to combine both. At each time step k, the agent assigns a belief state ψ_k, representing its best guess of ψ, together with an uncertainty parameter ϵ_k. The agent then enacts an energy-extraction protocol tailored to $\rho_k = (1-\epsilon_k)\psi_k + \epsilon_k\mathbb{1}/2$, and we show that the dissipation incurred during harvesting is characterized by the cumulative dissipation

$$W_{\text{diss}}(N) \coloneqq k_BT\sum_{k=1}^{N} D(\psi\|\rho_k)\ . \qquad (2)$$

Our main result demonstrates that, by striking a delicate balance between exploration (learning more about $|\psi\rangle$) and exploitation (extracting work using that knowledge), the agent can asymptotically learn ψ, such that $\rho_k \to \psi$ in the limit of large k. Meanwhile, the cumulative dissipation–which quantifies the thermodynamic cost of learning ψ–scales as

$$W_{\text{diss}}(N) = O\left(\ln^2(N)\ln\left(\frac{N}{\delta}\right)\right)\ , \qquad (3)$$

with probability at least $1-\delta$. This constitutes the first explicit characterization of optimal work extraction from unknown quantum states beyond the asymptotic many-copy limit.

In the asymptotic regime, as $N\to\infty$, our result implies that the average dissipation per copy vanishes, recovering the behavior of canonical state-aware protocols [2, 4, 8, 24] and state-agnostic protocols in the infinite-copy regime [16, 29], thereby bridging the gap between regimes of ignorance, perfect knowledge, and unlimited training data.

Protocol Specifics

We formalize extracting work from free energy within a quantum system by measuring how well we can use it to charge an external battery when in contact with a thermal bath Fig. 2) [15, 24]. The physical setup consists of the following components:

- Battery (System B): The storage medium for extracted work. It is modeled by a classical weight that can be raised and lowered, reflecting its potential energy.
- Thermal reservoir (System R): Reservoir held at *fixed* temperature T, which supplies us with qubits in Gibbs state, $\gamma = Z^{-1}e^{-H_R/k_BT}$, where H_R is the Hamiltonian and Z the partition function. Following standard conventions in quantum resource theory, we assume these Gibbs states are available at no thermodynamic cost [7, 8, 14, 20].

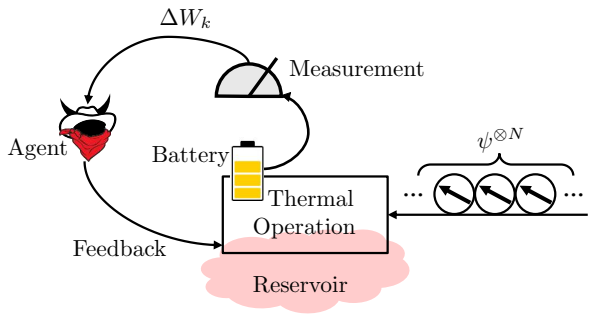


Figure 2. **Sketch of the sequential work extraction protocol with a thermal reservoir.** At each time step $k \in [N]$, the agent receives a copy of an unknown qubit state ψ and performs a thermal operation involving the reservoir and a battery. A measurement of the battery system is carried out to determine the extracted work ΔW_k, which is then used as feedback to improve the extraction strategy in subsequent rounds.

Under this backdrop, the agent can interact with a stream of qubit, all initialized in the same unknown quantum state ψ. Since the agent begins without prior knowledge of the state ψ, it cannot implement a perfect extraction protocol from the outset. Instead, it must gradually improve its strategy by learning from the observations on the battery system. A schematic diagram can be found in Fig. 2. The agent extracts free energy from A and stores it in the battery B through interactions with the thermal reservoir R.

At each round $k \in [N]$, the agent's physical procedure to harvest energy from ψ is governed by *two* key parameters: the target state ψ_k the agent to gamble ψ is in, and $\epsilon_k \in \mathbb{R}$ representing the agents *uncertainty* in their choice. In methods, we show that this results in a work extraction protocol that achieves the theoretically optimal work extraction when the input state is precisely $\rho_k = \Delta_{\epsilon_k}(\rho) = (1-\epsilon)\psi_k + \epsilon\mathbb{1}/2$.

Application of the protocol (see methods for details)

Algorithm 1: Adaptive Work Extraction

Require: A source that emits ψ, a battery system, a reservoir at temperature T, inital guess ψ_1 and initial error ϵ_1.

Procedure: At each round $k \in [N]$

1. Receive unknown ψ.
2. Design the extraction protocol tailored for $\rho_k = (1-\epsilon)\psi_k + \epsilon_k\mathbb{1}/2$.
3. Apply the designed protocol onto ψ. This process couples the system ψ and the battery B while interacting with the reservoir R.
4. Observe and record the battery's energy, μ_k.
5. Compute work extracted $\Delta W_k = \mu_k - \mu_{k-1}$.
6. Compute new estimate $\psi_{k+1}, \epsilon_{k+1}$ based on the whole history $\{\Delta W_{k'}\}_{k'=1}^{k}$ and $\{\psi_{k'}\}_{k'=1}^{k}$.

results in a discrete probability distribution for the work extracted, such that

$$\begin{aligned} \Pr(\Delta W_k = w_{k,0}) &= |\langle\psi|\psi_k\rangle|^2, \\ \Pr(\Delta W_k = w_{k,1}) &= 1 - |\langle\psi|\psi_k\rangle|^2, \end{aligned} \tag{4}$$

with

$$\begin{aligned} w_{k,0} &:= k_BT\big[\ln 2 + \ln(1-\epsilon_k/2)\big] \ , \\ w_{k,1} &:= k_BT\big[\ln 2 + \ln \epsilon_k/2\big] \ . \end{aligned} \tag{5}$$

The expected extracted work at round $k \in [N]$ is

$$\mathbb{E}[\Delta W_k] = k_BT\left[\ln 2 - \mathrm{D}(\psi\|\rho_k)\right] \ . \tag{6}$$

The first term is the non-equilibrium free energy of the unknown state. This is attained when the agent is fully committed to their gamble ($\epsilon_k = 0$) and the gamble is precisely on point ($\psi_k = \psi$). Otherwise, the agent incurs a penalty in the form of dissipation in the second term. Accordingly, the expected dissipation is

$$W_{\mathrm{diss}}^k := k_BT\mathrm{D}(\psi\|\rho_k). \tag{7}$$

This expression highlights a key trade-off: reducing dissipation toward 0 requires the agent to fully commit to his gamble, such that $\epsilon_k \to 0$. However, ϵ_k cannot be made arbitrarily small unless the agent's guess ψ_k is precisely on point ($\psi_k = \psi$); otherwise, the second term in Eq. (6) diverges and the agent's risks losing all work extracted so far. Consequently, achieving sub-linear heat dissipation in N requires our agent to asymptotically learn ψ to perfect precision, ensuring that they can fully commit to their gambles in the limit of large k.

Extracting work while learning

Traditional methods for achieving sublinear dissipation involve tomography of the first $\sqrt{N}$ qubit–essentially using these qubits to maximize exploration (learning about

ψ while sacrificing its free energy completely), and using this knowledge to maximize exploitation (energy extraction) on the qubits that remain. Our innovation here is to observe that there is no need for such a stark classification. It is possible to balance performance and weight for both. In Methods, we show that the work distribution in Eq. (4) serves as a projective measurement: by observing how much work is extracted, we have effectively measured the qubit along the $\{\psi_k, \psi_k^\perp\}$ basis. Our agent can use this information to learn more about ψ.

The key is for our agent to balance exploration vs. exploitation–taking inspiration from the multi-armed bandit framework [3, 5, 6, 10, 22]. That is, our agent must adaptively choose ψ_k to balance information gain and work extraction. In essence, the agent chooses directions close to the true state ψ that still provide useful learning information.

To decide which direction ψ_k we should *gamble* on, we leverage the optimistic principle from reinforcement learning [5, 18] adapted to our thermodynamic setting: among all pure-state estimates still compatible with the data collected so far, we pick probe directions that are expected to (i) *remain nearly aligned* with the state (to keep dissipation small), while (ii) being *maximally informative* in the directions where our current knowledge is most uncertain (to speed up learning).

More concretely, given a target success probability $1-\delta$, the agent first groups $4t$ consecutive rounds as a single stage. It then operates as follows:

- **Estimation and confidence region.** At the beginning of stage K^*, where $k = 4tk^* + 1$, the agent already extracted work from the last $4tk$ qubits. Thus, it has a record of all work values extracted so far (i.e., $\{\Delta W_s\}_{s=1}^{k}$) and the associated target state $|\psi\rangle_k$. From this, the agent forms a weighted least-squares estimate and constructs a confidence region $\tilde{\mathcal{C}}_k$ over which ψ resides with probability at least $1 - k\delta/N$. Geometrically, we can represent this region by the projection of an ellipsoid C_k on the surface of the Bloch sphere. (see Fig. 3 and methods for details).

- **Probe Selection**. Our agent then selects *four* states $\{|\hat{\psi}_{i,k}\rangle\}_{i=1}^{4}$ on $\tilde{C}_k$, geometrically corresponding to picking appropriate extremal points on C_k and projecting them onto the Bloch sphere. The agent will then attempt work extraction with a target state of $\psi_{i,k}$ and uncertainty ϵ_k over t consecutive qubits.

- **Confidence Update**. The agent then collates the resulting work values, and thus now has information on all work values extracted up $\}_{s=4k+1}^{4t(k^*+1)}$. There is then a systematic method for the agent to define a new confidence region C_k, with $k = 4t(k^*+1)$ (see methods).

Our agent initially begins with C_1 as the entirety of the Bloch sphere, whereby the initially $|\psi_{i,k}\rangle$ can be almost

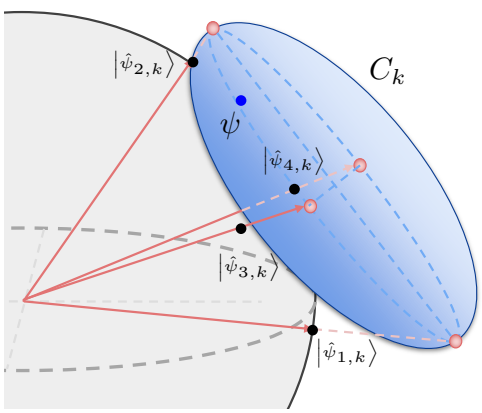


Figure 3. **Geometric representation of the adaptive learning strategy on the Bloch sphere.** At round k, the algorithm constructs a confidence region $\mathcal{C}_k$ (shaded blue ellipsoid) centered on a least-squares estimator $|\hat{\psi}_k\rangle$. This region is bounded to contain the true unknown state $|\psi\rangle$ with high probability. The protocol then selects four new probe estimators, $\{|\hat{\psi}_{i,k}\rangle\}_{i=1}^{4}$, corresponding to the extreme points of $\mathcal{C}_k$ projected onto the surface of the Bloch sphere. These boundary states serve as the targeted inputs for the next work extraction rounds, keeping the estimates close to the true state to minimize dissipation while simultaneously probing the axes of highest uncertainty.

arbitrarily chosen. We then have theoretical guarantees on how close the estimates ψ_k get to the unknown state ψ. Specifically, their infidelity $1 - |\langle\psi_k|\psi\rangle|^2$ scales as $O(\ln(N)/k)$ (see Supplemental Material [1, Section II]), for all $k \in \{1, 2, \ldots, N\}$. The estimates thus become increasingly more aligned with the true state as the rounds progress.

This bound on infidelity implies that our agent can become more confident in his guesses over time. In methods, we show that by setting ϵ_k to scale as $O(\ln(N)/k)$, and bounding $\mathrm{D}(\psi\|\Delta_{\epsilon_k}(\psi_k))$, we obtain

$$W_{\mathrm{diss}}(N) \leq k_B T \sum_{k=1}^{N} 8\epsilon_k(2 - \ln \epsilon_k/2). \tag{8}$$

As k increases towards N in the limit of large N, $\epsilon_k \to 0$, reflecting our agents' growing confidence in their guess as they gain more information on ψ. Since we can set $\epsilon_k \sim 1/k$, the cumulative contribution to the dissipation in (8) leads to the desired polylogarithmic behavior as outlined in Equation 3.

We illustrate our theoretical analysis with numerical simulations that explicitly illustrate the polylogarithmic dissipation predicted by our bounds. Figure 4 reports both the cumulative dissipation and the dissipation rate, defined as $W_{\mathrm{diss}}(N)/N$, as functions of the number of available copies. The simulations confirm that the cumulative dissipation grows polylogarithmically with N, while the dissipation rate decays to zero, in agreement with the dissipation guarantees of our protocol. Importantly, this behavior is already visible in the finite-copy

regime accessible to realistic implementations, indicating that the polylogarithmic scaling is not purely asymptotic. Together, these results demonstrate that our exploration–exploitation strategy achieves simultaneously accurate learning and minimal thermodynamic dissipation in practice, thereby providing strong numerical support for the theoretical polylogarithmic dissipation guarantees established above.

Finally, we verify that our results remain robust when accounting for the thermodynamic costs of memory erasure incurred during the agent's operation. Since our model involves measurements in the battery's energy eigenbasis, the associated reduction in entropy implies an energetic cost for memory erasure (Landauer's principle). In Supplementary Material [1, Section IV], we provide an explicit proof that incorporating these information-processing costs does not alter the scaling: the total dissipation remains $O(\text{polylog}(N))$.

DISCUSSION

Here, we considered the problem of state-agnostic sequential work extraction–where an agent is tasked to extract free energy from a sequence of N qubits all initialized in some unknown quantum state $|\psi\rangle$. To do this, the agent must necessarily learn $|\psi\rangle$ during the process, necessitating extra heat dissipation. Prior state-of-the-art methods incurred heat dissipation of $\Theta(\sqrt{N})$; such that $O(1/\sqrt{N})$ of all available free energy is wasted. Here, we observed that the act of observing the amount of energy we extract in itself allows an agent to learn about $|\psi\rangle$. By introducing a family of energy-extraction protocols that balanced the energy extracted with how much information our agent could learn, we demonstrated an energy extraction protocol that reduces heat dissipation of learning to $O(\text{polylog}(N))$–an exponential improvement over prior state-of-art.

There are several natural directions of further research. The first is to consider the relation of our work with that of state-agnostic energy extraction protocols that operate in the thermodynamic ($N \to \infty$) limit [28, 29], where it was shown that the heat dissipation per qubit can be made arbitrarily small. Our agentic approach naturally recovers such results; we scale N, and our cumulative heat dissipation per qubit scales to $O(\text{polylog}(N)/N) \to 0$. Such a comparison would naturally also require us to translate results across different models of quantum work extraction. Indeed, our studies in this area already yield some promising results. In the supplementary materials [1, Section III], we illustrate our adaptive algorithm in an alternative setting where the battery is modeled as a uniform energy ladder. Our cumulative dissipation similarly exhibits $O(\text{polylog}(N))$ scaling, suggesting that our protocol's advantage is robust across different thermodynamic realizations.

It would also be interesting to see if these asymptotic energy-extraction protocols can be adapted to the finite-copy domain, and if so, their resulting cumulative heat dissipation as a function of N. The second is in terms of general agnostic information ratchets, where the source is no longer a sequence of identically prepared unknown quantum states, but rather a source of correlated quantum states generated by a hidden Markov model. Such scenarios have been extensively studied in settings where the hidden Markov models are known, but what if our agent had to learn them in real time? Meanwhile, free energy is just one potential resource. It would certainly be exciting to explore if our reasoning here can extend to the extraction and distillation of other resources–such as entanglement, coherence, or magic.

METHODS

Work extraction protocol

The details of the work extraction protocol first proposed in [24] will be reviewed here. Aside from an unknown source, this protocol models the battery with a semi-classical realization. It is a semi-classical weight described by a continuous-variable state $\varphi(x) \in L^2(\mathbb{R})$. The battery Hamiltonian is defined as $H_B\varphi(x) = x\varphi(x)$, where x represents the height of the weight. The energy of the battery can be changed by translating the weight up by a certain height $\mathrm{d}x$, described by the translation operator $\Gamma^B_{\mathrm{d}x}\varphi(x) = \varphi(x - \mathrm{d}x)$.

The protocol also assumes access to a thermal reservoir, R, at a *fixed* temperature T with a *tunable* Hamiltonian $H_R(\nu) = \nu\,|1\rangle\langle 1|$. Here, $|0\rangle$ and $|1\rangle$ are energy eigenstates and ν is an energy gap that can be tuned. This reservoir supplies qubit thermal states $\gamma_\beta(\nu) = Z(\nu)^{-1}e^{-H_R(\nu)/k_BT}$.

In every round $k \in [N]$, the agent first performs a unitary on the system qubit in the form of $U_k = |0\rangle\langle\psi_k| + |1\rangle\langle\psi_k^\perp|$, satisfying $[H_A, U_k] = 0$. (Note that, if $\psi_k = \psi$, this unitary simply diagonalizes the system qubit in the computational basis.)

The agent then performs M thermal operations by repeatedly applying an energy-conserving unitary on a combined system A_kBR_ℓ to transfer energy from a fresh reservoir qubit R_ℓ to the battery B mediated by the increase in entropy of A_k. Specifically, for each iteration $\ell \in [M]$, the agent:

- sets the energy gap to $\nu(\ell, \epsilon_k)$ (see (10)) and selects a reservoir qubit in the thermal state $\gamma_\beta(\nu(\ell, \epsilon_k))$;
- applies the energy-conserving unitary swap operation

$$V_{\psi_k,\ell} = \sum_{i,j} |i\rangle\langle j|_{A_k} \otimes |j\rangle\langle i|_{R_\ell} \otimes \Gamma^B_{(i-j)\nu(\ell,\epsilon_k)}\,; \quad (9)$$

- discards the reservoir qubit.

The unitary $V_{\psi_k,\ell}$ in Eq. (9) swaps the states of the unknown system A_k and the thermal reservoir R_ℓ. This operation induces an energy exchange determined by the

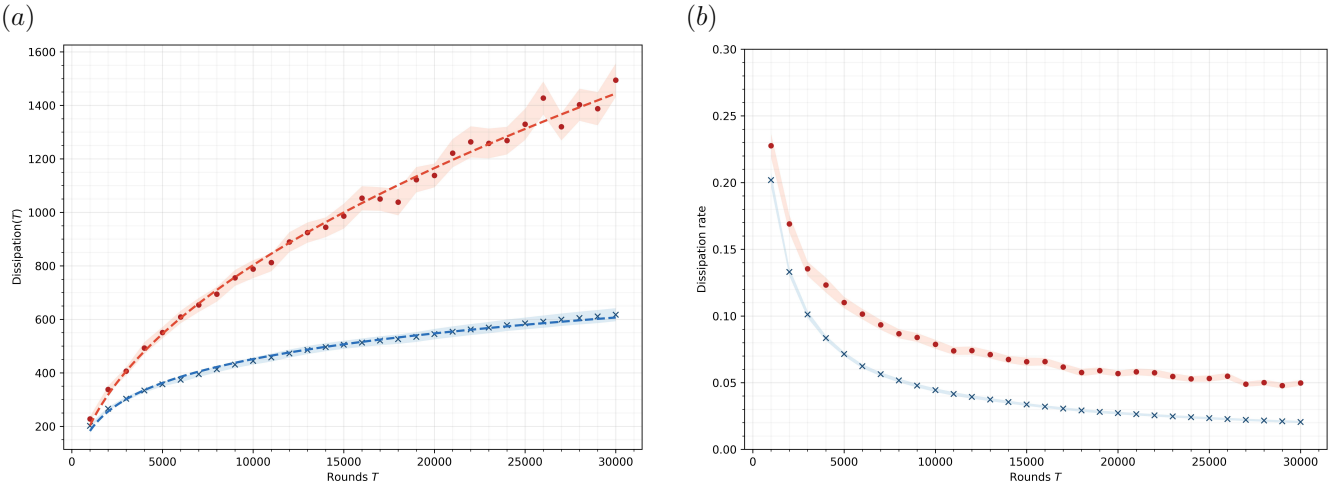


Figure 4. **Performance scaling of the adaptive work extraction protocol.** Cumulative dissipation (a) and dissipation rate (b) versus the number of rounds T (rate = average dissipation per copy). Blue: our adaptive protocol. Red: a tomography-first baseline that uses $O(1/\sqrt{T})$ of the available copies for learning and then applies the state-aware extraction protocol on the remainder. For our protocol we probe four directions per update and repeat each direction for $t = 5$ trials, hence the estimator is updated every $4t = 20$ rounds. While t is mildly horizon-dependent in the theory (to guarantee high-probability confidence bounds), empirically $t = 5$ already yields the same scaling, making the implementation effectively horizon-independent. Points are averages over 50 independent trials with a random pure state per trial; for the tomography-first baseline, each T is simulated as an independent experiment (since the strategy depends explicitly on the horizon). Dashed lines show least-squares fits: polylogarithmic $a(\log_{10} T)^2 + b$ for our protocol ($R^2 = 0.9958$, $a = 38.4 \pm 0.47$, $b = -163 \pm 8$) and square-root $a\sqrt{T} + b$ for tomography-first ($R^2 = 0.9930$, $a = 8.76 \pm 0.14$, $b = -74.0 \pm 17$). Shaded regions indicate 95% confidence intervals; regression statistics were computed with `scipy.stats.linregress`.

energy gap $\nu(\ell, \epsilon_k)$ between the two systems. The energy difference is transferred to the battery conserving the energy of the combined system, i.e., $[H_A + H_B + H_R, V_{\psi_k,\ell}] = 0$.

To maximize the extractable non-equilibrium free energy from system A_k, it is essential to minimize dissipation during the energy transfer to the battery. This can be achieved by ensuring that the process is quasi-static—that is, changes occur slowly enough. This is achieved in the limit of repetitions $M \to \infty$. Under such conditions, in each iteration ℓ except for the first one, the qubit in A_k remains close to the thermal state R_ℓ of the reservoir's Hamiltonian, thereby suppressing heat flow during the interaction.

However, the process is not entirely quasi-static—it is quasi-static for all but the first repetition. During the first repetition, the system qubit $|0\rangle\langle 0|$ is swapped with the reservoir qubit $(1-\epsilon_k)|0\rangle\langle 0| + \epsilon_k|1\rangle\langle 1|$. This swap generates entropy $\Delta S = -\epsilon_k \ln \epsilon_k - (1-\epsilon_k)\ln(1-\epsilon_k)$ and yields work extraction $\Delta U = -k_B T \epsilon_k \nu(1, \epsilon_k) = -k_B T \epsilon_k \ln \frac{1-\epsilon_k}{\epsilon_k}$. Combining both, the resulting change in free energy is $\Delta U - k_B T \Delta S = \ln(1-\epsilon_k)$, which precisely accounts for the logarithmic term observed in $w_{k,0}$ (Eq. (4)). A similar argument holds for $w_{k,1}$.

Conversely, if the process is carried out rapidly, i.e., the system state deviates from the reservoir's thermal state, the system will appear out of thermal equilibrium with the reservoir, resulting in heat exchange, which contributes to irreversible entropy production during the protocol [26, 27]. For practical implementation, it suffices to choose $M = DN^2/\ln^3(N)$ for some constant $D > 0$, ensuring that the total error probability does not exceed that of the tomography algorithm; see Supplemental Material [1, Section IIIA].

At each repetition $\ell \in [M]$, a slight mismatch between the energy levels of the system A_k and the reservoir R_ℓ is introduced by varying the energy gap. The gap is parameterized as follows

$$\nu(\ell, \epsilon_k) = k_B T \ln \left(\frac{1 - \frac{\ell}{2M} - \left(1 - \frac{\ell}{M}\right)\epsilon_k}{\frac{\ell}{2M} + \left(1 - \frac{\ell}{M}\right)\epsilon_k} \right), \qquad (10)$$

where $\epsilon_k \in [0, 1]$ controls the mismatch. Further details on this gap parametrization can be found in Supplementary Material [1, Section IIA]. The full interaction of systems A_k, B, and R_ℓ through the unitary $V_{\psi_k,\ell}$ is illustrated in Fig. 5.

To compute the total extracted work, the amount of energy transferred to the battery at each repetition $\ell \in [M]$ can be evaluated. This is done by considering the two initial input states that can trigger a swap operation, namely $|0\rangle_A |1\rangle_R$ and $|1\rangle_A |0\rangle_R$. When the unitary Eq. (9) is applied repeatedly, there is a certain probability that the states of the system and reservoir qubits are swapped, thereby enabling energy exchange with the battery. In the quasi-static limit $M \to \infty$, each of these initial states contributes a fixed amount of energy $\Delta W_k = \mu_{k+1} - \mu_k$ to the battery. The possible values of ΔW_k is shown in

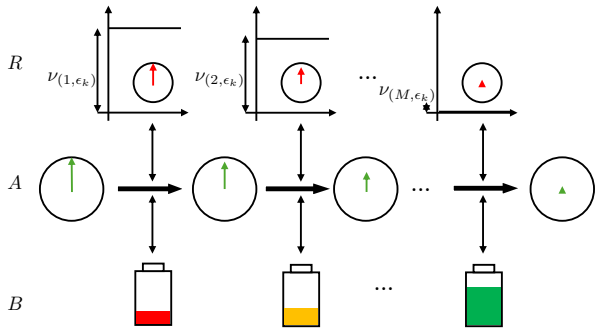


Figure 5. Illustration of the iterations of the thermal operation in the full system ABR, where arrows represent Bloch vectors of states, showing that the system qubit becomes more and more mixed as the process goes. The energy gaps $\{\nu_{k,i}\}_i$ of successive reservoir Hamiltonians form a strictly decreasing sequence, making the successive thermal states more mixed. At each step, a new reservoir qubit is supplied from the reservoir and swapped with the system qubit; the energy from the reservoir will flow into the battery. At the end of the process, the qubit in system A is the thermal state.

Eq. (4). See Supplementary Material [1, Section III] for rigorous derivation as well as rate of convergence of work values with respect to M.

Learning algorithm

The exact update rule for ψ_k at each time step $k \in \{1, 2, \ldots, N\}$ will now be reviewed. Each ψ_k is updated adaptively based on the previous binary measurement outcomes $\{r_l\}_{l=1}^{k-1}$ and measured directions $\{\psi_l\}_{l=1}^{k-1}$ such that they are distributed according to Born's rule i.e., $\Pr(r_l = 1) = 1 - |\langle\psi|\psi_l\rangle|^2$. The update rule is formulated in the Bloch vector representation. The algorithm proceeds in $K^* \in \mathbb{N}$ stages, where at each stage $k^* \in [K^*]$, it performs t measurements along the Bloch vector directions $\{a_{k^*,i}\}_{i=1}^4 \subset \mathbb{R}^3$ which will be defined below. The total number of rounds is then $N = 4tK^*$. The sampled rewards are used to build the t weighted least-square estimators

$$\widetilde{\theta}^{\mathrm{w}}_{k^*,j} = V_{k^*}^{-1} \sum_{s=1}^{k^*} \frac{1}{\hat{\sigma}_s^2} \sum_{i=1}^{4} a_{s,i}(2r_{s,i,j} - 1), \tag{11}$$

where $r_{s,i,j} \in \{0, 1\}$ is the outcome obtained from the measurement direction $a_{s,i}$ for $s \in [K^*]$ and $j \in [t]$. Here, $\hat{\sigma}_s^2$ is an upper bound on the variance of $r_{s,i,j}$ and $V^{\mathrm{w}}_{k^*}$ is the design matrix defined as

$$V^{\mathrm{w}}_{k^*} = \mathbb{1} + \sum_{s=1}^{k^*} \frac{1}{\hat{\sigma}_s^2} \sum_{i=1}^{4} a_{s,i}(a_{s,i})^{\intercal}. \tag{12}$$

As studied in [21, 22], the weighted least-squares estimator boosts learning along directions with low variance, which are precisely those close to ψ, since the variance satisfies $\mathrm{Var}(r_k) \leq 1 - |\langle\psi_k|\psi\rangle|^2$. For the specific update used by the algorithm, it can be shown that choosing $\hat{\sigma}_l^2 \sim 1/\sqrt{\lambda_{\max}(V_{k^*-1})}$, (where $\lambda_{\max}(\cdot)$ denotes the maximum eigenvalue) provides an upper bound on the outcome variances and ensures well-defined least-squares estimators. However, a single estimator is generally insufficient to guarantee proximity to the unknown state with high probability. To address this, the algorithm constructs multiple estimators and combines them into a normalized weighted median-of-means $\theta^{\mathrm{wMoM}}_{k^*}$, which satisfies the desired probabilistic guarantees to be close to ψ. Details on the construction of $\theta^{\mathrm{wMoM}}_{k^*}$ are provided in Supplementary Material [1, Section I]. Finally, it outputs the measurement directions in the Bloch sphere representation $\{a_{k^*+1,i}\}_{i=1,2,3,4}$, in the form of

$$\tilde{a}_{k^*+1,i} = \theta^{\mathrm{wMoM}}_{k^*} - \frac{(-1)^i}{\sqrt{\lambda_{\min}(V_{k^*})}} v_{k^*,\lceil \frac{i}{2} \rceil}, \tag{13}$$

where $\lambda_{\min}(V_{k^*})$ is the minimum eigenvalue of V_{k^*} and $\{v_{k^*,0}, v_{k^*,1}\}$ are the two eigenvectors of V_{k^*} with smallest eigenvalues. The normalized Bloch vectors $a_{k^*+1,i}$ are used to update the measurement directions ψ_k for round k in stage $k^* + 1$. More details can be found in Supplementary Material [1, Section I].

[1] See Supplemental Material at URL-will-be-inserted-by-publisher for detailed proofs and formal definitions.

[2] J. Åberg. *"Truly work-like work extraction via a single-shot analysis"*. Nature communications **4**(1): 1925 (2013).

[3] S. Agrawal and N. Goyal. *"Analysis of thompson sampling for the multi-armed bandit problem"*. In *Conference on learning theory*, pages 39–1, (2012).

[4] A. E. Allahverdyan, R. Balian, and T. M. Nieuwenhuizen. *"Maximal work extraction from finite quantum systems"*. Europhysics Letters **67**(4): 565 (2004).

[5] P. Auer, N. Cesa-Bianchi, and P. Fischer. *"Finite-time analysis of the multiarmed bandit problem"*. Machine learning **47**: 235–256, (2002).

[6] P. Auer, N. Cesa-Bianchi, Y. Freund, and R. E. Schapire. *"The Nonstochastic Multiarmed Bandit Problem"*. SIAM J. Comput. **32**(1): 48–77 (2003).

[7] F. Brandao, M. Horodecki, N. Ng, J. Oppenheim, and S. Wehner. *"The second laws of quantum thermodynamics"*. Proceedings of the National Academy of Sciences **112**(11): 3275–3279, (2015).

[8] F. G. Brandao, M. Horodecki, J. Oppenheim, J. M. Renes, and R. W. Spekkens. *"Resource theory of quantum states out of thermal equilibrium"*. Physical review letters **111**(25): 250404 (2013).

[9] S. Chen, B. Huang, J. Li, A. Liu, and M. Sellke. *"When Does Adaptivity Help for Quantum State Learning?"*. In *2023 IEEE 64th Annual Symposium on Foundations of*

Computer Science (FOCS), pages 391–404, (2023).

[10] V. Dani, T. P. Hayes, and S. M. Kakade. *"Stochastic Linear Optimization under Bandit Feedback."*. In *Proceedings of the 21st Conference on Learning Theory*, volume 2, page 3, (2008).

[11] C. Elouard and A. N. Jordan. *"Efficient quantum measurement engines"*. Physical review letters **120**(26): 260601 (2018).

[12] S. T. Flammia and R. O'Donnell. *"Quantum chi-squared tomography and mutual information testing"*. Quantum **8**: 1381 (2024).

[13] J. Haah, A. W. Harrow, Z. Ji, X. Wu, and N. Yu. *"Sample-Optimal Tomography of Quantum States"*. IEEE Transactions on Information Theory **63**(9): 5628–5641 (2017).

[14] M. Horodecki and J. Oppenheim. *"Fundamental limitations for quantum and nanoscale thermodynamics"*. Nature communications **4**(1): 2059, (2013).

[15] R. C. Huang, P. M. Riechers, M. Gu, and V. Narasimhachar. *"Engines for predictive work extraction from memoryful quantum stochastic processes"*. Quantum **7**: 1203 (2023).

[16] D. Šafránek, D. Rosa, and F. C. Binder. *"Work Extraction from Unknown Quantum Sources"*. Phys. Rev. Lett. **130**: 210401 (2023).

[17] A. Kolchinsky and D. H. Wolpert. *"Dependence of dissipation on the initial distribution over states"*. Journal of Statistical Mechanics: Theory and Experiment **2017**(8): 083202, (2017).

[18] T. Lai and H. Robbins. *"Asymptotically efficient adaptive allocation rules"*. Advances in Applied Mathematics **6**(1): 4–22 (1985).

[19] S. Lloyd. *"Thermodynamics + Natural Selection = Bayesian Inference"*. arXiv preprint arXiv:2511.17641 , (2025).

[20] M. Lostaglio, D. Jennings, and T. Rudolph. *"Description of quantum coherence in thermodynamic processes requires constraints beyond free energy"*. Nature communications **6**(1): 6383 (2015).

[21] J. Lumbreras, M. Terekhov, and M. Tomamichel. *"Learning pure quantum states (almost) without regret"*. arXiv preprint arXiv:2406.18370 , (2024).

[22] J. Lumbreras and M. Tomamichel. *"Linear bandits with polylogarithmic minimax regret"*. In *Proceedings of Thirty Seventh Conference on Learning Theory*, volume 247 of *Proceedings of Machine Learning Research*, pages 3644–3682, (2024).

[23] P. M. Riechers and M. Gu. *"Initial-state dependence of thermodynamic dissipation for any quantum process"*. Phys. Rev. E **103**: 042145 (2021).

[24] P. Skrzypczyk, A. J. Short, and S. Popescu. *"Work extraction and thermodynamics for individual quantum systems"*. Nature communications **5**(1): 4185 (2014).

[25] L. Szilard. *"Über die Entropieverminderung in einem thermodynamischen System bei Eingriffen intelligenter Wesen"*. Zeitschrift für Physik **53**(11): 840–856 (1929).

[26] P. Taranto, F. Bakhshinezhad, A. Bluhm, R. Silva, N. Friis, M. P. Lock, G. Vitagliano, F. C. Binder, T. Debarba, E. Schwarzhans, F. Clivaz, and M. Huber. *"Landauer Versus Nernst: What is the True Cost of Cooling a Quantum System?"*. PRX Quantum **4**: 010332 (2023).

[27] T. Van Vu and K. Saito. *"Finite-Time Quantum Landauer Principle and Quantum Coherence"*. Phys. Rev. Lett. **128**: 010602 (2022).

[28] K. Watanabe and R. Takagi. *"Black box work extraction and composite hypothesis testing"*. Physical Review Letters **133**(25): 250401 (2024).

[29] K. Watanabe and R. Takagi. *"Universal work extraction in quantum thermodynamics"*. arXiv preprint arXiv:2504.12373 , (2025).

ACKNOWLEDGEMENTS

JL thanks Jan Seyfried for discussions about quantum state tomography with quantum relative entropy loss. This work is supported by the National Research Foundation of Singapore through the NRF Investigatorship Program (Award No. NRF-NRFI09-0010), the National Quantum Office, hosted in A*STAR, under its Centre for Quantum Technologies Funding Initiative (S24Q2d0009), grant FQXi-RFP-IPW-1903 ("Are quantum agents more energetically efficient at making predictions?") from the Foundational Questions Institute (FQxI) and Fetzer Franklin Fund (a donor-advised fund of Silicon Valley Community Foundation), the Singapore Ministry of Education Tier 1 Grant RT4/23, MT, YH and JL are supported by the National Research Foundation, Singapore and A*STAR under its CQT Bridging Grant and its Quantum Engineering Programme (NRF2021-QEP2-02-P05). MT is supported by the NRF Investigatorship award (NRF-NRFI10-2024-0006).

AUTHOR CONTRIBUTIONS

JL and RCH contributed equally to this work. JL, RCH, YH, MG and MT wrote the manuscript. JL, RCH, and YH derived the technical results. JL, RCH, and MG came up with the initial idea. All authors contribution to discussion. MH and MT directed the project

CODE AVAILABILITY

The source code for the numerical simulations and data analysis presented in this work is available at the following GitHub repository: `https://github.com/josep-lumbreras/agnostic-work-extraction/tree/main`.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- SI.pdf