Referee Report

Hayden–Preskill recovery at finite temperature on a quantum processor: dynamics and initial state from the SYK model

Recommendation: Major revision. The SWAP-injected decoder is a useful and potentially publishable variation of finite-temperature Hayden–Preskill recovery, and the combination of an analytic treatment, SYK numerics, and a superconducting-processor demonstration is appealing. In its present form, however, the manuscript has not yet established that the logical-qubit SWAP used after parity reduction is a physically defined local operation in the SYK model; the central late-time formulas rely on an isotropy assumption that does not treat the Hamiltonian-commuting operator sector and is only narrowly tested; and the hardware mitigation uses an unvalidated transfer of a one-parameter noise model from a different, SWAP-removed circuit. The disorder averaging of a conditional ratio and the operational meaning of "successful recovery" at very small postselection probability also require quantitative treatment. These issues affect the main theoretical and experimental conclusions and should be resolved before publication in PRD or JHEP.

Synopsis and overall assessment

The paper modifies the finite-temperature Hayden–Preskill/Yoshida–Kitaev setup by replacing the usual appended diary with a SWAP injection. A diary qubit A, initially entangled with a reference R, is exchanged with a chosen subsystem b of a black-hole register B. The pre-existing degree of freedom expelled into A is then counted as radiation. The black-hole register and its early-radiation copy are initially in the thermofield-double state of equation (2), and the same Hamiltonian that defines this state generates the forward evolution in equation (6). On the decoder side, a mirror SWAP and the conjugated evolution are applied. Successful Bell projections on AA' and DD' define the probability in equation (10), while the Bell overlap on RR' defines the conditional fidelity in equation (12).

This construction changes the Hilbert-space relation from $d_A d_B = d_C d_D$ to $d_B = d_C d_D$. It also gives the decoder the expelled register A and requires an additional Bell projection on AA'. The paper derives exact block formulas for the probability and joint probability in Appendices A and B. At late times, it assumes uniform spreading of the traceless block operators, as summarized by equation (16), and obtains equations (17)–(18) for the average postselection probability in terms of the global thermal factor η_β and the injection-dependent factor ξ_β. The analogous calculation gives equations (22)–(23) for the joint probability and conditional fidelity. Equations (25)–(26) supply the initial values, and equation (27) introduces a thermally regularized OTOC used as a qualitative operator-growth diagnostic.

The numerical study uses the even-parity sector of an $N = 16$ SYK model, represented as seven qubits, and compares large and small late-radiation subsystems. It finds the expected tradeoff: after scrambling the postselection probability falls while the conditional fidelity rises; low temperature mainly suppresses the probability when D is large, whereas a small D also reduces the fidelity. The manuscript then studies binary sparse SYK ensembles and observes unstable oscillatory behavior under very strong sparsification.

For the hardware demonstration, the authors select the $N = 8$, $K = 10$ binary sparse Hamiltonian displayed in equation (32), prepare its TFD state with a variational circuit, and implement a single-step first-order product formula on ten qubits of `ibm_marrakesh`. Equations (35)–(36) extract the probability and conditional fidelity from counts. A circuit with the two SWAP injections removed is then used to define the multiplicative correction in equations (37)–(39) and the conditional depolarizing correction in equations (40)–(43). Figure 11 shows that the raw and corrected data retain the qualitative decrease of P and increase of F at two temperatures.

The topic is timely and the paper is clearly structured. Its most distinct contribution relative to the finite-temperature HP literature, Hamiltonian recovery studies, SYK/Petz-map analyses, and earlier scrambling experiments is the fixed-Hilbert-space SWAP channel and the associated factor ξ_β. This is more than a notational change, but it is also a different information-processing task from appending a diary. The paper will be significant if that distinction is made precise, the SYK embedding is defined independently of an arbitrary logical encoding, and the claimed late-time and hardware conclusions are supported with controlled errors.

Logical structure and principal assumptions

The conclusions rest on the following chain.

- The TFD state in equation (2), the forward evolution, and the mirror evolution are related by a fixed anti-linear identification of B with B'. The SWAPs in equation (8) act on matched tensor factors $b \subset B$ and $b' \subset B'$.
- Treating the expelled register as radiation and postselecting on AA' as well as DD' defines a new recovery channel. Large conditional Bell fidelity is interpreted as recovery, provided the success probability is not prohibitively small.
- After projection to one fermion-parity sector, that sector is assumed to possess a physically meaningful tensor-product qubit structure. A SWAP with one of these logical qubits is then treated as injection into a local SYK degree of freedom, and the same logical factorization is used to define C, D, and the OTOC.
- At late times, disorder averaging is assumed to distribute the squared coefficients of each traceless block uniformly over all $d_B^2 - 1$ traceless operator directions, as in equation (71). The identity direction is the only time-independent direction treated separately.
- The mean conditional fidelity is approximated by the ratio of mean joint and marginal probabilities. This requires small sample-to-sample fluctuations and weak correlations in precisely the regime where the denominator can be small.
- Hardware noise in the target and SWAP-removed reference circuits is assumed to have the same multiplicative attenuation for P and the same one-parameter depolarizing action on the postselected RR' state.

Several of these assumptions may be reasonable for the selected small instances, but they are not consequences of scrambling alone. They should be stated as hypotheses, tested quantitatively, and separated from exact identities.

Major comments

1. **The SWAP construction defines a different recovery channel; its relation to conventional Hayden–Preskill recovery needs a precise theorem or a more limited interpretation.**

   Equation (5) does avoid choosing an extension H_{AB} of H_B, but it does so by changing both the injection and the decoder resources. The diary replaces a degree of freedom of B, the displaced degree of freedom is supplied to Bob as A, and the decoder postselects on AA'. Dimension matching in Table I does not make this channel equivalent to the appended HP channel, and the Haar decoupling bound in equation (4) is explicitly not shown to apply. Thus, the claim that the ambiguity of the conventional construction has been "resolved" is too strong: a new, well-defined task has been proposed instead.

The revised paper should formulate the two protocols as quantum channels or isometries, list their input and accessible output registers, and state a recovery criterion for the SWAP channel. Ideally, the authors should prove a decoupling or entanglement-fidelity statement for this channel, including its dependence on d_A, d_D, and the thermal state. At minimum, they should explain exactly what operational statement follows from equations (10)–(12) and what does not follow from the qualitative trace-norm plot. A fair numerical comparison also needs a controlled appended Hamiltonian: the dashed baseline in Figure 4 averages the TFD preparation and scrambling over independent disorder realizations, so it changes Hamiltonian correlations at the same time as it changes the injection rule. The protocol comparison should hold those choices fixed or analyze their effect separately.

This clarification is also important for novelty. Prior work cited by the authors already treats finite-temperature decoding, Hamiltonian HP recovery, SYK recovery maps, qubit-transport evaporation models, and scrambling experiments. The new result is the SWAP channel and its ξ_β term, not a general resolution of finite-temperature HP dynamics. The introduction and conclusion should make that boundary explicit.

2. **A qubit in the parity-reduced SYK Hilbert space is not yet a specified local SYK subsystem.**

   Section III.1 projects the N-Majorana theory to its even-parity sector and then identifies this $2^{N/2-1}$-dimensional space with $N/2-1$ computational qubits. Equation (15) defines physical Jordan–Wigner qubits before projection, but the projected sector has no canonical tensor factorization into the logical qubits subsequently called b, C, and D. Changing the isometry from the even sector to $(\mathbb{C}^2)^{\otimes(N/2-1)}$ changes the logical SWAP, the partial trace in ξ_β, and the subsystem OTOC. Moreover, an ordinary SWAP between the diary and one pre-projection occupation qubit generally changes the fermion parity of B; the corresponding parity-preserving logical operation is typically nonlocal in the physical Majoranas.

   The authors should give the encoding isometry explicitly and write the logical SWAP as an operator on the full fermionic Hilbert space. They should state its Majorana support and circuit cost, explain in what sense it is a local injection, and test the dependence of the conclusions on the encoding and injected logical factor. A complementary calculation using both parity sectors or a genuinely parity-preserving fermionic-mode injection would be particularly useful. If the intention is only to use the projected SYK matrix as an abstract chaotic qubit Hamiltonian, that is legitimate, but the paper must stop identifying the operation with a local qubit of the original SYK system. This point bears directly on the physical interpretation of ξ_β and on the claimed black-hole analogy.

3. **The uniform-spreading estimate omits conserved operator components and is not validated at the scope required by equations (17)–(23).**

   The exact reduction to block norms in equations (55)–(70) is useful. The nontrivial step is equation (71), which assigns equal disorder-averaged weight to every traceless operator direction. Removing only the identity does not remove the full commutant of H. For a fixed Hamiltonian, every energy-diagonal component of $\mathcal{N}_0^{(ar)}$ is stationary under $U(t)\mathcal{N}_0^{(ar)}U^\dagger(t)$, even when the spectrum is nondegenerate. At finite temperature the blocks themselves contain $\rho_\beta^{1/2} = f(H)$, so their correlations with this conserved sector are not obviously erased by disorder averaging. "No residual degeneracy" therefore does not justify isotropy over all $d_B^2 - 1$ directions.

   Appendix C tests only the surviving $\alpha = 0$ coefficients for the single block $(a, r) = (0, 0)$, at $T = 1$, one partition, one system size, and two times. Figure 12 is a useful visualization, but it neither tests the other blocks entering P nor the distinct operator $\mathcal{S}$ entering Q; it also gives no numerical error norm or size scaling. The authors should decompose the blocks into the Hamiltonian commutant and its orthogonal complement, derive the exact infinite-time or diagonal-ensemble contribution, and state when the remaining part may be modeled by a Haar adjoint second moment. They should then report a quantitative residual for equation (16), across all (a, r), for the Q block, across the temperatures and C/D partitions used in Figure 4, and for increasing N. Error bars on equations (17) and (23), rather than markers alone, are needed. If the simple formulas remain empirical finite-size fits, the text should call them that rather than general late-time saturation values.

4. **The disorder average of a conditional ratio and the claim of successful low-temperature recovery need a statistical and operational analysis.**

   For each disorder realization, equation (21) gives $F = Q/P$. Consequently, $\overline{F} = \overline{Q/P}$ is not generally $\overline{Q}/\overline{P}$. Equation (23), or equivalently equation (100), makes this replacement under a small-fluctuation assumption. The issue is most acute at low temperature, where Figure 4 shows that P can become very small and rare realizations or covariance between P and Q can dominate a ratio. Twenty disorder samples are not enough to assume this away without displaying the distribution.

   Please report, for every temperature and partition, both $\overline{Q/P}$ and $\overline{Q}/\overline{P}$, their difference, sample variances, covariance, and confidence intervals. The average convention used for each plotted fidelity and each analytic marker must be stated. Similar uncertainty information should be added to the disorder-averaged Figures 3–8.

   The operational claim also needs a success-weighted quantity. Conditional fidelity close to unity can coexist with an exponentially or thermally suppressed success probability; in that case recovery is accurate only on a very rare branch. The authors should show $Q = PF$, the expected number of trials $1/P$, and the scaling of P with N, β, and d_D, and then define quantitatively what they mean by "non-negligible." Figure 3 does not show decoupling at the lowest temperatures, while Figure 4(a) shows high conditional fidelity there; this is not a contradiction, but it does mean that the two notions of recovery must not be conflated. The abstract's statement that both P and F are "proportional to temperature" is not supported by either the formulas or the plots and should be replaced by a precise statement of temperature dependence.

5. **The reference-circuit correction is a model-dependent calibration, not yet validated error mitigation.**

   Equations (38)–(39) assume that noise multiplies the target and SWAP-removed reference probabilities by the same factor. The circuits do not have the same gates, and their state trajectories and accepted postselection ensembles differ. Equations (40)–(43) make a stronger assumption: noise before Bell postselection is replaced by a depolarizing channel on the already conditioned RR' state, with the same $\lambda(t)$ for the target and reference. In general, conditioning is nonlinear; errors can change which shots are accepted and need not reduce to a channel on RR' with a transferable scalar parameter. Agreement after using the known noiseless target curve is therefore not by itself independent evidence that the correction is accurate.

   The authors should show the raw reference data, $M_P(t)$, and $\lambda(t)$, with propagated uncertainty. The transfer assumption should be tested in backend-calibrated noise simulations and, preferably, against an independent mitigation procedure such as several noise-scaled circuits. A withheld set of times or Hamiltonian instances would distinguish calibration from post hoc agreement. The analysis should verify physicality ($0 \leq P^{\mathrm{mit}}, F^{\mathrm{mit}} \leq 1$) and account for the correlations created by reusing the reference estimates in ratios. Until this is done, Figure 11 supports the qualitative trend already visible in the raw data, but the claims that the mitigation "restores" either observable should be softened.

6. **The hardware demonstration needs independent validation and reproducibility.**

   The $N = 8$, $K = 10$ instance in equation (32) is selected because it is said to be chaotic and because one first-order Trotter step reproduces P and F over $0 \leq t \leq 6$. Figure 9 compares those same target observables, so it does not exclude instance selection tailored to the desired signal or show that the implemented channel approximates e^{-iHt} more generally. The paper should display the level-spacing or spectral-form- factor diagnostic for this exact instance, state how many instances were screened and by what criterion, and quantify product-formula error using a state, unitary, or set of independent observables not used in the selection. A convergence comparison with two or more Trotter steps in noiseless simulation is also needed.

For reproducibility, please provide the parity-reduced Pauli Hamiltonian, the ordered product formula, the logical-to-physical qubit layout, the transpilation settings and seeds, the full VQA ansatz and optimized parameters at both temperatures, the processor calibration snapshot, raw counts for target and reference circuits, disorder seeds, and analysis code. The five experimental repetitions should be accompanied by shot-noise intervals and a description of whether they were interleaved in time. These details are necessary to assess whether the observed temperature difference exceeds device drift and whether the experiment can be reproduced.

Minor comments

1. Equation (2) uses $|n^*\rangle$ on B' but says $H_{B'} = H_B$, while equation (8) later uses the complex-conjugated evolution. For a complex SYK matrix these statements require a chosen basis and an anti-linear identification. Please state explicitly whether the matrix on B' is H_B^* and verify the TFD invariance used in equation (49).

2. The claim after equation (3) that the extension to $N_A > 1$ is straightforward should specify that b must then be a subsystem with dimension $d_b = d_A$, rather than the single qubit used throughout the derivation.

3. Equation (32) should be reconciled with the factor $i^{q/2} = -1$ in equations (13) and (28). If the global sign has been absorbed into the displayed binary couplings, say so explicitly, since the sign of a fixed Hamiltonian matters for the finite-temperature state.

4. The time units in all plots should be stated in terms of J or $\mathcal{J}$. The choices $J = \sqrt{2}$ and $\mathcal{J} = 1/\sqrt{5}$ otherwise make comparisons between dense, sparse, and hardware figures ambiguous.

5. Figure 5 shows that mismatching b and b' suppresses recovery for one pair of logical sites. This is not a verification of a general mirror condition. Please show all site pairs, or state the narrower numerical observation.

6. The OTOC in equation (27) is described appropriately as qualitative in Section III.4, but the conclusion says that recovery "follows" operator growth. The paper should report a quantitative distance or correlation between the normalized curves and should not infer growth of F from an agreement tested only against P.

7. For the probability and fidelity extracted in equations (35)–(36), give binomial or bootstrap intervals in addition to the standard deviation across five runs. The denominator of equation (36) becomes small at low temperature, so symmetric error bars can be misleading.

8. The manuscript alternates between "saturation," a long-time value, and a finite scrambling window. At finite N a unitary system has recurrences, and Figure 8(b) shows persistent oscillations. Define whether each quoted value is a disorder average at a fixed time, a time-window average, or an infinite-time average.

9. The paper would benefit from a final edit of the source. Remove large commented-out passages and obsolete figures, define every averaging bar at first use, and distinguish consistently among exact identities, analytic estimates, and empirical observations.