Random Unitaries, Robustness, and Complexity of Entanglement

J. Odavić, G. Torre, N. Mijić, D. Davidović, F. Franchini, and S. M. Giampaolo

Ruđer Bošković Institute, Bijenička cesta 54, 10000 Zagreb, Croatia

September 13, 2023

It is widely accepted that the dynamic of entanglement in the presence of a generic circuit can be predicted by the knowledge of the statistical properties of the entanglement spectrum. We tested this assumption by applying a Metropolis-like entanglement cooling algorithm generated by different sets of local gates, on states sharing the same statistic. We employ the ground states of a unique model, namely the one-dimensional Ising chain with a transverse field, but belonging to different macroscopic phases such as the paramagnetic, the magnetically ordered, and the topological frustrated ones. Quite surprisingly, we observe that the entanglement dynamics are strongly dependent not just on the different sets of gates but also on the phase, indicating that different phases can possess different types of entanglement (which we characterize as purely local, GHZ-like, and W-state-like) with different degree of resilience against the cooling process. Moreover, in some circumstances, we observe a scrambling effect by the algorithm that produces a Wigner-Dyson entanglement spectrum statistics on a state that does not obey a volume law for the entanglement entropy.

1 Introduction

Entanglement is the most distinctive mark of quantum mechanics [1, 2] and an essential resource for many technological devices currently in development [3–5]. Therefore, it is easy to understand the growing interest in characterizing entanglement, especially in quantum many-body systems, as they represent the platform on top of which such devices are to be implemented. Characterization of entanglement is paramount in this context, as it is not just the sheer amount of entanglement which plays a crucial role in quantum applications. This is because some entangled states can be described efficiently also through classical resources. To provide an example, it is known that quantum circuits starting from factorized states diagonal in the computational basis and made by gates from Clifford’s group [6], can be efficiently simulated on a classical computer despite the amount of entanglement of the output state [7, 8]. The resulting states are known as “stabilizer states”. However, adding gates outside Clifford’s group, such as T-gates, makes it impossible to simulate the circuits efficiently on a classic computer [9–11]. The difference between the two cases can be characterized by looking at the statistical properties of the entanglement spectra. While in the second case, the output generally develops a Wigner-Dyson distribution in the entanglement spacing statistics, in the first one, the Poisson distribution is always obtained [9, 10, 12]. This difference plays a key role in the theory of quantum information, since only circuits doped with T-gates are capable of universal computation, as they can reach any state in the Hilbert space independently of the initial state [13, 15].

An alternative way of analyzing this difference is to use the concept of stochastic irreversibility (or robustness) of entanglement. The idea behind this approach is to obtain information on the entanglement structure of a state by observing its evolution under the action of an entanglement cooling algorithm. At its core, this algorithm is the usual Metropolis Monte Carlo protocol, with the cost function played not by the energy but by an entanglement measure. A transformation, chosen from a predefined set, is applied to the initial state, and the entanglement value of the new state is determined. The new state is accepted (retained) if the cost function has decreased and with a certain probability otherwise. While states with a Poissonian statistic of the entanglement spectrum are not very resistant to this approach, i.e. after a few steps, the total amount of entanglement in the system tends to vanish, the cooling algorithm proves to...
be almost ineffective in states where this statistic follows a Wigner-Dyson distribution [9, 10]. The different robustness of entanglement against a cooling algorithm has been linked to a concept of “complexity”, coherently with the usual picture that a Wigner-Dyson distribution, stemming from the existence of strong correlations between the entanglement eigenvalues, indicate a higher complexity as well as higher robustness [14, 15]. This difference in robustness has proved to be extremely useful for distinguishing between the different dynamical phases present in quantum many-body systems [14] and falls within the recent interest of the quantum many-body community in random quantum circuits [16, 17].

In the present paper, we test how strong is the relationship between stochastic irreversibility and the entanglement spectrum statistic, by placing a particular emphasis on the choice of the initial states. In previous works (e.g., see Ref. [14]), the states to which the cooling algorithm was applied were randomly generated. On the contrary, our starting states are the ground states of the one-dimensional quantum Ising model in its different macroscopic phases. As we will show, the entanglement spectrum of all these states follows a Poisson distribution for the level-spacing statistics, once some degeneracies are properly dealt with. Therefore, the action of the cooling algorithm could be expected to be independent of the initial state phase. Instead, quite surprisingly, our results show a very different picture.

The most peculiar behavior we observe is associated with the ground states of models in a topologically frustrated phase [20–25]. They are obtained by imposing frustrated boundary conditions (that is, periodic boundary conditions with an odd number of sites) on systems with antiferromagnetic interactions. The resulting ground states can be largely characterized as a linear superposition of single-dressed kink states [26, 27] (topological solitons) and this representation allows us to describe, both qualitatively and quantitatively, the behavior of various physical quantities even in the presence of integrability breaking terms [26–29]. In these states, after the cooling algorithm, we observe a stabilization of the entanglement to a reduced baseline value (finite, but strongly dependent on the size of the system), and any further reduction appears to be statistically unlikely.

In the present work, we consider two sets of operations within the cooling algorithm. They are made of one- and two-body gates, with the latter acting only on neighboring spins. In this sense, our algorithm can be considered made of local gates. In the first set, we include only operations preserving the parity symmetry of the Hamiltonian which, as a consequence, cannot explore every state in the Hilbert space, thus generating a violation of ergodicity. On the contrary, in the second, the set of operations is extended to a complete set to ensure, at least in principle, access to any state.

While the end values of the entanglement obtained by the cooling algorithm starting from states with topological frustration are not qualitatively influenced by the choice of employed gates, this is not true for the ground states of the paramagnetic phase. On them, the action of the cooling algorithm is practically negligible if the set of operations is limited to the one in which elements commute with the parity of the Hamiltonian, while their entanglement is quickly reduced if the larger set of gates is considered. This dependence on the gate set also allows distinguishing the paramagnetic phase from the ferromagnetic one, since on the latter the cooling algorithm is unable to destroy most of their entanglement, regardless of the gate set taken into consideration.

The difference between the two cases lies in the presence of entanglement of purely local nature in the paramagnetic case, while both the ground states in the ordered and topologically frustrated phases have long-range quantum correlations [26, 30]. Long-range entanglement is less affected by the action of local gates, and even less so as the system size increases. On the other side, when the entanglement is local, local gates can easily reduce it, although the impossibility of exploring the whole Hilbert space with a reduced set of gates can still prevent its complete suppression.

As noted above, the entanglement spectrum statistics of the initial states are always (mostly) Poissonian, and indeed this is also the case at the end of the cooling if the non-universal gate set is employed. However, when the universal one is used and when there is sufficient local entanglement to act on, the final states display a Wigner-Dyson (WD) entanglement spectrum statistics. Before our work, WD statistics has always been observed in states with volume law entanglement, but indeed we can confirm that this is not the case for our states.

The paper is organized as follows: In Sect. 2 we introduced the model used to generate the input states of our cooling algorithm. Then, in Sect. 3 we describe in detail the cooling algorithm and the different sets of local gates. Afterward, in Sect. 4, we describe the results obtained, with a particular focus on how the size of the system, its quantum phase, and the gate set affects the evolution of the different states under the action of the cooling algorithm. In Sect. 5 we draw our conclusions.

2 The Model

As highlighted in the introduction, our goal is to apply an entanglement cooling algorithm (see also Sect. 3) to states that are ground states of the same Hamiltonian but in different macroscopic phases. We focus on the ground states of the one-dimensional spin-$1/2$ transverse field Ising model (TFIM), since it is a prototypical example of a many-body system possessing
Figure 1: Half-chain Rényi-2 entropy of the transverse-field Ising model. The quantum phase transitions (black dashed lines) separate the paramagnetic phase ($|J/h| < 1$) from the ferromagnetic ($|J/h| < 1$) and the frustrated antiferromagnetic one ($|J/h| > 1$). The excess entanglement in the frustrated AFM regime is an increasing function of the system size, which saturates at the thermodynamic limit [26]. In the inset, we plot the nearest-neighbor concurrence [33], where the excess concurrence in the frustrated phase decreases with the chain length. In the thermodynamic limit, the curves would be symmetric around the origin.

different macroscopic phases. It is defined by the following Hamiltonian,

$$H = J \sum_{i=1}^{N} \sigma_i^x \sigma_{i+1}^x - h \sum_{i=1}^{N} \sigma_i^z,$$

where the $\sigma^\alpha$ with $\alpha = x, y, z$ are the Pauli operators. Limiting our analysis to systems with periodic boundary conditions ($\sigma^x_{N+1} = \sigma^x_0$) made by an odd number of spins ($N = 2M + 1$ with $M \in \mathbb{N}$), the so-called frustrated boundary conditions (FBC), the model admits three distinct phases. When the local field dominates over the interaction term between spins, the system is in the paramagnetic phase (PARA), characterized by a gapped spectrum and a vanishing spontaneous magnetization in the longitudinal directions. On the contrary, when the interaction term dominates over the local field ($|J| > |h|$), we can realize two inequivalent phases depending on the sign of the interaction $J$. If the interaction is ferromagnetic (FM), i.e. when it favors parallel alignments ($J < 0$), the system shows a gapped magnetically ordered phase, with a finite magnetization in the $x$-direction [31]. On the other hand, in the case of an antiferromagnetic interaction (AFM, $J > 0$) the system is in a gapless topological frustrated phase [20–25], where the spontaneous magnetization is destroyed by the presence of a delocalized kink excitation. We remark that the TFIM is also integrable. We do not use this feature in our analysis, and we do not expect it to influence our results: to check this assumption, in Sec. 4 we will also add an integrability breaking term and apply the same algorithm to the resulting ground state.

A way to discriminate among the three model’s phases is by looking at the various kinds of bipartite entanglement [3, 32] that can be quantified by the Rényi entropies, defined as,

$$S_\alpha(\rho_\Lambda) = \frac{1}{1-\alpha} \log_2 \text{Tr}[\rho_\Lambda^\alpha],$$

which depends on the parameter $\alpha \in [0,1) \cup (1,\infty]$. In eq. (2), $\rho_\Lambda$ is the reduced density matrix obtained by tracing out from the ground state $|\Psi\rangle$ all the degrees of freedom of spins that lie outside the subset $A$. In the limit $\alpha \to 1^-$, the Rényi entropies reduce to the von Neumann entropy [3]: $S_1(\rho_\Lambda) = -\sum_k \lambda_k \log_2 \lambda_k$, where $\{\lambda_k\}$ is the set of eigenvalues of the reduced density matrix $\rho_\Lambda$.

In Fig. 1 we present the behavior of $S_2(\rho_\Lambda)$ as a function of the ratio $J/h$ for $h > 0$ in the case in which $A$ is made by $(N - 1)/2$ contiguous spins. For the sake of simplicity, we refer to the string made by $(N - 1)/2$ contiguous spins as the half chain. Differently from the other phases, in the topologically frustrated one, we observe a relevant dependence of the entanglement on the size of the chain that can be explained by taking into account that in such a phase the ground states are characterized by a delocalized excitation that increases the total amount of entanglement [26].

While the half-chain Rényi-0 entanglement entropy captures the non-local nature of entanglement, the concurrence [33] measures its local contribution. It is defined starting from the reduced density matrix obtained by tracing out the degrees of freedom of every site of the chain but two, which in this case we take as neighboring sites. In the inset of Fig. 1, we observe that contrary to what happens to $S_2$, the contribution of the delocalized excitation in the frustrated phase decreases as the chain length increases and vanishes in the thermodynamic limit. As a consequence, for large systems, the concurrences of the FM and the AFM phases coincide. However, due to the complexity of the cooling algorithm, our system will always be well below such a limit. Hence, for the analyzed cases, the amount of local entanglement of a ground state in the presence of a topological frustrated system will be relatively higher than the one coming from the FM phase.

Despite the differences in the entanglement entropies discriminating the various phases of the model, the statistical properties of the entanglement are almost the same. To highlight this fact, we focus on the probability density function of the consecutive level spacing ratio. Ordering the set of eigenvalues $\{\lambda_k\}$ of the reduced density matrix $\rho_A$ from the smallest to the biggest, the set of the consecutive level spacing ratios are

$$r_k = \frac{\lambda_{k+1} - \lambda_k}{\lambda_k - \lambda_{k-1}}, \quad k = 2, 3, 4, ..., 2^{[N/2]} - 1. \quad (3)$$

The elements of this set are distributed accordingly with different statistics, see Appendix B. The results obtained are shown in Fig. 2 for different parameter choices. We observe a significant deviation
Figure 2: Histograms for consecutive entanglement spectrum spacing of the reduced density matrix (RDM) eigenvalues of the ground states in different regimes of the TFIM Hamiltonian for $N = 35$ spins. Number of bins used is 125. We superimposed the analytical curves for the Wigner-Dyson (dashed) and Poisson (continuous) distribution for comparison. Due to the large system/subsystem sizes, we obtain good statistics within each single shot realization, which is what is plotted. The different regimes taken into account are: (a) Deep ferromagnetic (FM) regime with $J/h = -2.5$; (b) Ferromagnetic (FM) regime close to the quantum phase transition point at $J/h = -1.05$; (c) Paramagnetic (PARA) regime with $J/h = 0.75$; (d) Frustrated antiferromagnetic (frus. AFM) regime with $J/h = 2.5$. Except for results in panel (a), the entanglement spectrum spacing statistic always follows a Poisson distribution. In panel (a), the majority of the spacing falls in a narrow bin close to the origin because of the degeneracies of the entanglement spectrum for subsystems much bigger than the correlation length. For more details on the computation, see Appendix B.

3 The Entanglement Cooling Algorithm

Even if the Hamiltonian in eq. (1) is integrable, we recover the ground states in the different phases by exact diagonalization [35]. The reason is two-fold. First, with the explicit form in the spin basis (as compared to the form in terms of Bogoliubov fermions) the GS is more suitable for generic gate applications. A second reason for this choice comes from the computational bottleneck associated with the large number of gates applications needed to resolve the behavior, which also ready for the system sizes accessible by exact diagonalization requires several days of computation.

Once we have obtained the GS, we apply to it the entanglement cooling algorithm to test the stochastic irreversibility of its entanglement. This algorithm is nothing else but a Metropolis Monte Carlo (MC) algorithm [36] in which each step is made of two successive operations. The first one is the application of a single randomly chosen gate to the state coming from the previous step, i.e. a unitary ‘time’ evolution operator that transforms the state and can modify the value of the entanglement. The gates can be written as $U(t) = \exp(iO(t)\Delta t)$ with $\Delta t = \pi/10$ and are applied at randomly selected neighboring pairs of spins $\{l, l + 1\}$. The different kinds of gates generator $O^{(k)}$ are selected with uniform probability from the sets of deterministic operators tabulated in Table 1. Concerning the Hamiltonian in eq. (1), the two sets hold very different properties. The first is made by gates that preserve the parity symmetry of the Hamiltonian along $z$-direction. On the contrary, the gates in the second set, either violate the parity symmetry or are generated by non-integrable operators. It is worth noting that, while both sets of gates alone do not constitute complete sets of operations able to provide universal quantum computation, the two sets together do (see Ref. [37]).

The second operation is a filtration procedure, in which the new step can be accepted or rejected depending on the change of a cost function. The main difference between our algorithm and the traditional Metropolis MC ones is that in these last the cost function is represented by the energy, while in our case is the half-chain entanglement. We remark that this choice implements an interesting interplay between the locality of the applied gates and the global nature of the cost function. The new state is accepted or rejected with a probability equal to $\min\{1, \exp(-\Delta S_n/T)\}$ where $\Delta S_n = S_n^{\text{new}} - S_n^{\text{old}}$ denotes the difference between the averaged Rényi-$\alpha$ entropies ($S_n = N^{-1}\sum_i S_i$) where the average is made over all sets made of $(N - 1)/2$ contiguous spins. Even if in principle the average is taken over $N$ different partitions, at each step only two entropies need to be recomputed since local operations inside a single partition do not change the value of the entanglement.

Using the von Neumann entropy ($\alpha = 1$) or any of

<table>
<thead>
<tr>
<th>Set 1</th>
<th>Set 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>$O_{t}^{(1)} = \sigma_{l}^{x} \otimes \overline{l}<em>{l+1} + \overline{l}</em>{l} \otimes \sigma_{l+1}^{x}$</td>
<td>$O_{t}^{(4)} = \sigma_{l}^{x} \otimes \overline{l}<em>{l+1} + \overline{l}</em>{l} \otimes \sigma_{l+1}^{x}$</td>
</tr>
<tr>
<td>$O_{t}^{(2)} = \sigma_{l}^{z} \otimes \sigma_{l+1}^{z}$</td>
<td>$O_{t}^{(5)} = \sigma_{l}^{x} \otimes \overline{l}<em>{l+1} + \overline{l}</em>{l} \otimes \sigma_{l+1}^{y}$</td>
</tr>
<tr>
<td>$O_{t}^{(3)} = \sigma_{l}^{z} \otimes \sigma_{l+1}^{z}$</td>
<td>$O_{t}^{(6)} = \sigma_{l}^{y} \otimes \sigma_{l+1}^{z}$</td>
</tr>
</tbody>
</table>

Table 1: Table of local deterministic gate generators. Set 1 is parity preserving. Sets 1 & 2 taken together form the universal set [37].
the Rényi entropies ($\alpha \neq 1$) the entropy cooling algorithm leads to similar behavior. In particular, we observe that statistically, the inequality $S_q \geq S_{q'}$ for $q < q'$ is satisfied at any point of the algorithm’s execution [38]. For our analysis we employ the $\alpha = 2$ Rényi entropy [3, 39–41], since it can be computed faster than the von Neumann entropy since it can be computed directly through matrix-matrix multiplication, without the need for a diagonalization algorithm required by the von Neumann entropy. The main advantage of our approach is that it can be, and it has been implemented very efficiently (and in a parallelized way) using the architecture of modern graphical processing units (GPU) achieving a speed-up in matrix-matrix operation compared to standard CPUs (see Appendix A and Ref. [42].) Therefore, Rényi-2 entropy is an efficient choice on top of being observable via atomic and optical experiments [43–48]. Moreover, the $S_q$ is also connected to the information scrambling in quantum systems with averaged out-of-time-ordered correlations (OTOC), that are used to diagnose the onset of quantum chaos [49].

The cooling part of the algorithm refers to a fictitious temperature $T$ decreasing as a function of the Metropolis MC steps. Indeed, the temperature is lowered in an evenly spaced logarithmic temperature grid spanning the range $T \in [10^{-4}, 10^{-8}]$. In the literature, this procedure is usually referred to as the simulated annealing method and is often used in the efficient search for global minimum in various high-dimensional problems [9]. Moreover, as is typical in these kinds of approaches, we perform an average of over $M$ different Metropolis MC trajectories or stochastic quantum circuits.

4 Results

Having described both the entanglement cooling algorithm and the model used to generate the input states of the algorithm, we move to illustrate our results. The different types of evolutions of the entanglement obtained with the algorithm are summarized in Fig. 4. The plots in the left columns are obtained by choosing local gates uniformly at random from the first set, while in the right one, all local gates are considered to ensure a universal quantum computation (1 & 2). For more information about our efficient computational implementation, allowing us to reach systems up to $N = 21$ spins with a considerable amount of Metropolis MC steps and statistics, see Appendix A.

Deep in the FM cases, we observe that except for the smaller systems, the state displays almost perfect robustness of the entanglement regardless of the set of gates employed. This can easily be explained since in the FM phase, the ground state is well approximated by a $N$-qubit Greenberger–Horne–Zeilinger (GHZ) state [51]. Moving closer to the quantum phase transition adds local correlations on which the algorithm can act, but their contribution is hardly noticeable on the Rényi-2 for the parameters we chose. The GHZ state represents a global symmetric superposition of two orthogonal fully factorized states. Such states are well-known in the theory of quantum information, since they maximize the multipartite entanglement [52] while retaining no bipartite entanglement after local measurement on one of its parts [53]. In such a state, all the quantum correlations come out from the peculiar global superposition [30] and, therefore, it comes as no surprise that local gates acting at most on two neighboring sites of the chain are unable to change this entanglement. To simulate a global interaction with local interactions, we need a series of the latter which, taken individually, will increase the value of the entanglement of the state. The probability that this happens, considering how the cooling algorithm selects the different states, decreases exponentially as the size of the chain increases.

In the PARA regime, the dynamic of the entanglement is strongly affected by the chosen sets of local gates. After applying the first set, we observe that, after a small transient in which the total entanglement is slightly reduced, we arrive at an almost size-independent plateau for the entanglement. Conversely, when the complete set of local gates is taken into account the entanglement is considerably reduced, hence signaling a clear difference from the FM phase, where the entanglement cannot be removed.

In the final case, when the cooling algorithm is applied to the frustrated AFM ground state, both similarities and differences compared with the previous cases are evident. Similar to the PARA case, the entanglement is not completely robust, while akin to the
In panels (a) and (b), the results are obtained for the deep ferromagnetic (FM) ground-state ($J/h = -2.5$). In the panels (c) and (d), we present the results in the ferromagnetic (FM) regime close to the quantum phase transition point ($J/h = -1.05$); see Fig. 1. Next, in panels (e) and (f), the evolution obtained starting from ground states of the paramagnetic (PARA) phase ($J/h = 0.75$) is displayed. In the bottom panels (e) and (f), the evolutions obtained for ground states in the topologically frustrated antiferromagnetic (frus AFM) regime ($J/h = 2.5$) are shown.

FM phase the overall behavior of the entanglement evolution does not depend on the set of gates used in the algorithm. In particular, the entanglement remains initially locked to its initial value before being reduced. However, this reduction does not proceed until the entanglement is completely removed, but it rather stops at a size-dependent baseline value. Moreover, the rate of the entanglement cooling scales with the system size $N$. Increasing the system size, the rate of decay becomes less steep, indicating that the entanglement becomes extensively robust, in agreement with the delocalized single-particle interpretation [26].

In Fig. 5, we present a finite-size analysis of the (average) values reached by the entanglement entropy at the end of the cooling algorithm. In all cases but the frustrated one it is clear that we have reached a plateau, which constitutes a more or less significant reduction of entanglement, although the difference from the starting value decreases with the system size. For the AFM frustrated phase, it is not possible to extrapolate from our data the asymptotic regime. Nevertheless, we expect that the final value will keep increasing algebraically with the system size towards a value that is only slightly reduced compared to the initial one. This is consistent with the progressive difficulty of the algorithm in destroying the long-range entanglement in the W-state structure of the frustrated states as the number of sites is increased (while the local entanglement in the dressed kink states can be attacked and reduced) [55]. Interestingly, the initial plateau for which the entanglement remains within 5% of its initial value seems to increase with the system size in all considered cases, as shown in the right panel of Fig. 5.

Changing the cooling procedure, e.g. modifying the temperature gradient in terms of the MC steps, would yield slightly different results for the averaged final entanglement entropy. Nevertheless, the scaling of robustness with system size $N$ behavior looks ubiquitous and not influenced by these choices.

Thus, though all considered initial states share largely the same initial statistic for the entanglement spectrum, we see very different behaviors under the cooling evolution. This fact indicates that the reduced density matrix eigenvalue statistics are not sufficient to completely capture the underlying complexity of entanglement, in the sense of its robustness [14, 15]. However, the various phenomenologies can be explained easily. Let us start by considering the FM and PARA phases. Both phases present a finite energy gap that separates the ground states from the overlying excited states. Therefore, the ground states can be
Figure 5: Finite-size analysis of the plateaus reached when the entanglement cooling algorithm is applied using the universal gates set. Left (a) panel: Relative difference for the various states between their half-chain Rényi-2 entropy before and after the cooling, i.e. start and end, respectively. The end value of the entanglement is averaged over the different MC trajectory realizations. Right (b) panel: Initial plateaus size, defined as the number of MC steps after which the entanglement deviates by more than 5% from its starting value for the different cases considered in Fig 4.

obtained starting from their relative classical points ($\lambda = 0$ for the PARA phase and $\lambda = -\infty$ for the FM phase) using a quasi-adiabatic continuation [54]. This process generates entanglement of local nature that can be easily removed by an entanglement-cooling algorithm such as the one used in [30]. However, the two classical points present a notable difference. In the case of the paramagnetic phase, the classical ground state is single, and therefore the local entanglement generated by the adiabatic continuation process is the only one present in the system. In principle, one could therefore expect complete destruction of entanglement in the process. Our results indicate that a plateau at a finite value is reached also in this case. We speculate that this is due to the simulated annealing: after some local entanglement is removed by the algorithm, the systems reach an equilibrium between the instances in which moves that increase the entropy are accepted and those that are able to reduce the entanglement. Conversely, at the classical ferromagnetic point, the ground state is doubly degenerate. Adding a small transverse field breaks the symmetry by selecting a particular global superposition, producing a state similar to a GHZ state [27]. Therefore, the quasi-adiabatic deformation process can be seen starting from this state and the local entanglement thus generated is added to that associated with the global superposition.

Furthermore, by progressively increasing the magnetic field, one can follow a quasi-adiabatic deformation of the ground state that acquires additional local entanglement in addition to the non-local one associated with the global superposition. While the local entanglement can be directly removed by the cooling algorithm, the global one is resilient against its action, since it is based on local gates, unless very small sizes are considered. Hence, the resilience of entanglement in the FM phase that we see in our data to a value close to $\log_2(2) = 1.$

On the contrary, in the AFM phase, due to the simultaneous presence of an odd number of spins, periodic boundary conditions, and antiferromagnetic interactions, the system exhibits topological frustration. In this phase (due to the absence of an energy gap) the quasi-adiabatic continuation cannot be straightforwardly applied. However, as shown by previous works [26, 55], in several respects the ground state can still be seen as a global state with multiparticle entanglement on top of which, through a local deformation, an entanglement of local nature has been generated. Unlike the FM phase, the global state is no more represented by a GHZ state but by a sort of $W$-state, that is, in physical terms, a linear superposition of single kink states [55]. This fact is crucial since, differently from the first kind of states, part of

Figure 6: Consecutive entanglement spectrum spacing ratio histogram of the reduced density matrix (RDM) eigenvalues of the ground state of the non-integrable Hamiltonian in eq. 4. We also report the average spacing ratio. The Hamiltonian parameters used are: $J = -1.25, J' = 0.2$ and $\hbar = 1$. The histogram has 125 bins and, having a single ground state for this model, the number of sampled RDM eigenvalues is 1024.

Figure 7: Evolution of the entanglement entropy under the cooling algorithm, starting with the ground state of the non-integrable model with Hamiltonian in eq. (4) in the paramagnetic phase with parameters $J = -1.25, J' = 0.2$ and $\hbar = 1$. Data points represent the averaged half-chain Rényi-2 entanglement entropy over $M = 96$ Metropolis MC trajectories. Both the non-universal (panel (a)) and universal (panel (b)) gate sets (see Table 1) are considered in the entanglement cooling procedure.
the entanglement of the $W$-states is local, since they are the family of states with multipartite entanglement [56–65] that maximize the amount of bipartite entanglement after local measurement on one of its part [52, 53]. The presence of such bipartite entanglement, absent in the states coming from the FM phase, provides a path along which the entanglement cooling algorithm can act, reducing, but not removing, the amount of entanglement in the system. Moreover, for every single pair of spin, the amount of bipartite entanglement in a $W$-state reduces with the system size, hence explaining why the difference between the values of the initial and final plateaus in the dynamics of the $W$-state reduces as $N$ increases.

A Non-integrable Extension: the ANNNI Model

As we mentioned above, the TFIM is also known for being exactly solvable. While we do not expect this feature to influence the behavior of the entanglement under the cooling algorithm, to make sure that this is indeed the case, we consider now a non-integrable extension of the TFIM, namely the ANNNI model [66–68]:

$$H = J \sum_{j=1}^{N} \sigma_j^x \sigma_{j+1}^x + J' \sum_{j=1}^{N} \sigma_j^z \sigma_{j+2}^z - h \sum_{j=1}^{N} \sigma_j^z,$$  \hspace{1cm} (4)

where we assume periodic boundary conditions and focus on the ferromagnetic regime $J = -1.25$ with $h = 1$. The next-to-nearest neighbor interaction we added compared to the traditional TFIM in eq. (1) breaks the integrability. We fix this additional coupling to be antiferromagnetic $J' = 0.2$, to add some amount of frustration and correlation in the system, while keeping it subdominant with respect to the larger nearest neighbor coupling to ensure we do not cross a transition to a extensively frustrated phase. The zero-temperature phase diagram of this model has been worked out over the years in the literature [68–72]. The relatively high magnetic field pushes the system into a paramagnetic phase. While the integrability of the clean TFIM allowed us to construct its ground states for relatively large chains, for this model we are not able to push the system size sufficiently: in Fig. 6 we report the level spacing statistics for the initial state entanglement spectrum, which looks Poissonian, although the limited statistics give rise to some spurious fluctuations. In Fig. 7, we present the results of the entanglement cooling algorithm using both available gate sets. We observe similar results as in the paramagnetic phase of the integrable TFIM (see panels (e) and (f) in Fig. 4). Overall, the universal gate set is better at entanglement cooling compared to the non-universal one. We decided to test the paramagnetic phase of the ANNNI

Figure 8: Consecutive entanglement spectrum spacing ratio histogram of the reduced density matrix (RDM) eigenvalues for the final states at the end of the entanglement cooling algorithm for the integrable TFIM and its non-integrable extension, the ANNNI model in its paramagnetic phase. The states used for these averages over $M = 96$ samples are those obtained in Fig. 4 and 7. The histograms have been generated using 125 bins. In the left panels, resulting from an algorithm employing only parity preserving gates, the final entanglement spectrum statistics is always Poissonian as it was at the beginning (see Fig. 2 and 6). In the right panels, obtained using universal gates, the entanglement spectrum has always evolved toward a Wigner-Dyson distribution, except for the deep ferromagnetic case, characterized by very little initial local entanglement.
model in order to have purely local entanglement for our algorithm to act upon, and we notice that starting with a higher value compared to the TFIM, in the end, the algorithm settles also on a higher plateau.

**Entanglement Spectrum Statistics**

As discussed before, the entanglement spectrum statistics of all initial states we considered follow a Poisson distribution (with deviations related to exact degeneracies), as is often the case with the ground states of the local Hamiltonian. In Fig. 8 we report the average entanglement spectrum statistics obtained at the end of the cooling algorithm for all cases we considered so far (namely the final points for the larger system sizes in Fig. 4 and 7).

While using only the first gate set, the entanglement spectrum remains Poissonian, using universal gates we always observe an evolution toward exact Wigner-Dyson statistics in all cases, except when starting with a state with negligible local entanglement, such as in the deep ferromagnetic regime. Our interpretation is that, in the process of removing local entanglement, our algorithm also performs a sort of scrambling which results in the observed “chaotic” spectrum. But this can happen only if there is sufficient local entanglement to act upon, meaning that a sufficiently large number of moves are accepted before reaching the final plateaus. In Fig.9 we compare the consecutive RDM level spacing statistics for the final states obtained using universal gates in the deep ferromagnetic case and the frustrated AFM one to highlight their difference and to show the evolution of the entanglement ratio during the cooling process as a function of the number of Montecarlo steps. We stopped the evolution after an equivalent amount of Metropolis MC step applications, and we choose this number to be $10^5$ steps, but the results are the same if we evolve the state even further.

We remark that, up to now, the appearance of a Wigner-Dyson entanglement spectrum has always been associated with the presence of a volume law for the corresponding entanglement entropy. In Fig. 10 we present the values of the Rényi-2 entropy as a function of the subsystem length for the initial and final states in our analysis. As expected, the initial ground states of our Hamiltonians in their various phases always satisfy an area law, with a relative violation for the frustrated case where the entanglement can grow with the system size, while remaining upper-bounded, as explained in Ref. [26]. Similarly, the final states obtained at the end of the algorithm never approach a volume law and are compatible with an exact area law. To the best of our knowledge, these results are unprecedented and deserve additional analysis in the future, since they imply that it is possible to construct local models with area law whose entanglement spectrum follows a Wigner-Dyson distribution, although they have never been found so far.

**5 Conclusion**

In our work, we tested the evolution of entanglement induced by applying a cooling algorithm to different states, all but one being the ground states of Hamiltonians in the same class of integrable models, but in different macroscopic phases. For completeness, to check the independence of our results from integrability, we also considered a non-integrable extension. Each of the chosen states is characterized by an initial entanglement spectrum following the same Poissonian distribution. Quite surprisingly, states in different phases evolve differently, depending on the macroscopic phase and on the different sets of gates employed. In the paramagnetic phase, the ground state can be obtained by adiabatic continuation from a fully factorized, classical one. Thus, it contains only
local correlations whose entanglement the cooling algorithm can progressively remove when a complete (universal) set of local unitaries are employed. In the ferromagnetic phase, in agreement with the quasiadiabatic continuation approach, the ground states can be seen as locally deformed GHZ states [30, 54]. Hence, after the cooling algorithm has destroyed the local entanglement, the residual global entanglement (characteristic of a GHZ state) is resilient to such a process. Its destruction would require a series of moves that are accepted even if they are increasing the entanglement until they overcome the barrier separating a GHZ state from a separable one, but such an event is exponentially suppressed with the system size.

Finally, in the presence of topological frustration, we are not in the range of validity of the quasiadiabatic continuation. However, it was proved [26, 55] that the ground state can still be seen as a locally deformed W-state. The W-states, as well as the GHZ states, are states that, while maximizing the multipartite entanglement, also show a large bipartite entanglement, unlike the latter states. Its presence, being also somewhat non-local, provides an amount of entanglement that the cooling algorithm can destroy, but less efficiently as the system size increases.

Overall, our results show a much more varied picture than what one might have expected. The reason why it remained unnoticed until now is mostly due to the fact that previous analyses have started from randomly generated states in which geometrical and topological properties are absent. Indeed, the three phases we considered present entanglements of three different types which are not reflected by the entanglement spectrum statistics: purely local, GHZ-like, and W-state-like, with the topologically frustrated chain exhibiting all of them. Indeed, our data suggest that the evolution induced by the cooling algorithm is affected by the geometrical and topological properties of the ground state manifold of the initial state, as well as the presence, and the kind, of multipartite entanglement. Moreover, we observe that the choice of gate set affects significantly the amount of entanglement that the algorithm is able to destroy and thus the value of the final plateaus.

In fact, in all cases, we observe that, contrary to what one could have expected, the entanglement cooling algorithm is never able to completely destroy all local entanglement. It rather settles on plateaus in which some sort of meta-stable equilibrium is reached between the amount of entanglement that is reduced at a given step and the one that is added by a move accepted because of the finite temperature in the annealing process.

A striking, completely unexpected, behavior, which we believe is connected with the nature of the plateaus we just discussed, is that when a sufficient amount of local entanglement is present in the initial state and eventually destroyed by the algorithm. In these cases, our results show that the final state features a Wigner-Dyson distribution of the level statistics of its entanglement spectrum. Effectively, while cooling, the algorithm has also scrambled the entanglement spectrum and increased its complexity. As Wigner-Dyson statistics has been connected with higher robustness for it, at this point the algorithm remains stuck in a plateau, unable to destroy additional entanglement effectively. Note that, since this statistic has been observed before only in states obeying a volume law for the entanglement entropy, our findings disprove the expectation that the two are connected. It would be interesting to understand what are the conditions for which local Hamiltonians can generate area law ground states with the Wigner-Dyson entanglement spectrum, but this is a subject for future work.

Before concluding, we underline that our work is relevant from a methodological point of view as well. In particular, we use GPU-enhanced calculations to accelerate and scale the entanglement cooling protocol. The numerical codes, data, and plotting scripts employed for our cooling algorithm are provided in Refs. [79, 80].

Acknowledgments

S.M.G., F.F., and G.T. acknowledge support from the QuantiXLie Center of Excellence, a project co-financed by the Croatian Government and European Union through the European Regional Development Fund – the Competitiveness and Cohesion (Grant KK.01.1.1.01.0004). F.F. and S.M.G. also acknowledge support from the Croatian Science Foundation (HRZZ) Projects No. IP–2019–4–3321. J.O., N. M. and D.D. recognize the support of the Croatian Science Foundation under grant number HRZZ-UIP-2020-02-4559 and the European Regional Development Fund under the grant KK.01.1.1.01.0009 - DATACROSS. The authors gratefully acknowledge the HPC RIVR consortium (www.hpc-rivr.si) and EuroHPC JU (eurohpc-ju.europa.eu) for funding this research by providing computing resources of the HPC system Vega at the Institute of Information Science (www.izum.si). Part of this research was performed using the advanced computing service - supercomputer Supek - provided by the University of Zagreb University Computing Centre - SRCE.

A Computational details and speed-up of execution

In the cooling algorithm, the computation of the Rényi-2 entropy (see Sect. II) is the major computational barrier. Although diagonalization is replaced by matrix-matrix multiplication to calculate entropy,
it still remains the main bottleneck, accounting for most of the execution time. Since the size of the matrix and the number of matrix-matrix multiplications (i.e. the number of Monte Carlo steps) cannot be reduced, the algorithm can be speed-up by executing the matrix multiplication on graphical processing units (GPU). GPUs are best suited for matrix operations due to their massively parallel (many cores) architectures, delivering up to an order of magnitude better performance compared to (multithreaded) CPUs. The biggest disadvantage is that GPUs cannot perform well if the matrices are too small. In this case, a slowdown can even be observed compared to CPU.

In our experiments, we display systems with \( N = 21 \) spins using matrices of size \( 2048 \times 1024 \), which are too small to achieve good utilization of GPU resources. The solution is to pack multiple matrix multiplications into a single and large matrix-matrix multiplication operation called batched operation to achieve high performance. For this particular research, we have proposed a different approach based on multiple concurrent tasks instead of a single batched operation. The detailed implementation, speed-up achieved and performance discussion can be found in [42], while we give only a brief description below.

The mutual independence of the Metropolis MC simulations provides a trivial parallelization, as no synchronization or communication is required, making the problem embarrassingly parallel. The Message Passing Interface (MPI) is used to associate each trajectory with a single MPI process, while multiple processes can be run concurrently on the same device using the NVIDIA Multi-Processing Service (MPS). This approach allows spatial sharing of device resources with the disadvantage of slowing down the computation of a single Metropolis MC trajectory (i.e. matrix multiplication), but this has the advantage of allowing multiple simulations to be computed in parallel, effectively overlapping the computations from different trajectories. Analysis has shown [42] that overlapping has a much better effect on overall performance than slowing down the single Metropolis MC trajectory. For the largest use case with a system size of \( N = 21 \) spins and 96 MC trajectories, we found that the best configuration is 12 procedures (MPI processes) per NVIDIA A100 GPU, which took about 116 hours on 8 GPUs. Overall, we achieved a speed-up of two orders of magnitude on a single Metropolis Monte Carlo (MC) trajectory compared to the standard CPU implementation.

B Entanglement spectrum statistics details

Consecutive level spacing ratios: In the manuscript, we focus on the consecutive level spacing ratio probability density function \( P(r) \). This is a suitable quantity of interest in the case of quantum many-body applications where statistics quality is heavily bound by the Hilbert space dimension [73–75]. In particular, using an ascending (from smallest to the biggest) ordered set of eigenvalues \( \{\lambda_k\} \) of the reduced density matrix, we compute the consecutive ratios as

\[
r_k = \frac{\lambda_{k+1} - \lambda_k}{\lambda_k - \lambda_{k-1}}, \quad k = 2, 3, 4, ..., 2^{[N/2]} - 1.
\]

Taking the values of the generated list of \( \{r_1, r_2, ...\} \) we plot a normalized histogram, where we disregard any spurious (and rare) values that are larger than \( r_j > 10.0 \) for proper normalization and binning of the histogram.

Averaged consecutive spacing ratio: Additionally, we consider the averaged consecutive spacing ratio defined as

\[
\bar{r} = \left\langle \frac{\min(s_{k,j}, s_{k+1,j})}{\max(s_{k,j}, s_{k+1,j})} \right\rangle_{2^{[N/2]}/2M},
\]

where we define the spacings as \( s_{k,j} = \lambda_{k+1,j} - \lambda_{k,j} \), and the index \( j = 1, 2, ..., M \) refers to the different reduced density matrices considered (ensemble).

From Random Matrix Theory (RMT) it is known that, in the Poisson case, the ratio follows the probability distribution \( P_{\text{Poisson}}(r) = (1 + r)^{-2} \). On the contrary, for the Wigner-Dyson distribution, we have \( P_{\text{WD}} = Z^2_\beta^{-1}(r + r^2)^{\beta/2}(1 + r^2)^{-1-3/2\beta} \) with \( \beta = 2 \) (Gaussian Unitary Ensemble - GUE) and \( Z_{\beta=2} = 4\pi/81\sqrt{3} \). The predictions for the averages are thus \( \bar{r}_{\text{Poisson}} = 2\ln2 - 1 \approx 0.386 \) and \( \bar{r}_{\text{WD}} = 2\sqrt{3}/\pi - 1/2 \approx 0.602 \) [74].

Advantages for the integrable TFIM: The integrable TFIM can be mapped to free-fermions using the Jordan-Wigner transformation [31] and solved explicitly [76]. The half-chain reduced density matrix (RDM) \( \rho_A \) of the ground state for this model can then be computed using the prescription detailed in Ref. [77]. The domain \( A \), due to translational invariance of the spin chain, comprises any \([N/2]\) (half-chain) contiguous spins. Note that with this approach, we are able to push our numerical considerations much beyond the Exact Diagonalization (ED) efficiency [35]. Additionally, using mpmath [78] Python library we boost the precision of the fermionic approach, as double precision does not suffice for this kind of analysis. For the largest example, we shall consider, i.e. \( N = 35 \) and half-chain of size of the RDM of \( 17 \) the number of eigenvalues of the RDM is \( 131072 \). In Fig. 11 we show the results for the entanglement spectrum statistics for a single (\( M = 1 \)) realization and different chain lengths, to show how important it is to reach adequate system sizes to obtain reliable statistics.
The chain sizes are $N = 13, 15, 21, 25, 31, 35$; while the half-chain sizes of the RDM are $\lfloor N/2 \rfloor = 6, 7, 10, 12, 15, 17$. We only focus on the range $0 < r < 10$ where we consider 50 bins to make the histogram. Inside the plot, we provide the averaged consecutive spacing ratio number we compute from the given spectrum. Parameters for this case are $h = 1.0$ and $J = 2.5$ which is the antiferromagnetic frustrated regime. This is single-shot data ($M = 1$), and no ensemble averaging is performed.

Figure 11: Consecutive entanglement spectrum spacing ratio histogram of the RDM eigenvalues for different systems size. The chain sizes are $N = 13, 15, 21, 25, 31, 35$; while the half-chain sizes of the RDM are $\lfloor N/2 \rfloor = 6, 7, 10, 12, 15, 17$. We only focus on the range $0 < r < 10$ where we consider 50 bins to make the histogram. Inside the plot, we provide the averaged consecutive spacing ratio number we compute from the given spectrum. Parameters for this case are $h = 1.0$ and $J = 2.5$ which is the antiferromagnetic frustrated regime. This is single-shot data ($M = 1$), and no ensemble averaging is performed.

References


[63] D. Girolami, T. Tufarelli, and C. E. Susa, Quantifying genuine multipartite correlations and


